ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Πρόβλεψη τιμής κλεισίματος μετοχών με τεχνικές Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Του

ΣΠΥΡΙΔΩΝΟΣ ΣΠΥΡΟΠΟΥΛΟΥ

Επιβλέπων : Βασίλειος Ασημακόπουλος
Καθηγητής Ε.Μ.Π.
Υπεύθυνος : Ευάγγελος Σπηλιώτης
Διδάκτωρ Ε.Μ.Π.

Αθήνα, Ιούνιος 2019

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Πρόβλεψη τιμής κλεισίματος μετοχών με τεχνικές Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Του

ΣΠΥΡΙΔΩΝ ΣΠΥΡΟΠΟΥΛΟΥ

Επιβλέπων : Βασίλειος Ασημακόπουλος
Καθηγητής Ε.Μ.Π.
Υπεύθυνος : Ευάγγελος Σπηλιώτης
Διδάκτωρ Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12η Ιουλίου 2019

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Βασίλειος Ασημακόπουλος	Ιωάννης Ψαρράς	Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2019

(Υπογραφή)

(Υπογραφή)

.....

ΣΠΥΡΙΔΩΝ ΣΠΥΡΟΠΟΥΛΟΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Σκοπός της παρούσας διπλωματικώς εργασίας είναι η ανάπτυξη ενός αλγορίθμου που διενεργεί αγοραπωλησίες μετοχών του δείκτη S&P500 αξιοποιώντας τεχνικές μηχανικής μάθησης. Οι αυτοματοποιημένες αγοραπωλησίες μετοχών αποτελούν ένα πολύ σημαντικό κομμάτι του σύγχρονου χρηματοπιστωτικού συστήματος καθώς τα τελευταία χρόνια η λήψη αποφάσεων που σχετίζεται με αγοροπωλησίες μετοχών βασίζεται σε πολύ μεγάλο βαθμό σε τεχνικές προβλέψεων και ειδικότερα σε μεθόδους μηχανικής μάθησης.

Έτσι, λαμβάνοντας υπ'όψιν την προβλεπτική ικανότητα διαφόρων τεχνικών μηχανικής μάθησης αλλά και εξετάζοντας διάφορος τύπους χαρτοφυλακίων μετοχών, στα πλαίσια της παρούσας διπλωματικής αναπτύχθηκε ένας αλγόριθμος που αυτοματοποιεί την όλη διαδικασία και παράγει προτάσεις για διάφορα σενάρια αγοροπωλησίας.

Αρχικά ο αλγόριθμος βελτιστοποιήθηκε με βάση τις παραμέτρους των εξεταζόμενων τεχνικών μηχανικής μάθησης και στη συνέχεια χρησιμοποιήθηκε προκείμενου να εξαχθούν προβλέψεις για διάφορα χαρτοφυλάκια. Η τελική κερδοφορία εξετάστηκε ανά περίπτωση μέσω αντίστοιχων προσομοιώσεων. Η εργασία παρουσιάζει αναλυτικά τις μετρήσεις που έγιναν για τα επιμέρους βήματα της προτεινόμενης μεθοδολογίας και αποδεικνύει την αξία των εξεταζομένων τεχνικών για την υποστήριξη αποφάσεων σε χρηματιστήρια.

Λέξεις κλειδιά

Μηχανική Μάθηση, Προβλέψεις, S&P500, Χρηματιστήριο, Τιμή κλεισίματος, Χαρτοφυλάκιο

Abstract

The scope of this thesis is the development of a tool that implements stock trading, using stocks listed on S&P500, supported by machine learning decision methods. The automated stock trading is a very important section of the financial world since the recent years decision making in stock markets is greatly based on forecasting methods and especially on machine learning methods.

So, taking into consideration the forecasting ability of machine learning methods and Testing many types of stock portfolios, on this thesis, an algorithm have been developed which automates the process and produces propositions for such scenarios.

Initially, the algorithm has been optimized based on the parameters of the used machine learning methods and then it has been used to predict for several portfolios. This thesis presents both the experiments and the individual steps of the methodology and proves the value of the methods used in order to support decision making processes in the stock market.

Keywords

Machine learning, predictions, S&P500, Stock Market, Closing Price, Portfolio

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Ασημακόπουλου για την ανάθεση της παρούσας διπλωματικής εργασίας καθώς μου έδωσε την ευκαιρία να ασχοληθώ με το ενδιαφέρον αντικείμενο της μηχανικής μάθησης. Επίσης θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Ψαρρά και τον Καθηγητή κ. Ασκούνη για την τιμή που μου κάνουν, να είναι μέλη της τριμελούς επιτροπής. Ακόμα, θα ήθελα να ευχαριστήσω τον ερευνητή κ. Βαγγέλη Σπηλιώτη για τις πολύτιμες συμβουλές και καθοδήγηση που μου προσέφερε όλο αυτό το διάστημα εκπόνησης της παρούσας διπλωματικής. Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου για την στήριξη κατα την διάρκεια των σπουδών μου.

Περιεχόμενα

Περιεχόμενα Εικόνων

Περιεχόμενα Πινάκων

ΚΕΦΑΛΑΙΟ 1 ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο και σκοπός της εργασίας

Η τεχνητή νοημοσύνη, οι εφαρμογές και οι επιπτώσεις της, αποτελούν ένα από τα σημαντικότερα θέματα συζήτησης της επιστημονικής, της οικονομικής, της πολιτικής και εν γένει της παγκόσμιας κοινότητας, για τη διαμόρφωση του μέλλοντος, με την μηχανική μάθηση να κατέχει εξέχουσα θέση στην όλη διαδικασία, ως ένας από τους παλαιότερους ερευνητικούς τομείς της. Ως ερευνητικός τομέας, η μηχανική μάθηση έχει ασχοληθεί με πλήθος επιμέρους προβλημάτων και έχει προτείνει διάφορους μεθόδους επίλυσης. Τα τελευταία χρόνια έχουν αναπτυχθεί πολλές σχετικές εφαρμογές που ποικίλουν ως προς το περιεχόμενο, από συστήματα εύρεσης χρήσης κλεμμένων πιστωτικών καρτών έως τα αυτόνομα οχήματα.

Σε αυτή την διπλωματική εργασία ,λοιπόν, εξετάζεται ο τρόπος με τον οποίο αυτό το σύγχρονο εργαλείο μπορεί να φανεί χρήσιμο στην χρηματοοικονομική ζωή και συγκεκριμένα στις χρηματιστηριακές αγοραπωλησίες. Συγκεκριμένα γίνεται προσπάθεια πρόβλεψης της τιμής κλεισίματος κάποιων μετοχών της Wall Street με τεχνικές μηχανικής μάθησης και συγκεκριμένα νευρωνικών δικτύων, CART, Random Forests, Support Vector Machine. Στο τέλος γίνεται σύγκριση των αποτελεσμάτων μεταξύ τους. Επίσης χαράσσονται πλάνα επενδύσεων με βάση τις εν λόγω προβλέψεις.

Μια καλή προσέγγιση της μελλοντικής τιμής των περιουσιακών στοιχείων και συγκεκριμένα των μετοχών αποτελεί την πηγή της εξασφάλισης κέρδους από την χρηματιστηριακή αγορά , είτε για ιδιώτες επενδυτές είτε για επαγγελματικά επενδυτικά εταιρικά σχήματα δηλαδή επενδυτικές τράπεζες, funds κ.λ.π. . Η ανάγκη λοιπόν, από τον κόσμο των χρηματαγορών για όσο το δυνατόν καλύτερη πρόβλεψη της μελλοντικής αξίας των χαρτοφυλακίων τους, καθώς και η ραγδαία αύξηση της μελετητικής δραστηριότητας στον χώρο της μηχανικής μάθησης, οδήγησε στην χρήση τέτοιων προβλεπτικών μοντέλων. Οι προβλέψεις που εξάγονται από το μοντέλο αποτελούν ένα εργαλείο στα χέρια των ειδικών, ακριβέστερο σε σχέση με τις κλασικές τεχνικές στατιστικής πρόβλεψης, έτσι ώστε να μεγιστοποιήσουν το κέρδος τους απο τις συναλλαγές αυτές και εν γένει να καθορίζουν είτε την βραχυχρόνια είτε και την μεσοπρόθεσμη στρατηγική τους στην αγορά και κατά συνέπεια την ίδια την ύπαρξη τους στον πολύ ανταγωνιστικό κλάδο των χρηματοοικονομικών υπηρεσιών.

Επιπρόσθετα, αξίζει να σημειωθεί, πως ένας ακόμα παράγοντας που οδήγησε τους επαγγελματίες επενδυτές να στραφούν στην στατιστική και την τεχνολογία για παραγωγή προβλέψεων, είναι ο αντίκτυπος των ανθρώπινων συναισθημάτων στην λήψη αποφάσεων στις χρηματιστηριακές συναλλαγές. Οι αγοραπωλησίες μετοχών

λοιπόν, γίνονται αυτόματα με βάση το κατάλληλα διαμορφωμένο μοντέλο. Έτσι ο αλγοριθμικός τρόπος μπορεί να εξαλείψει τις ανθρώπινες προκαταλήψεις αλλά και να πραγματοποιήσει την συναλλαγή πολύ γρηγορότερα. Τον ήδη γνωστό αλγοριθμικό τρόπο έρχεται να βελτιώσει ακόμα περαιτέρω η μηχανική μάθηση. Πλέον υπάρχει η δυνατότητα να γίνεται παρακολούθηση ακόμα περισσότερων αγορών παγκοσμίως σε πραγματικό χρόνο και να γίνει αποδοτικότερη, άρα και πιο επικερδής τελικά, η επεξεργασία των ιστορικών στοιχείων που θα οδηγήσει σε μια σωστή πρόβλεψη.

Η εφαρμογή τέτοιων τεχνικών γίνεται εντονότερη τα τελευταία χρόνια. Πλέον πολλοί επενδυτές κάνουν χρήση της μηχανικής μάθησης με τον τρόπο που περιγράφηκε παραπάνω, αλλά αξίζει να σημειωθεί πως υπάρχουν funds (π.χ. Two Sigma investments) που χαράσσουν την στρατηγική τους αποκλειστικά με τέτοια μοντέλα μηχανικής μάθησης.

1.2 Οργάνωση Εργασίας

Αρχικά στο 2ο Κεφάλαιο παρουσιάζονται τα διάφορα είδη προβλέψεων γίνεται περιγραφή γενικότερα των διαδικασιών ποσοτικής πρόβλεψης, δίνονται στοιχεία σχετικά με διάφορες τεχνικές προβλέψεων και ορίζονται οι βασικοί δείκτες ακρίβειας που χρησιμοποιούνται γενικά από τους ερευνητές.

Ύστερα, στο 3ο Κεφάλαιο ορίζεται η μηχανική μάθηση, τα είδη μηχανικής μάθησης που χρησιμοποιούνται ευρέως και διατυπώνεται η αναλυτική λειτουργία των Τεχνητών Νευρικών Συστημάτων, που χρησιμοποιήθηκαν και στην παρούσα διπλωματική.

Στο 4ο Κεφάλαιο ορίζονται τα δένδρα αποφάσεων. περιγράφεται η υλοποίηση του αλγορίθμου CART και οι λειτουργίες του καθώς και η υλοποίηση του αλγορίθμου Random Forest και άλλα που είναι μια διαφοροποίηση του CART. Επίσης διατυπώνονται συμπεράσματα για τους δύο αυτούς αλγορίθμους.

Στο 5ο κεφάλαιο δίνεται η λειτουργία των τεχνικών SVM (Support Vector Machine) που είναι μια ακόμα διαδεδομένη τεχνική μηχανικής μάθησης.

Στο 6ο Κεφάλαιο παρουσιάζονται οι μετρήσεις που έγιναν προκειμένου να αναπτυχθεί το εργαλείο προβλέψεων, περιγράφονται αναλυτικά οι παράμετροι που δοκιμάστηκαν προκείμενου να επιτευχθεί η βελτιστοποίηση της τεχνικής. Ακόμα δίνεται και μια περιγραφή των δεδομένων που χρησιμοποιήθηκαν (dataset).

Στο 7ο Κεφάλαιο παρουσιάζονται οι στρατηγικές επενδύσεων που προέκυψαν με βάση τα αποτελέσματα των μετρήσεων που προηγήθηκαν στο 6ο κεφάλαιο και διάφορα χαρτοφυλάκια μετοχών.

Στο 8ο γίνεται μια κριτική ανάλυση των αποδόσεων των χαρτοφυλακίων και τέλος, εξάγονται συμπεράσματα για την παρούσα διπλωματική στο σύνολο της.

ΚΕΦΑΛΑΙΟ 2 ΤΕΧΝΙΚΕΣ ΠΡΟΒΛΕΨΕΩΝ ΚΑΙ ΔΕΙΚΤΕΣ ΑΚΡΙΒΕΙΑΣ

2.1 Προβλέψεις

Οι προβλέψεις αποτελούν βασικό κομμάτι της καθημερινότητας κάθε ανθρώπου, καθώς αποτελούν έναν γνώμονα για τις αποφάσεις που καλείται να πάρει είτε ως άτομο είτε ως μέλος της κοινωνικής πραγματικότητας. Αποφάσεις με βάση τις προβλέψεις παίρνονται σε όλες τις πτυχές της ζωής.

Οι προβλέψει ωστόσο, είναι πολύ σημαντικό κομμάτι και για τις στρατηγικές αποφάσεις που καλούνται να πάρουν οι επιχειρήσεις, ατομικές, μικρές ή μεγάλες. Από την πρόβλεψη ζήτησης του σουπερ μάρκετ της γειτονιάς έως την πρόβλεψη πωλήσεων της Apple. Η ακρίβεια της πρόβλεψης λοιπόν είναι ζωτικής σημασίας καθώς κρίνει, πολλές φορές την ίδια τη βιωσιμότητα του οργανισμού.

Τα τελευταία 30 χρόνια υπάρχει συντελείται ραγδαία ανάπτυξη στον τομέα της επιχειρησιακής έρευνας που έχει ως επακόλουθο και το ιδιαίτερα αυξημένο ενδιαφέρον για το επιστημονικό πεδία των προβλέψεων. Πρακτικά η μελέτη των προβλέψεων χρησιμοποιείται για την ανάγκη εξασφάλισης έναντι των κινδύνων που πηγάζουν απο την αβεβαιότητα του μέλλοντος. Όλα τα παραπάνω στοιχειοθετούν την σημασία των προβλέψεων σε όλα τα πεδία που αντιμετωπίζουν το ζήτημα της λήψης απόφασης.

Από την άλλη μεριά, η τομέας της πρόβλεψης έχει δεχτεί δυσμενείς κριτικές για την ανικανότητα του να προβλέψεις έγκαιρα επερχόμενες αλλαγές καθώς και για τα μεγάλα σφάλματα που κάποιες φορές προκύπτουν. Όμως μεγάλο μέρος της κριτικής είναι ουσιαστικά ανεδαφική καθώς η πρόβλεψη και είναι προφητεία. Το μέλλον κρύβει πολλές εκπλήξεις και τα σφάλματα είναι αναπόφευκτα.

2.2 Κατηγορίες προβλέψεων

1. Ποσοτικές μέθοδοι

Προκειμένου να πραγματοποιήσουμε προβλέψεις με τις ποσοτικές μεθόδους είναι αναγκαίο να έχουμε στη διάθεση μας μεγάλο όγκο πληροφοριών αλλά και να θεωρήσουμε ότι το μέλλον θα λειτουργεί όπως και το παρελθόν. Υπάρχουν δύο υποκατηγορίες ποσοτικών μεθόδων πρόβλεψης: το μοντέλο χρονοσειρών (time series model) και το αιτιοκρατικό ή επεξηγηματικό μοντέλο (causal relationship).

Το μοντέλο χρονοσειρών, είναι το πιο διαδεδομένο είδος ποσοτικού μοντέλου πρόβλεψης. Για την εφαρμογή του πρέπει να υπάρχουν ιστορικά στοιχεία για το μέγεθος που θα επιχειρηθεί να προβλεφθεί. Το μοντέλο χρονοσειρών βασίζεται στην υπόθεση ότι η μεταβολή της τιμές του μεγέθους ακολουθεί ένα συγκεκριμένο λανθάνον πρότυπο που επαναλαμβάνεται στο χρόνο και παραμένει σταθερό. Οι προβλέψεις παράγονται με την αναγνώριση αυτού του προτύπου και την προέκταση του στο μέλλον. Παράδειγμα τέτοιων μεθόδων είναι η εξομάλυνση (smoothing), η αποσύνθεση (decomposition), και οι αυτοπαλινδομικές μέθοδοι (ARIMA)

Το αιτιοκρατικό μοντέλο στηρίζεται στην βασική υπόθεση ότι υπάρχει μια σταθερή σχέση μεταξύ του υπό πρόβλεψη μεγέθους (εξαρτημένη μεταβλητή) και ορισμένων παραμέτρων (ανεξάρτητη μεταβλητή) που το επηρεάζουν. Το πιο σημαντικό πλεονέκτημα των αιτιοκρατικών μεθόδων είναι ότι προσφέρουν στον χρήστη την δυνατότητα να προβλέψει την μελλοντική τιμή κάποιου μεγέθους, για διάφορους συνδυασμούς των μεταβλητών εισόδου. Έτσι έχοντας στην διάθεσή του διάφορα εναλλακτικά σενάρια μπορεί να καταλήξει ευκολότερα στην επιλογή της βέλτιστης επιλογής. Παρόλ' αυτά τα αιτιοκρατικά μοντέλα έχουν και μερικά σημαντικά μειονεκτήματα**.** Ένα από αυτά είναι ο μεγάλος όγκος δεδομένων που πρέπει να έχουμε στην κατοχή μας, αφού θα πρέπει να έχουμε ισάξια δεδομένα για την μεταβλητή που μελετάμε (μεταβλητή πρόβλεψης) και για κάθε επιπλέον μεταβλητή που βρίσκεται στο περιβάλλον της και την επηρεάζει (ανεξάρτητες μεταβλητές). Στις αιτιοκρατικές μεθόδους ανήκουν οι μέθοδοι παλινδρόμησης (απλή γραμμική παλινδρόμηση, πολλαπλή γραμμική παλινδρόμηση) και οι οικονομετρικές μέθοδοι.

2. Ποιοτικές μέθοδοι

Είναι οι μέθοδοι πρόβλεψης, που βασίζονται στην κρίση, δηλαδή σε υποκειμενικές εκτιμήσεις και στην εμπειρία. Η ανάγκη, βέβαια για επιτυχημένες προβλέψεις με μικρά σφάλματα τις καθιστούν ολοένα και λιγότερο προτιμητέες. Συνεπώς, χρησιμοποιούνται στην πλειοψηφία τους για προβλέψεις σε μεσοπρόθεσμο ή μακροπρόθεσμο σχεδιασμό ή ακόμη καλύτερα σε συνδυασμό με μια ποσοτική μέθοδο.

Οι μέθοδοι πρόβλεψης επίσης μπορούν να χωριστούν σε κατηγορίες ανάλογα με τον ορίζοντα πρόβλεψης

1. Βραχυπρόθεσμες, συνήθης μέθοδοι Προεκβολής
2. Μεσοπρόθεσμες, χρονικός ορίζοντας 6-12 μήνες συνήθως Προεκβολές ή Αιτιακές

3. Μακροπρόθεσμες, ορίζοντας ετών, συνήθως Αιτιακές ή Ποιοτικές

Νευρωνικά Δίκτυα

Ένα τεχνητό νευρωνικό δίκτυο μπορεί να βοηθήσει σε περιπτώσει που υπάρχουν μη γραμμικές διαδικασίες, όπου η συσχέτιση δεν είναι γνωστή εξ΄ αρχής και άρα είναι δύσκολο να επιτευχθεί βέλτιστη προσαρμογή. Η κύρια ιδέα των νευρωνικών δικτύων είναι το φιλτράρισμα των εισόδων, που αποτελούν και τις ανεξάρτητες μεταβλητές, μέσω ενός ή περισσοτέρων κρυφών επιπέδων, πρωτού παραχθεί η τελική έξοδος. Τα νευρωνικά δίκτυα έχουν εφαρμοσθεί σε διάφορες πτυχές των προβλέψεων, από απευθείας παραγωγή προβλέψεων έως την βελτιστοποίηση συγκεκριμένων παραμέτρων.

2.3 Βασικά βήματα της διαδικασίας της πρόβλεψης

Εδώ περιγράφονται εν συντομία τα πέντε βασικά βήματα που είναι απαραίτητα σε μια διαδικασία παραγωγής και αξιολόγησης προβλέψεων, σύμφωνα με τους Μακρυδάκη, Wheelright και Hyndman (1998).

1. Καθορισμός του προβλήματος. Πολλές φορές αυτό το βήμα είναι πολύ δύσκολο καθώς πρέπει να καθοριστεί τι πρέπει να προβλεφθεί , από ποιους και πως .
2. Συλλογή των δεδομένων. Πρέπει να συλλεχθούν και να συντηρηθούν σωστά τα δεδομένα. Επίσης πέρα από τα μετρήσιμα αριθμητικά δεδομένα πρέπει να αξιοποιηθεί και η γνώση των εργαζομένων στην επιχείρηση και η κριτική του εμπειρία για την συγκεκριμένη χρονική στιγμή.
3. Προετοιμασία των χρονοσειρών. Εδώ πρέπει να αποκτηθεί μια ολοκληρωμένη αίσθηση των διαθέσιμων δεδομένων ώστε να γίνουν οι απαραίτητες προσαρμογές, αν χρειάζεται (π.χ. εποχικότητα, τάση) πριν εισαχθούν στο μοντέλο πρόβλεψης
4. Επιλογή μεθόδων πρόβλεψης. Εδώ γίνεται από τους ειδικούς η επιλογή του μοντέλου αλλά και η επιλογή των παραμέτρων, κυρίως βάση ιστορικών δεδομένων.
5. Χρήση και αξιολόγηση των μοντέλων πρόβλεψης. Στο τελευταίο στάδιο δείχνει το πόσο ικανοποιητικές είναι οι παραχθείσες προβλέψεις μέσω εξειδικευμένων δεικτών, κάποιοι εκ των οποίων παρουσιάζονται παρακάτω.

2.4 Μέθοδοι στατιστικών προβλέψεων

- Naïve

Αποτελεί την πιο απλή στατιστική μέθοδο. Δεν παράγει ακριβείς προβλέψεις, αλλά πολλές φορές χρησιμοποιείται ως σημείο αναφοράς για άλλες πιο πολύπλοκες μεθόδους. Η πρόβλεψη που προκύπτει από τη μέθοδο naive για μια χρονική στιγμή t είναι ίσε με την πραγματική παρατήρηση της ακριβώς προηγούμενης περιόδου t-1:

$$F_t = Y_{t-1}$$

- Μέθοδοι Εκθετικές εξομάλυνσης

Οι συγκριμένες μέθοδοι αναπτύχθηκαν την δεκαετία του 1950. Οι μέθοδοι εξομάλυνσης είναι κατάλληλες κυρίως για βραχυχρόνιες και μεσοπρόθεσμες προβλέψεις μεγάλου όγκου χρονοσειρών. Αποδίδουν καλύτερα σε δεδομένα με στασιμότητα ή μικρό ρυθμό ανάπτυξης.

- Μοντέλα παλινδρόμησης

Η ανάλυση της παλινδρόμησης μελετά τη σχέση μεταξύ μια εξαρτημένης μεταβλητής με συγκεκριμένες ανεξάρτητες μεταβλητές. Τέτοια μοντέλα θεωρούνται κατάλληλα για παραγωγή μακροπρόθεσμων προβλέψεων.

- Μέθοδος Theta

Το μοντέλο πρόβλεψης Theta (Asimakopoulos και Nikolopoulos, 2000) βασίζεται στην τροποποίηση των τοπικών καμπυλοτήτων της χρονοσειράς. Η αρχική χρονοσειρά αποσυντίθεται σε δύο ή περισσότερες γραμμές Theta. Κάθε μια από αυτές προεκτείνεται ξεχωριστά και οι προβλέψεις τους συνδυάζονται.

2.5 Δείκτες Ακρίβειας

Σκοπός είναι η μέτρηση της ακρίβειας των προβλέψεων μέσω στατιστικών δεικτών. Για το λόγο αυτό αρχικά ως σφάλμα ορίζουμε την διαφορά μεταξύ πραγματικής τιμής και πρόβλεψης για μια περίοδο:

$$e_i = Y_i - F_i$$

Παρακάτω παρουσιάζονται επιπλέον χρήσιμοι δείκτες (χάριν ευκολίας σφάλματα) που χρησιμοποιούνται ευρέως στις προβλέψεις.

- Μέσο Σφάλμα (Mean error) :

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n} Yi - Fi$$

Το μέσο σφάλμα υπολογίζεται από τον απλό προσημασμένο μέσο όρο των σφαλμάτων και εκφράζει ένα μέτρο συστηματικότητας του σφάλματος. Όσο η τιμή αυτού είναι κοντά στο μηδέν, τόσο τα σφάλματα είναι τυχαία και όχι συστηματικά. Αν ο δείκτης παίρνει θετικές τιμές δηλώνει απαισιοδοξία στις προβλέψεις, μιας και οι προβλέψεις είναι κατά μέσο όρο μικρότερες των πραγματικών τιμών. Από την άλλη αρνητικές τιμές του δείκτη δηλώνουν αισιοδοξία. Συχνά ο δείκτης αναφέρεται και ως bias.

- Μέσο Απόλυτο Σφάλμα (Mean Absolute error) :

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|Yi-Fi|$$

Το μέσο απόλυτο σφάλμα εκφράζει ένα μέτρο της ακρίβειας της πρόβλεψης έναντι των πραγματικών τιμών διατηρώντας τις μονάδες μέτρησης αρχικής χρονοσειράς. Δηλώνει ένα μέσο μέτρο της αστοχίας της πρόβλεψης χωρίς να δίνεται έμφαση στην κατεύθυνση της πρόβλεψης. Όσο μεγαλύτερη είναι η τιμή του δείκτη τόσο μικρότερη προκύπτει η ακρίβεια της μεθόδου που εφαρμόσθηκε.

- Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error) :

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(Yi-Fi)^2$$

Το μέσο τετραγωνικό σφάλμα όπως και το μέσο απόλυτο σφάλμα είναι ένα μέτρο ακρίβειας της πρόβλεψης το οποίο όμως δίνει πολύ μεγαλύτερο βάρος στα μεγάλα σφάλματα και μικρότερο βάρος στα μικρά σφάλματα (λόγω τετραγωνισμού).

- Ρίζα μέσου τετραγωνικού σφάλματος (Root Mean Squared Error) :

$$\text{RMSE}=\sqrt{MSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Yi-Fi)^2}$$

Η ρίζα μέσου τετραγωνικού σφάλματος υπολογίζεται άμεσα από την τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος. Έχει τις ίδιες ιδιότητες με αυτό ,αλλά είναι εκφρασμένο στις μονάδες της αρχικής χρονοσειράς.

- Μέσο απόλυτο ποσοστιαίο σφάλμα (Mean Absolute Percentage Error) :

$$\text{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{Yi-Fi}{Yi}\right|\times 100\%$$

Ορισμένες φορές είναι πιο χρήσιμος ο υπολογισμός των σφαλμάτων πρόβλεψης σε καθαρά ποσοστιαία μορφή. Αυτό για παράδειγμα θα ήταν χρήσιμο όταν θέλουμε να συγκρίνουμε την ακρίβεια μιας μεθόδου πρόβλεψης που έχει εφαρμοστεί σε παραπάνω από μία χρονοσειρές όπου η κάθε μία έχει διαφορετικό επίπεδο μέσης τιμής. Επίσης είναι πολύ χρήσιμο όταν οι πραγματικές τιμές είναι ιδιαίτερα υψηλές. Το μέσο απόλυτο ποσοστιαίο σφάλμα είναι εκφρασμένο επί τοις εκατό και λαμβάνει τιμές μεγαλύτερες ή ίσες του μηδενός με τις μικρότερες τιμές να υποδηλώνουν και καλύτερη απόδοση της μεθόδου πρόβλεψης.

- Συμμετρικό μέσο απόλυτο ποσοστιαίο σφάλμα (Symmetric Mean Absolute Percentage Error) :

$$\text{sMAPE}=\sum_{i=1}^{n}\left|\frac{2(Yi-Fi)}{Yi+Fi}\right|\times 100\%$$

Όπως φαίνεται από τα παραπάνω το sMAPE είναι μια παραλλαγή του MAPE. Το sMAPE, όπως φαίνεται και από τον τύπο του, αποκτά και πάνω όριο και πλέον μπορεί να πάρει τιμές στο διάστημα [0%, 200%]. Όμως, με αυτή την αλλαγή το sMAPE έχει πρόβλημα καθώς είναι πλέον όχι και τόσο συμμετρικός δείκτης, παρόλο το όνομα του. Πιο συγκεκριμένα, οι αισιόδοξες και οι απαισιόδοξες προβλέψεις δεν έχουν την ίδια μεταχείριση. Αυτό φαίνεται και από το παρακάτω παράδειγμα:

- Αισιόδοξη πρόβλεψη: Y_t=100 και F_t=110 έχουμε sMAPE =4,76%
- Απαισιόδοξη πρόβλεψη: Y_t=100 και F_t=90 έχουμε sMAPE=5,26%

Στην παρούσα διπλωματική εργασία, για την αξιολόγηση των προβλέψεων έγινε χρήση αποκλειστικά του sMAPE.

ΚΕΦΑΛΑΙΟ 3 Νευρωνικά Δίκτυα

3.1 Εισαγωγή στην μηχανική μάθηση

Ένα φυσικό ή τεχνητό σύστημα επεξεργασίας πληροφορίας, συμπεριλαμβανομένων εκείνων με δυνατότητες αντίληψης, μάθησης, συλλογισμού, λήψης απόφασης, επικοινωνίας και δράσης ονομάζεται γνωστικό σύστημα (cognitive system). Η έννοια της μάθησης σε ένα γνωστικό σύστημα όπως γίνεται αντιληπτή στην καθημερινή ζωή μπορεί να συνδεθεί με δύο βασικές ιδιότητες:

- Την ικανότητα του στην πρόσκτηση γνώσης κατά την αλληλεπίδραση του με το περιβάλλον, μέσα στο οποία δραστηριοποιείται και
- Την ικανότητα του να βελτιώνει με την επανάληψη τον τρόπο με τον οποίο εκτελεί μια ενέργεια (και συνεπώς την απόδοση του).

Ορίζοντας, πιο συγκεκριμένα την μηχανική μάθηση μπορούμε να την ορίσουμε με βάση την ανθρώπινη εμπειρία. Ο άνθρωπος προσπαθεί να κατανοήσει το περιβάλλον του παρατηρώντας το και δημιουργώντας μια απλοποιημένη, άρα αφαιρετική εκδοχή του, το λεγόμενο μοντέλο. Η δημιουργία ενός τέτοιου μοντέλου ονομάζεται επαγωγική μάθηση ενώ η διαδικασία γενικότερα ονομάζεται επαγωγή (induction). Επιπρόσθετα ο άνθρωπος μπορεί να οργανώνει και να συσχετίζει τις εμπειρίες και τις παραστάσεις του δημιουργώντας νέες δομές που ονομάζονται πρότυπα (patterns). Η δημιουργία μοντέλων ή προτύπων απο ένα σύνολο δεδομένων, από ενα υπολογιστικό σύστημα ονομάζεται μηχανική μάθηση (machine learning).

Στο πέρασμα του χρόνου έχουν διατυπωθεί διάφοροι ορισμοί για την μηχανική μάθηση, όπως των:

- Carbonell (1987), "…η μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης, νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης".
- Mitchell (1997), "Ένα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία E σε σχέση με μια κατηγορία εργασιών T και μια μετρική απόδοης P, αν η απόδοση του σε εργασίες της T, όπως μετριούνται από την P, βελτιώνονται με την εμπειρία E.

3.2 Είδη μηχανικής μάθησης

- Μάθηση με επίβλεψη (supervised learning) ή μάθηση με παραδείγματα (learning from examples).

Σε αυτό το είδος μάθησης το σύστημα καλείται να μάθει μια έννοια ή μια συνάρτηση από ένα σύνολο δεδομένων, η οποία αποτελεί περιγραφή του μοντέλου. Ονομάστηκε έτσι καθώς θεωρείται ότι υπάρχει τρόπος κάποιος «επιβλέπων» ο οποίος παρέχει τη σωστή τιμή εξόδου της συνάρτησης, για τα δεδομένα που εξετάζονται

- Μάθηση χωρίς επίβλεψη (unsupervised learning) ή μάθηση από παρατήρηση (learning from observation).

Σε αυτό το είδος μάθησης το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποιά είναι.

Τα μοντέλα που περιγράφηκαν παραπάνω μπορούν είτε να χρησιμοποιηθούν ως μοντέλα πρόβλεψης (predictive models) μιας και προβλέπουν την τιμή μιας μεταβλητής είτε να δώσουν ποιοτικές πληροφορίας για τα δεδομένα. Όμως τα πρότυπα έχουν τοπικό χαρακτήρα, δηλαδή το καθένα περιγράφει ένα μέρος δεδομένων και χαρακτηρίζονται ως πρότυπα πληροφόρησης (informative patterns) επειδή περιγράφουν συσχετίσεις μεταξύ των δεδομένων.

Όσον αφορά την μάθηση με επίβλεψη πρέπει να τονιστεί ότι υπάρχουν δύο μοντέλα:

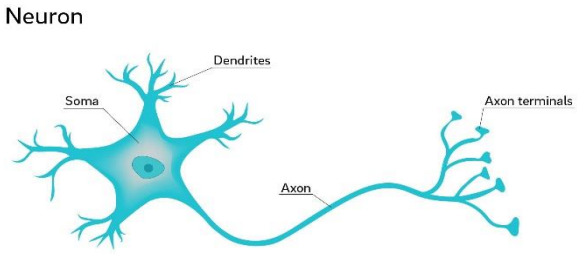
- Τα μοντέλα κατάταξης (classifiers)
- Τα μοντέλα παλινδρόμησης (regression)

  Τα μοντέλα παλινδρόμησης αντιστοιχούν τον χώρο εισόδου σε ένα πραγματικό - αποτιμημένο τομέα δηλαδή έχουμε συνεχείς τιμές ενώ οι ταξινομητές χαρτογραφούν τον χώρο εισόδου σε προκαθορισμένες κατηγορίες ή κλάσεις. Για παράδειγμα, στην παρούσα διπλωματική έχουμε μοντέλο παλινδρόμησής ενώ αν, ως έξοδοι στα πειράματα μας είχαμε απλά εντολή για αγορά ή πώληση μετοχών θα είχαμε μοντέλο κατάταξης.

3.2 Λειτουργία των Νευρωνικών Δικτύων

Τα νευρωνικά δίκτυα είναι μια ιδιαίτερη προσέγγιση στη δημιουργία συστημάτων τεχνητής νοημοσύνης καθώς δεν αναπαριστούν ρητά τη γνώση αλλά βασίζονται σε βιολογικά πρότυπα και μιμούνται τις διαδικασίες του ανθρώπινου εγκεφάλου.

Η ικανότητα του ανθρώπου να σκέφτεται, να έχει μνήμη και να επιλύει διάφορα προβλήματα εντοπίζεται στον εγκέφαλο του, με δοκιμή μονάδα αυτού τον νευρώνα (neuron)



Εικόνα 3. 1 Βιολογικός Νευρώνας

Ένας τυπικός νευρώνας (Σχήμα 3.1) αποτελείται απο το σώμα (soma ή body) που είναι και ο πυρήνας του, τους δενδρίτες (dendrites) και τον άξονα (axon). Μέσω των δενδριτών λαμβάνει ο νευρώνας σήματα από τους γειτονικούς νευρώνες. Ο άξονας είναι η έξοδος του νευρώνα και το μέσο σύνδεσης του με τους υπόλοιπους νευρώνες. Σε κάθε δενδρίτη υπάρχει ένα απειροελάχιστο κενό που λέγεται σύναψη (synapse). Η συνάψεις μέσω χημικών διαδικασιών επιταχύνουν ή επιβραδύνουν τη ροή ηλεκτρικών σημάτων στο σώμα του νευρώνα. Η ικανότητα μάθησης και μνήμης που παρουσιάζει ο εγκέφαλος οφείλεται στην ικανότητα των συνάψεων να μεταβάλουν την αγωγιμότητα τους. Τα ηλεκτρικά σήματα που εισέρχονται στο σώμα των νευρώνων μέσω των δενδριτών συνδυάζονται και αν το αποτέλεσμα ξεπερνά κάποια τιμή κατωφλίου το σήμα διαδίδεται με τη βοήθεια του άξονα προς τους άλλους νευρώνες.

Ο εγκέφαλος ενός τυπικού ανθρώπου αποτελείται περίπου απο 100 δισεκατομμύρια νευρώνες και καθένας νευρώνας συνδέεται με 1000 άλλους δηλαδή υπάρχουν περίπου 100 τρισεκατομμύρια συνάψεις που καθορίζουν τη λειτουργία του εγκεφάλου. Τα μοντέλα προσομοίωσης δεν αντιγράφουν, λόγω του μεγέθους, τους νευρώνες του εγκεφάλου αλλά περιορίζονται στο πολύ ένα εκατομμύριο συνάψεις. Ο εγκέφαλος είναι ικανός να λαμβάνει πολύπλοκες αποφάσεις εκπληκτικά γρήγορα επομένως έχει τεράστια υπολογιστική ικανότητα ως ένα παράλληλο και κατανεμημένο σύστημα. Αυτά του τα πλεονεκτήματα οδήγησαν την επιστήμη στην επιθυμία μοντελοποίησης του.

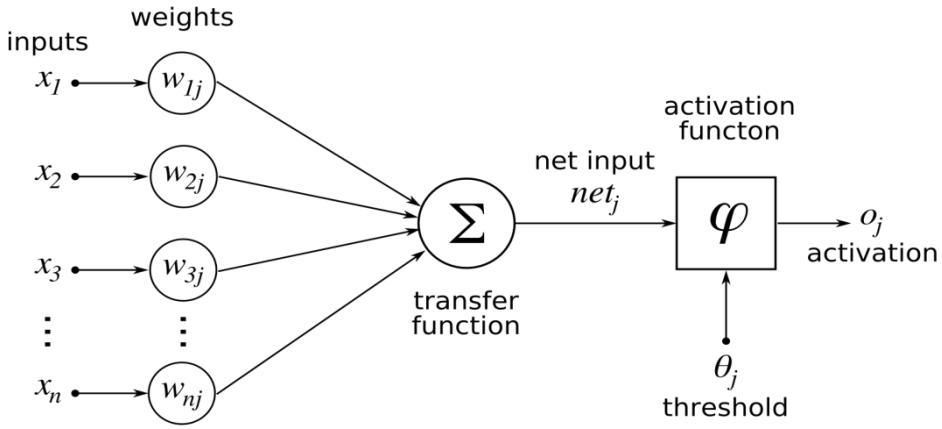
3.3 Μοντέλο τεχνητού νευρώνα

Ο τεχνητός νευρώνας (artificial neuron) είναι ένα υπολογιστικό μοντέλο τα μέρη του οποίου αντιστοιχούν σε αυτά του βιολογικού νευρώνα. Ο τεχνητός νευρώνας δέχεται κάποια σήματα εισόδου x_1, x_2, …, x_n . Κάθε τέτοιο σήμα εισόδου μεταβάλλεται από μια τιμή βάρους w_i (weight) ο ρόλος της οποίας είναι αντίστοιχος του ρόλου της

σύναψης στον βιολογικό νευρώνα. Η τιμή βάρους μπορεί να είναι είτε θετική είτε αρνητική, σε αντιστοίχιση με την επιταχυντική ή επιβραδυντική λειτουργία της σύναψης.

Το κύριο μέρος του νευρώνα, δηλαδή το σώμα, χωρίζεται σε δύο μέρη. Το ένα είναι ο αθροιστής (sum) οποίος προσθέτει τα επηρεασμένα από τα βάρη σήματα εισόδου παράγοντας της ποσότητα S και την συνάρτηση ενεργοποίησης (activation function), ένα είδος φίλτρου το οποίο διαμορφώνει την τελική τιμή της εξόδου y, σε συνάρτηση με την ποσότητα S και την τιμή κατωφλίου της συνάρτησης ενεργοποίησης. Ο νευρώνας μπορεί να έχει πολλές εξόδους αλλά όλες θα έχουν την ίδια τιμή.

Επιπρόσθετα μπορούμε να θεωρήσουμε πως εκτός από τα εισερχόμενα σήματα και τα αντίστοιχα βάρη, ο νευρώνας έχει και κάποιο βάρος w_0 , το οποίο ονομάζεται πόλωση (bias) ή παράγοντας προδιάθεσης του νευρώνα. Η διαφορά αυτού του βάρους από τα υπόλοιπα συνίσταται στο ότι επιδρά σε μια τιμή εισόδου $x_0 = 1$. Η πόλωση δεν είναι στοιχείο του εσωτερικού του νευρώνα αλλά για εξωτερικό ερέθισμα που προστίθεται μαζί με τα υπόλοιπα εισερχόμενα σήματα. Αρκετές φορές χρησιμοποιείται για να καθοριστεί έμμεσα και δυναμικά η θέση της συνάρτησης ενεργοποίηση στο καρτεσιανό επίπεδο.



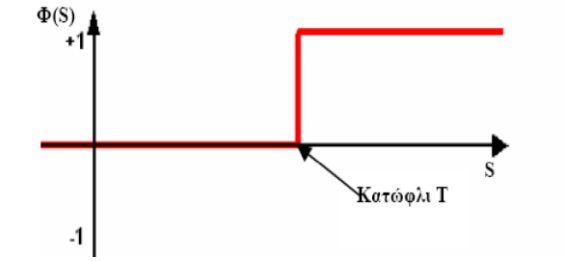
Εικόνα 3. 2 Τεχνητός Νευρώνας

3.4 Συναρτήσεις ενεργοποίησης

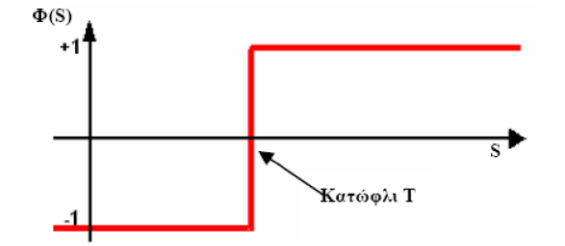
Παρακάτω παρουσιάζονται τρεις τυπικές συναρτήσεις ενεργοποίησης:

- Η βηματική συνάρτηση (step function)(Σχήμα 3.3), η οποία δίνει στην έξοδο αποτέλεσμα 1 μόνο αν η τιμή που υπολογίζει ο αθροιστής είναι μεγαλύτερη της τιμής κατωφλίου T.
- Η συνάρτηση προσήμου (sign function)(Σχήμα 3.4), η οποία δίνει στην έξοδο αρνητική ή θετική πληροφορία αν η τιμή του αθροιστή είναι μικρότερη ή μεγαλύτερη από μια τιμή κατωφλίου T

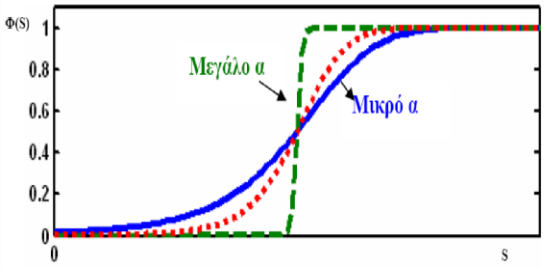
- Η λογιστική συνάρτηση (logistic function)(Σχήμα 3.5) η οποία εκφράζεται από τη σχέση $f(S) = 1/(1 + e^{-aS})$ με το α να είναι ένας συντελεστής ρύθμισης της ταχύτητας μετάβασης μεταξύ δύο ασυμπτωτικών τιμών.



Εικόνα 3. 3 Η βηματική συνάρτηση



Εικόνα 3. 4 Η συνάρτηση προσήμου

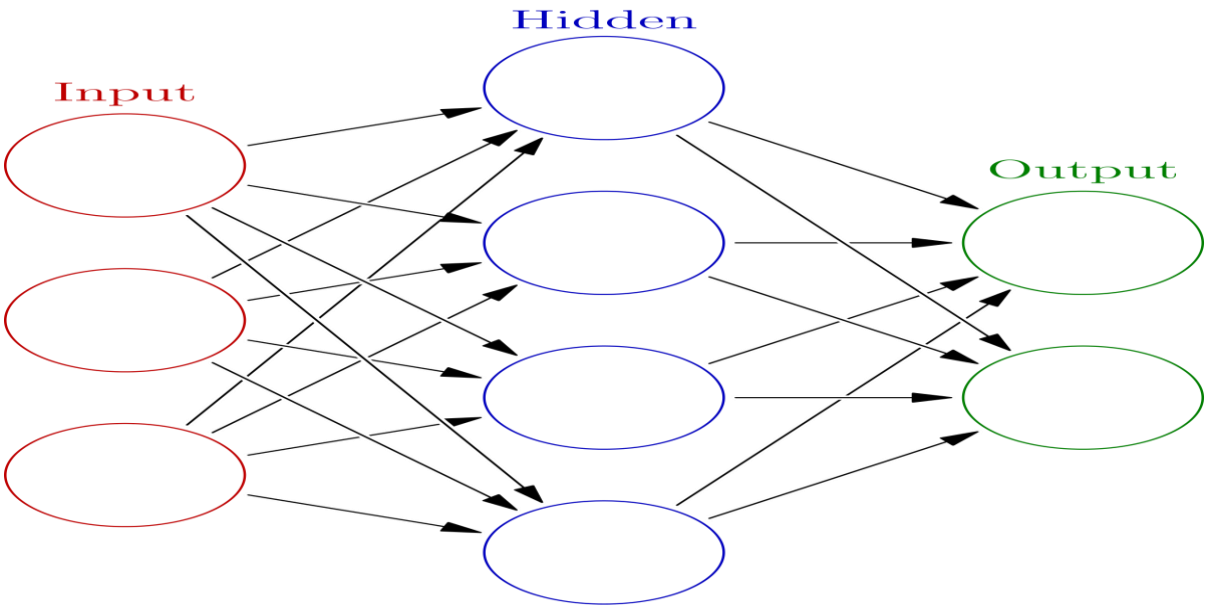


Εικόνα 3. 5 Η λογιστική συνάρτηση

3.5 Λειτουργία νευρωνικών δικτύων

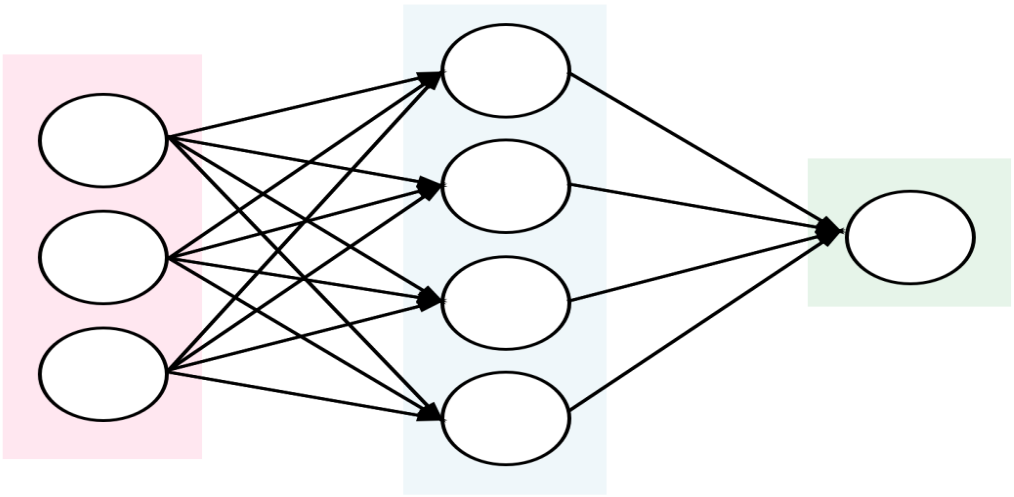
Τα τεχνητά νευρωνικά δίκτυα (artificial neural networks) ή πιο απλά ΤΝΔ, χαρακτηρίζονται ως συστήματα επεξεργασίας δεδομένων που αποτελούνται από ένα πλήθος τεχνητών νευρώνων οργανωμένων σε δομές παρόμοιες με αυτές του ανθρώπινου εγκεφάλου. Συνήθως οι τεχνητοί νευρώνες είναι οργανωμένοι σε μία σειρά από στρώματα ή επίπεδα (layers). Το πρώτο από αυτά τα επίπεδα ονομάζεται επίπεδο εισόδου (input layer) και χρησιμοποιείται για την εισαγωγή των δεδομένων. Τα στοιχεία, που συνιστούν το επίπεδο εισόδου, δεν είναι ουσιαστικά νευρώνες, καθώς δεν εκτελούν κάποιον υπολογισμό (δεν έχουν ούτε βάρη εισόδου, ούτε συναρτήσεις ενεργοποίησης). Στη συνέχεια μπορούν να υπάρχουν, προαιρετικά, ένα

ή περισσότερα ενδιάμεσα ή κρυφά επίπεδα (hidden layers). Τέλος ακολουθεί ένα επίπεδο εξόδου (output layer).

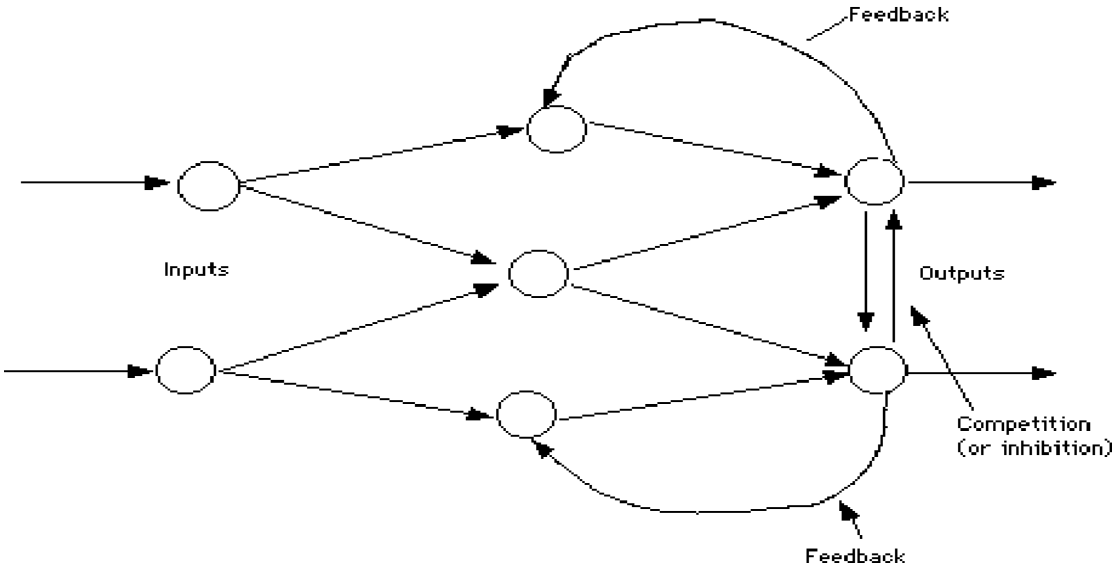


Εικόνα 3. 6 Τεχνητό Νευρωνικό Δίκτυο

Οι νευρώνες στα ΤΝΔ μπορεί να είναι είτε πλήρως είτε μερικώς συνδεδεμένοι. Πλήρως συνδεδεμένοι (fully connected) είναι εκείνοι οι οποίοι συνδέονται με όλους τους υπόλοιπους νευρώνες. Σε όλες τις άλλες περιπτώσεις οι νευρώνες είναι μερικώς συνδεδεμένοι (partially connected). Όταν δεν υπάρχουν συνδέσεις μεταξύ νευρώνων ενός επιπέδου και νευρώνων προηγούμενου επιπέδου (όταν δηλαδή η ροή πληροφορίας είναι μιας κατεύθυνσης) τα ΤΝΔ χαρακτηρίζονται ως δίκτυα με απλή (ή πρόσθια) τροφοδότηση (feedforward). Στην αντίθετη περίπτωση, καθώς και στην περίπτωση συνδέσεων μεταξύ νευρώνων του ίδιου επιπέδου, τα ΤΝΔ χαρακτηρίζονται ως δίκτυα με ανατροφοδότηση (feedback ή recurrent). Ο τύπος του δικτύου με ανατροφοδότηση διαφέρει από τον τύπο της απλούς τροφοδότησης στο ότι περιλαμβάνει ένα βρόχο ανάδρασης, όπου κάθε νευρώνας τροφοδοτεί το σήμα της εξόδου του στις εισόδους όλων των άλλων νευρώνων.



Εικόνα 3. 7 ΤΝΔ απλής τροφοδότησης



Εικόνα 3. 8 ΤΝΔ με ανατροφοδότηση

3.6 Βασικές ιδιότητες των Νευρωνικών Δικτύων

Οι παρακάτω ιδιότητες είναι άρρητα συνδεδεμένες με τα ΤΝΔ:

- Η ικανότητά τους να μαθαίνουν μέσω παραδείγματα (learn by example)

Τα ΤΝΔ, παρόλο που δεν είναι να μόνο συστήματα με ικανότητα μάθησης μέσω παραδειγμάτων, διακρίνονται για της ικανότητα τους να οργανώνουν την πληροφορία των δεδομένων εισόδου σε χρήσιμες μορφές, οι οποίες αποτελούν στην ουσία, ένα μοντέλο που αναπαριστά της σχέση μεταξύ της εισόδου και της εξόδου.

- Η μεγάλη ανοχή του σε σφάλματα (fault-tolerant)

Αυτό σημαίνει πως η καταστροφή ή η κακή λειτουργία ενός νευρώνα δεν διαταράσσει σημαντικά την αποτελεσματικότητα τους καθώς η πληροφορία που περιέχουν δεν είναι εντοπισμένη σε ένα μόνο σημείο αλλά σε ολόκληρο το δίκτυο.

- Η ικανότητα τους για αναγνώριση προτύπων (pattern recognition)

Αυτή τους η ιδιότητα προκύπτει από την μη επιρροή του από ελλιπή δεδομένα και θόρυβο. Όταν ένα ΤΝΔ εκπαιδευτεί στο να αναγνωρίζει συνθήκες και καταστάσεις, απαιτείται μόνο ένας κύκλος λειτουργίας για να προσδιοριστεί μια συγκεκριμένη κατάσταση.

- Η δυνατότητα τους ως κατανεμημένη μνήμη (distributed memory) και ως μνήμη συσχέτισης (associative memory)

Τα ΤΝΔ μπορούν να χαρακτηριστούν ως κατανεμημένη μνήμη καθώς η κωδικοποίηση που δημιουργούν είναι κατανεμημένη σε όλα τα βάρη της συνδεσμολογίας τους. Για το ίδιο λόγο χαρακτηρίζονται και ως μνήμη συσχέτισης καθώς η πληροφορία αποθηκεύεται μέσω συσχετίσεων που δημιουργούνται από τα δεδομένα εκπαίδευσης.

3.7 Μάθηση και ανάκλαση των ΤΝΔ

Τα ΤΝΔ έχουν δύο βασικές λειτουργίες: την μάθηση και την ανάκλαση. Μάθηση (learning) είναι η διαδικασία της τροποποίησης της τιμής των βαρών του δικτύου, ώστε δοθέντος συγκεκριμένου διανύσματος εισόδου να παραχθεί συγκεκριμένο διάνυσμα εξόδου. Η διαδικασία αυτή ονομάζεται επίσης εκπαίδευση (train) του ΤΝΔ. Ανάκληση (recall) είναι η διαδικασία του υπολογισμού ενός διανύσματος εξόδου για συγκεκριμένο διάνυσμα εισόδου και τιμές βαρών.

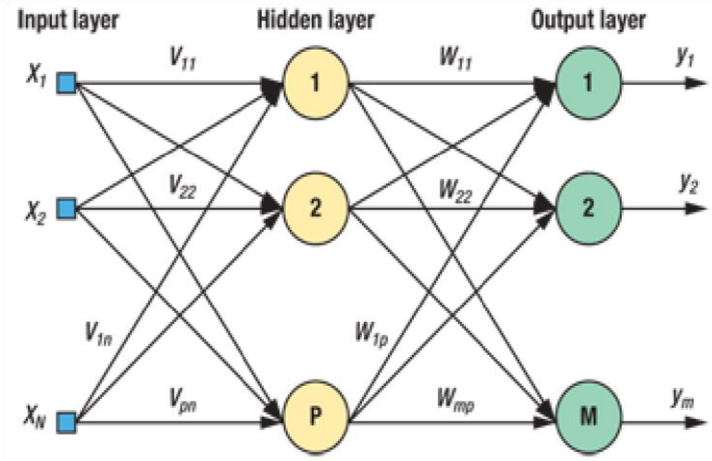
Η διαδικασία της μάθησης μπορεί να διακριθεί σε τρία είδη:

- Στη μάθηση με επίβλεψη (supervised learning) δίνονται στο δίκτο ζευγάρια διανυσμάτων εισόδου-επιθυμητής εξόδου και αυτό παράγει, με την τρέχουσα κατάσταση βαρών, μια έξοδο που αρχικά διαφέρει από την επιθυμητή. Αυτή η διαφορά ονομάζεται σφάλμα (error) και βάσει αυτής καθώς και του αλγορίθμου εκπαίδευσης γίνεται η αναπροσαρμογή των βαρών.
- Στη βαθμολογημένη μάθηση (graded learning) η έξοδος χαρακτηρίζεται ως καλή ή κακή με βάση μια αριθμητική κλίματα και τα βάρη επαναπροσαρμόζονται με βάση αυτόν τον χαρακτηρισμό.
- Στη μάθηση χωρίς επίβλεψη (unsupervised learning) η απόκριση του δικτύου βασίζεται στην ικανότητα του να αυτό-οργανώνεται με βάση τα διανύσματα εισόδου καθώς δεν υπάρχουν αντίστοιχα διανύσματα εξόδου. Στην πράξη τέτοια δίκτυα κατηγοριοποιούν τα δεδομένα εισόδου.

Με κριτήριο την αρχιτεκτονική τους τα νευρωνικά δίκτυα μπορούν χωριστούν στις εξής δύο κατηγορίες:

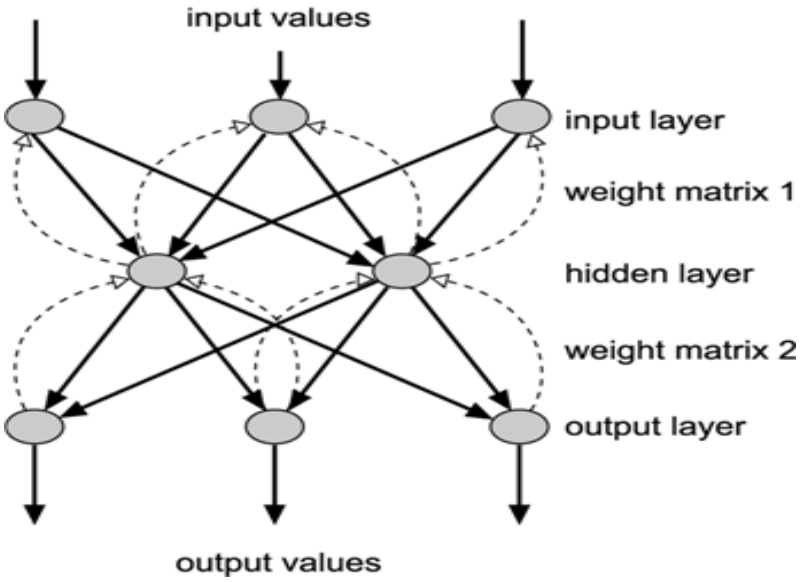
- Νευρωνικά δίκτυα πρόσθιας τροφοδότησης

Τα ΤΝΔ πρόσθιας τροφοδότησης (feedforward) είναι η πιο απλή μορφή νευρωνικών δικτύων και το όνομα τους οφείλεται στο ότι η ροή της πληροφορίας μέσα στο δίκτυο είναι μονής κατεύθυνσης. Σε αυτά υπάρχει ένα επίπεδο εισόδου, ένα επίπεδο εξόδου και προαιρετικά, ένα η περισσότερα ενδιάμεσα κρυφά επίπεδα.



Εικόνα 3. 9 ΤΝΔ με Feedforward

Στα ΤΝΔ με ανατροφοδότηση η ροή της πληροφορίας μπορεί να κινηθεί και προς τις δύο κατευθύνσεις, δηλαδή και προς την έξοδο και προς την είσοδο. Αυτού του τύπου τα δίκτυα είναι αρκετά περίπλοκα και η ανατροφοδότηση τους πολύ ισχυρή. Σε αυτή την περίπτωση ουσιαστικά τα κρυφά επίπεδα έχουν την δυνατότητα να κατασκευάσουν τα δικά τους δεδομένα εισόδου.



Εικόνα 3. 10 ΤΝΔ με Feedback

Η διαδικασία της μάθησης

Όπως αναφέρθηκε παραπάνω τα ΤΝΔ μπορούν να ταξινομηθούν ανάλογα με τον τρόπο μάθησης. Σκοπός αυτής της υποενότητας είναι να παρουσιαστεί αναλυτικά η διαδικασία της μάθησης.

Η όλη διαδικασία είναι κυκλική και μπορεί να χωριστεί στα εξής στάδια:

1. Διέγερση του ΤΝΔ από το εξωτερικό του περιβάλλον
2. Το ΤΝΔ κάνει τις αναγκαίες εσωτερικές αλλαγές εξαιτίας αυτής της διέγερσης
3. Μετά το πέρας των αλλαγών το ΤΝΔ απαντά στο περιβάλλον του ανάλογα.

Οι Mendel και McLaren ορίζουν της διαδικασία της εκπαίδευσης πιο φορμαλιστικά ως εξής: Η εκπαίδευση είναι η διαδικασία κατά την οποία αλλάζουν οι ελεύθερες μεταβλητές του Τεχνητού Νευρωνικού Δικτύου μέσω μιας διαρκούς κατάστασης διέγερσης από το περιβάλλον μέσα στο οποίο βρίσκεται. Το είδος της μάθησης εξαρτάται από την μέθοδο με την οποία γίνεται η αλλαγή των μεταβλητών .

3.8 Το Perceptron

Το perceptron είναι μια απλή τοπολογία δικτύου πρόσθιας τροφοδότησης χωρίς κρυφά επίπεδα. Προτάθηκε απο τον Rosenblatt το 1958, ως ένας μηχανισμός που μπορεί να εκπαιδευτεί στην κατηγοριοποίηση προτύπων και σε διάφορες παραλλαγές. Η μάθηση στο perceptron με ένα νευρώνα που ονομάζεται στοιχειώδες είναι καθοδηγούμενο απο το σφάλμα (error driven) και συνίσταται στον

υπολογισμών των κατάλληλων βαρών w_i ώστε να παραχθεί η επιθυμητή έξοδος. Πρόκειται δηλαδή για μια απλή μορφή μάθησης με επίβλεψη. Ο αλγόριθμος μεταβολής των βαρών έχει ως εξής:

Μέχρις ώτου ικανοποιηθεί η συνθήκη τερματισμού της εκπαίδευσης εκτέλεσε:

Για κάθε ζευγάρι εισόδου x και επιθυμητής εξόδου t απο το σύνολο εκπαίδευσης

1. Υπολόγισε την έξοδο t
2. Αν y=t τότε δε γίνεται καμία μεταβολή των βαρών
3. Αν y≠t τότε μετέβαλε τα βάρη των ενεργών γραμμών εισόδου (αυτών που έχουν σήμα ≠ 0) κατά την ποσότητα $\Delta w=d(t-y)x$ έτσι ώστε το y να πλησιάσει το t.

Γενικά στον αλγόριθμο, το d λέγεται ρυθμός μάθησης (learning rate), με τιμή συνήθως απο 0 έως 1 και καθορίζει τον ρυθμό μεταβολής των βαρών.

Αποδεικνύεται ότι αν υπάρχει ένα διάνυσμα βαρών W που παράγει την επιθυμητή έξοδο για όλα τα διανύσματα εκπαίδευσης, τότε ξεκινώντας από ένα τυχαίο διάνυσμα βαρών W_0 και μετά από πεπερασμένο αριθμό βημάτων ο αλγόριθμος perceptron θα συγκλίνει σε κάποιο διάνυσμα βαρών W^* το οποίο επίσης θα παράγει την επιθυμητή έξοδο για όλα τα διάνυσμα εκπαίδευσης. Δηλαδή ο αλγόριθμος συγκλίνει για κάθε πρόβλημα που μπορεί να αναπαρασταθεί έτσι. Η χαρακτηριστική ιδιότητα αυτής της κατηγορίας προβλημάτων ονομάζεται γραμμική διαχωρισιμότητα.

<u>Κανόνας Δέλτα</u>

Ο κανόνας δέλτα (delta rule) που αναπτύχθηκε από τους Widrow και Hoff τη δεκαετίατου ’60 αποτελεί γενίκευση του κανόνα εκπαίδευσης του perceptron καθώς η μάθηση είναι και εδώ καθοδηγούμενη από το σφάλμα. Ο συγκεκριμένος αλγόριθμος βασίζεται στην συνεχή μεταβολή των συντελεστών βάρους με σκοπό την ελαχιστοποίηση της διαφοράς μεταξύ των επιθυμητών αποτελεσμάτων εξόδου και αυτών που υπολογίστηκαν από τον Τεχνητό Νευρωνικό Δίκτυο και συγκεκριμένα την ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος.

Ο κανόνας Δέλτα ονομάζεται και κανόνας της επικλινούς καθόδου (gradient decent rule) εξαιτίας του ότι ακολουθεί την αρνητική κλίση της επιφάνειας του σφάλματος, με κατεύθυνση προς το ελάχιστο της. Σε αυτόν τον κανόνα η μεταβολή της τιμής του βάρους w_i, εξαιτίας της εκπαίδευσης με ένα μόνο από τα διανύσματα εκπαίδευσης, δίνεται από τη σχέση : $\Delta w_i = w_i' - w_i = d(t\text{-input})x_i$ με x_i η επιμέρους είσοδος της

οποίας το βάρος αναπροσαρμόζουμε, input το συνολικό σήμα εισόδου του νευρώνα, t η επιθυμητή έξοδος, w_i' το νέο βάρος και w_i η παλιά τιμή του βάρους. Η σταθερά d παίζει καθοριστικό ρόλο στην απόδοση του κανόνα και επηρεάζει την ταχύτητα σύγκλισης. Μεγάλες τιμές του d επιταχύνουν τη σύγκλιση προς το ελάχιστο σφάλμα αλλά αυξάνουν το κίνδυνο να προσπεραστεί το ελάχιστο με πιθανό αποτέλεσμα την παλινδρόμηση γύρω απο τις βέλτιστες τιμές βαρών. Αντίθετα μικρές τιμες του d αποτρέπουν το παραπάνω φαινόμενο αλλά απαιτούν περισσότερο χρόνο εκπαίδευσης.

Γενικευμένος κανόνας Δέλτα

Η παραπάνω σχέση για το Δw ισχύει και στην περίπτωση που αντί του συνολικού σήματος input εισόδου χρησιμοποιηθεί η πραγματική έξοδος y του νευρώνα, δηλαδή ληφθεί υπ' όψιν η δράση κάποιας βηματικής συνάρτησης ενεργοποίησης f. Τότε προκύπτει :

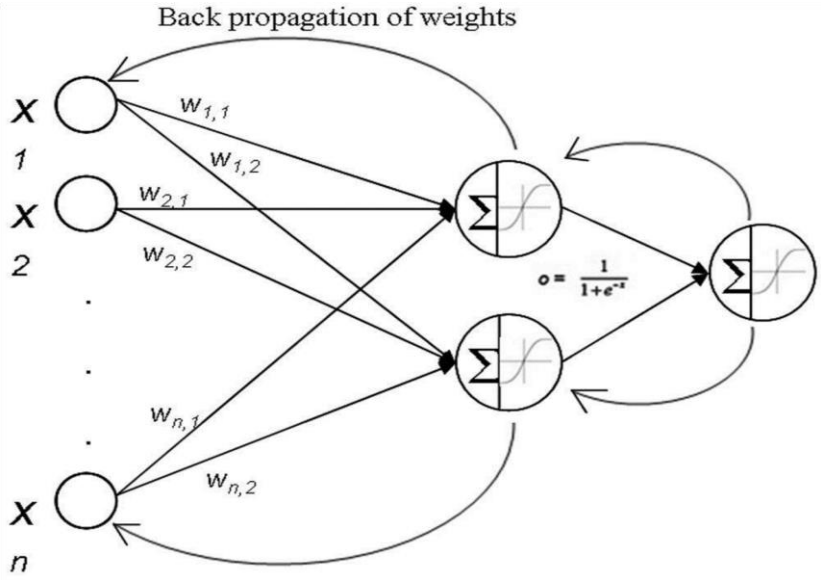
$\Delta w_i = w_i' - w_i = d(t-y)x_i$

Σε αυτή την περίπτωση προστίθεται ένας ακόμη όρος που σχετίζεται με την πρώτη παράγωγο f' της f. Ο όρος αυτός εκφυλίζεται. Η γενική αυτή περίπτωση ονομάζεται γενικευμένος κανόνας Δέλτα.

Τέλος η εκπαίδευση με τον κανόνα Δέλτα σταματά όταν το μέσο τετραγωνικό σφάλμα είναι μικρότερο απο κάποια επιθυμητη τιμή, διαφορετικά επαναλαμβάνεται, με ενδεχόμενη αλλαγή των διανυσμάτων εκπαίδευσης.

Ανάστροφη Μετάδοση Λάθους

Η ανάστροφη μετάδοση λάθους (back propagation) αποτελεί την πιο γνωστή μέθοδο εκπαίδευσης των νευρωνικών δικτύων πολλών επιπέδων. Ο αλγόριθμος αυτός βασίζεται στον γενικευμένο κανόνα Δέλτα ο οποίος επιτρέπει να καθοριστεί το ποσοστό του συνολικού σφάλματος που αντιστοιχεί στα βάρη του κάθε νευρώνα, ακόμη και αυτών που ανήκουν σε κρυφά επίπεδα, για τους οποίους η επιθυμητή έξοδος δεν είναι γνωστή. Με αυτόν τον τρόπο γίνεται δυνατό να υπολογίζονται οι διορθώσεις στα βάρη του κάθε νευρώνα ξεχωριστά.



Εικόνα 3. 11 Back Propagation

Η διαδικασία ενός κύκλου εκπαίδευσης του νευρωνικού δικτύου πολλών επιπέδου περιλαμβάνει δύο στάδια. Αρχικά εισάγονται τα δεδομένα από κάποιο διάνυσμα εκπαίδευσης, οπότε οι νευρώνες στο επίπεδο εισόδου παράγουν αποτέλεσμα το οποίο με τη σειρά του αποτελεί είσοδο για το επόμενο, κρυφό επίπεδο. Η διαδικασία αυτή επαναλαμβάνεται διαδοχικά για τα επόμενα επίπεδα, μέχρι το επίπεδο εξόδου και συνολικά ονομάζεται πρόσθιο πέρασμα (forward pass). Άρα η είσοδος ενός κρυφού νευρώνα j δίνεται από τη σχέση :

$$\text{input}_j = \sum_{i=1}^{n} VijXi$$

όπου v_{ij} είναι το βάρος της σύνδεσης μεταξύ των νευρώνων i,j, και το x_i το σήμα εισόδου του νευρώνα i.

Η έξοδος του κρυφού νευρώνα θα είναι:

$$z_j = f(\text{input}_j) = f(\sum_{i=1}^{n} VijXi)$$

και αποστέλλεται σε όλους τους νευρώνες του επόμενου σταδίου. Η συνάρτηση f είναι συνήθως μια σιγμοειδής συνάρτηση. Η παραπάνω σχέση εφαρμόζεται σε όλους τους νευρώνες εκτός από αυτούς του επιπέδου είσοδο.

Για τους νευρώνες του επιπέδου εξόδου θα έχουμε αντίστοιχα:

$$\text{Input}_k = \sum_{j=1}^{q} WjkZj \text{ και } y_k = f(\text{input}_k) = f(\sum_{i=1}^{q} WjkZj)$$

Πρέπει να ανφερθεί ότι τα ΤΝΔ με κρυφά επίπεδα έχουν καλύτερη δυνατότητα αναπαράστασης από τα δίκτυα ενός επιπέδου, μόνο αν χρησιμοποιούν μη γραμμική συνάρτηση ενεργοποίησης. Ειδικά για

δίκτυα που εκπαιδεύονται με ανάστροφη μετάδοση λάθους, η συνάρτηση ενεργοποίησης θα πρέπει επιπλέον να είναι μονότονα αύξουσα και παραγωγίσιμη σε όλο το φάσμα των τιμών εισόδου.

Το δίκτυο ξεκινά του υπολογισμούς με τυχαίες τιμές βαρών οι οποίες όμως θα πρέπει να αναπροσαρμοστούν ώστε να περιοριστεί το σφάλμα στην έξοδο. Δεδομένου ότι για τους νευρώνες εξόδου είναι γνωστά το επιθυμητό αποτέλεσμα και το ακριβές σφάλμα, είναι δυνατό να χρησιμοποιηθεί ο γενικευμένος κανόνας Δέλτα για να αναπροσαρμοστούν οι τιμές των βαρών μεταξύ του επιπέδου εξόδου και του προηγούμενου επιπέδου. Αποδεικνύεται ότι:

$$\Delta w_{jk} = d \times \delta_k \times z_j$$

Όπου d είναι ο ρυθμός μάθησης, δ_k ο ρυθμός μεταβολής του σφάλματος ως προς την είσοδο του νευρώνα k και z_j η πραγματική έξοδος του νευρώνα j. Για τους νευρώνες του επιπέδου εξόδου η ποσότητα δ_κ δίνεται από τη σχέση: $\delta_k=(t_k-y_k)f'(input_k)$ όπου t_k η επιθυμητή έξοδος για νευρώνα k και f' η πρώτη παράγωγος της συνάρτησης ενεργοποίησης. Για τους νευρώνες του κρυφού επιπέδου ισχύει η σχέση:

$$\delta j=f'(input_j)\times\sum_{k=1}^{m} \delta_k \times w_{jk}$$

Μέσω αυτής μπορούν να αναπροσαρμοστούν τα βάρη σύμφωνα με τη σχέση:

$$\Delta w_{ij} = d \times \delta_j \times x_i$$

Γίνεται πλέον φανερό ότι μπορεί να υπολογιστούν αρχικά οι νέες βαρών που συνδέουν το επίπεδο εξόδου με το προηγούμενο κρυφό επίπεδο, στη συνέχεια να υπολογιστούν τα βάρη που συνδέουν αυτό το κρυφό επίπεδο με το προηγούμενο, κ.ο.κ. μέχρι και τα βάρη μεταξύ επιπέδου εισόδου και του πρώτου κρυφού επιπέδου. Αυτό το στάδιο αναπροσαρμογής των βαρών ονομάζεται ανάστροφο πέρασμα (backward pass) ή ανάστροφη μετάδοση (back propagation).

Ο αλγόριθμος της ανάστροφης μετάδοσης λάθους είναι μια διαδικασία βελτιστοποίησης επικλινούς καθόδου η οποία ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα σύμφωνα με όσα ειπώθηκαν στην προηγούμενη παράγραφο και σχετίζονται με το κανόνα Δέλτα.

Κατά την παραπάνω προσπάθεια ελαχιστοποίησης μπορεί να θεωρηθεί σαν μια αναζήτηση του ολικού ελαχίστου της συνάρτησης σφάλματος, η οποία έχει σαν παραμέτρους τις τιμές των βαρών. Η διόρθωση που γίνεται κάθε φορά προσπαθεί να ελαχιστοποιήσει το σφάλμα διαλέγοντας να κάνει εκείνες τις αλλαγές που φαίνεται να το μειώνουν

τοπικά. Αυτή η αναζήτηση είναι τύπου αναρρίχησης λόφων. Σε ορισμένες περιπτώσεις το δίκτυο που έχει εκπαιδευτεί με αυτόν τον τρόπο ίσως να μην αποδώσει τα αναμενόμενα πρόκειται δηλαδή για δίκτυο που πέφτει σε τοπικά ελάχιστα (local minima) ή/και παραλύει τελείως (network paralysis).

Όταν το δίκτυο πέφτει σε τοπικά ελάχιστα πρόκειται για εγγενή αδυναμία της αναζήτησης αναρρίχησης λόφων ναι βρει το ολικό ελάχιστο, δηλαδή το κατάλληλο διάνυσμα βαρών που ελαχιστοποιεί το σφάλμα. Το πρόβλημα αυτό λύνεται συνήθως, με την χρήση άλλων στατιστικών μεθόδων. Επιπρόσθετα όταν εμφανίζεται το φαινόμενο της παράλυσης το νευρωνικό δίκτυο πέφτει σε στάσιμη κατάσταση γιατί κάποια βάρη έχουν σταθερά υψηλές απόλυτες τιμές και ταυτόχρονα δεν τροποποιούνται σημαντικά σε κάθε κύκλο διόρθωσης. Σε αυτή την περίπτωση ίσως χρειαστεί μεγάλος (π.χ. εκατομμύρια) αριθμός κύκλων εκπαίδευσης. Μια λύση σε αυτό το πρόβλημα είναι η αύξηση του ρυθμού μάθησης d.

Τέλος, συνήθως, γίνεται χρήση ενός συνόλου προτύπων αξιολόγησης μέσω του οποίου ελέγχεται η απόδοση του δικτύου μετά ή και κατά την διάρκεια της εκπαίδευσης.

ΚΕΦΑΛΑΙΟ 4 ΔΕΝΔΡΑ ΑΠΟΦΑΣΕΩΝ

4.1 Ορισμός δένδρου απόφασης

Τυπικά το δένδρο απόφασης (ΔΑ) ορίζεται ως εξής :

Δέντρο Απόφασης (ΔΑ) ή Δέντρο Κατηγοριοποίησης είναι ένα δέντρο με τις ακόλουθες ιδιότητες:

– Κάθε εσωτερικός κόμβος ονοματίζεται με το όνομα ενός χαρακτηριστικού X_i.

– Κάθε κλαδί/σύνδεση ονοματίζεται με ένα κατηγόρημα που μπορεί να εφαρμοστεί στο χαρακτηριστικό που αποτελεί το όνομα του κόμβου πατέρα.

– Κάθε φύλλο ονοματίζεται με το όνομα μιας κλάσης.

Επίσης πρέπει να αναφερθούν κάποιοι σημαντικοί ορισμοί σχετικά με τα ΔΑ.

- Χαρακτηριστικά διάσπασης (splitting features) Τα χαρακτηριστικά των παραδειγμάτων στη βάση D που χρησιμοποιούνται σαν ονόματα κόμβων του δέντρου, δηλ. επιλέχτηκαν ως καλύτερα χαρακτηριστικά.
- • Χαρακτηριστικό στόχου (target feature) Το χαρακτηριστικό που οι τιμές του αντιπροσωπεύουν τις κλάσεις κατηγοριοποίησης.
- Κατηγορήματα διάσπασης (splitting predicates) Τα κατηγορήματα που χρησιμοποιούνται σαν ονόματα των κλάδων του δέντρου.
- Κριτήριο διάσπασης (splitting criterion). Το κριτήριο με βάση το οποίο επιλέγεται το καλύτερο χαρακτηριστικό διάσπασης κάθε φορά.
- Κριτήριο τερματισμού (stopping criterion). Το κριτήριο με βάση το οποίο τερματίζεται ο αλγόριθμος.

Μια αρκετά διαδεδομένη μέθοδος μηχανικής μάθησης είναι εκείνη που βασίζεται σε δένδρα απόφασης. Σε αυτή την μέθοδο, επιχειρείται η προσέγγιση μιας συνάρτησης στόχου ακλουθώντας την μέθοδο «διαίρει και βασίλευε» (divide and conquer). Ο χώρος του προβλήματος χωρίζεται σε περιοχές από στιγμιότυπα που φέρουν την ίδια τιμή ως προς κάποια μεταβλητή χαρακτηριστικό, και η διαδικασία επαναλαμβάνεται αναδρομικά, αναπαριστώντας με τον τρόπο αυτό το παραγόμενο μοντέλο ως δένδρο απόφασης. Στα θετικά σημεία των δέντρων απόφασης συγκαταλέγονται: η γρήγορη εκπαίδευση και η δυνατότητα μεταφοράς του παραγόμενου μοντέλου από δένδρο απόφασης σε ένα σύνολο κανόνων συμπερασμού (if – then rules), προς διευκόλυνση της κατανόησής του.

4.2 Αλγόριθμοι Δένδρων Αποφάσεων

Υπάρχουν αρκετοί αλγόριθμοι δένδρων αποφάσεων οι οποίοι λίγο ή πιο πολύ διαφέρουν μεταξύ τους. Οι κυριότεροι είναι οι παρακάτω:

- Ο ID3 (Iterative Dichotomiser 3) αναπτύχθηκε από τον Ross Quinlan το 1986. Ο αλγόριθμος αυτός δημιουργεί ένα δένδρο πολλών δρόμων και βρίσκει για κάθε κόμβο εκείνο το χαρακτηριστικό στοιχείο που θα αποδώσει το μεγαλύτερο κέρδος πληροφορίας για τις στοιχεία «στόχους» (target). Τα αντίστοιχα δένδρα μεγαλώνουν στο μέγιστο δυνατό και τότε συνήθως συμβαίνει ένα «κλάδεμα» βήματος προκειμένου να βελτιωθεί τη ικανότητα του δένδρου να γενικεύει με νέα δεδομένα.
- Ο C4.5 είναι μεταγενέστερος του ID3. Στον αλγόριθμο αυτόν τα εκπαιδευμένα δένδρα αποφάσεων, π.χ. οι έξοδοι του ID3, μετατρέπονται σε σετ με κανόνες συμπερασμού (if-then rules). Τότε η ακρίβεια του κάθε κανόνα αξιολογείται έτσι ώστε να καθοριστεί η σειρά με την οποία πρέπει να εφαρμοστούν. Εδώ, γίνεται το «κλάδεμα» αφαιρώντας την προϋπόθεση ενός κανόνα αν η ακρίβεια του κανόνα βελτιώνεται χωρίς αυτή.
- Ο C5.0 είναι η τελευταία έκδοση του Quinlan. Χρησιμοποιεί λιγότερη μνήμη και φτιάχνει μικρότερα σετ κανόνων από τον C4.5. Επίσης έχει και μεγαλύτερη ακρίβεια.
- Ο αλγόριθμος CART (Classification and Regression Trees) είναι παρόποιος με τον C4.5 αλλά διαφέρει στο γεγονός ότι υποστηρίζει αριθμητικές τιμές στις μεταβλητές-στόχους (regression) και δεν υπολογίζει σετ κανόνων. Ο CART κατασκευάζει δυαδικά δένδρα χρησιμοποιώντας τα χαρακτηριστικά της εισόδου (input node αναλογικά με τα ANN) που μεγιστοποιεί το κέρδος από τις πληροφορίες της εισόδου, σε κάθε κόμβο.

Στην παρούσα διπλωματική χρησιμοποιήθηκαν οι εφαρμογές μηχανικής μάθασης του scikit-learn το οποίο υλοποιεί μια βελτιστοποιημένη έκδοση του CART, έτσι στα παρακάτω παρουσιάζεται διεξοδικά ο συγκεκριμένος αυτός αλγόριθμος.

Η μεθοδολογία CART προτάθηκε από τον Breiman το 1984 για το δημιουργία ΔΑ. Ο CART χρησιμοποιεί ένα σετ μάθησης (learning set) δηλαδή τα ιστορικά δεδομένα με καθορισμένα τα δεδομένα εξόδου. Ένας αλγόριθμος που είναι γνωστός ως «επαναληπτικός διαχωρισμός»(recursive partitioning) είναι το κλειδί για την λειτουργία του CART. Είναι μια διαδικασία βήμα προς βήμα κατά την οποία κατασκευάζεται ένα ΔΑ είτε διαχωρίζοντας είτε όχι κάθε κόμβο σε δύο «κόμβους-απογόνους». Αυτή η ιδιότητα του CART είναι πολύ σημαντική καθώς κάνει το αλγόριθμο απλό ως προς την κατανόηση και ερμηνεία των αποτελεσμάτων. Το μοναδικό σημείο αφετηρίας του ΔΑ (classification ή regression) ονομάζεται ρίζα (root node) και περιλαμβάνει όλο το σετ μάθησης L στην κορυφή του δένδρου. Ένα κόμβος είναι ένα υποσύνολο των μεταβλητών και μπορεί να είναι είτε τελικός κόμβος είτε όχι. Ένας μη τελικός κόμβος (γονεάς, parent) είναι ένας κόμβος που διασπάται σε δύο κόμβους-απόγονους (binary split). Αυτή η διάσπαση καθορίζεται από μια προϋπόθεση μιας μεταβλητής όπου

η συνθήκη ικανοποιείται ή όχι με βάση την παρατηρούμενη τιμή της μεταβλητής. Όλες οι παρατηρήσεις στο σύνολο L, που έχουν φτάσει σε ένα κόμβο-γονέα και ικανοποιούν την συνθήκη για αυτή τη μεταβλητή προχωρούν προς τα κάτω σε έναν από τους δύο κόμβους-απογόνους. Όσες από τις εναπομείναντες παρατηρήσεις (στον γονέα) δεν ικανοποιούν την συνθήκη προχωρούν προς τα κάτω στον άλλο κόμβο-απόγονο. Ο κόμβος ο οποίος δεν διαιρείται ονομάζεται τελικός κόμβος και παίρνει μια ετικέτα κλάσης (class label). Κάθε παρατήρηση στο L πάει υποχρεωτικά σε έναν τελικό κόμβο. Για να παραχθεί, με το παραπάνω τρόπο ένα μοντέλο ΔΑ, ο CART καθορίζει το καλύτερο σπάσιμο του σετ L για να γίνει η αρχή και στην συνέχει το καλύτερο σπάσιμο των υποσυνόλων με βάση διάφορα θέματα όπως τον καθορισμό των μεταβλητών για το νέο σπάσιμο καθορίζοντας πότε ένας κόμβος είναι τελικός. Το να καθοριστεί ένας τελικός κόμβος είναι σχετικό απλό ενώ η εύρεση του βέλτιστου ύψους του δένδρου ή ο τρόπος που θα γίνει ο διαχωρισμός είναι δυσκολότερες διαδικασίες. Παρακάτω ο CART αναλύεται διεξοδικά.

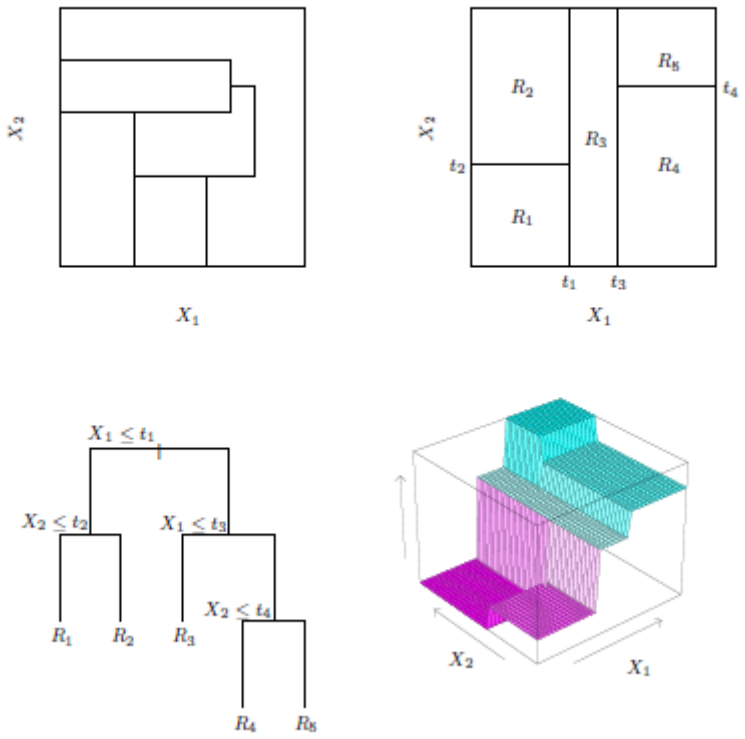
Γενικά το σύνηθες κριτήριο για ένα πρόβλημα παλινδρόμισης είναι το υπολοιπόμενο άθροισμα τετραγώνων (RSS) ενώ για ένα δένδρο ταξινόμισης ο δείκτης Gini.

Παράδειγμα δένδρου παλινδρόμησης

Έστω συνεχής απόκριση Y και εισόδου X_1 , X_2. Αρχικά χωρίζουμε το σύνολο σε ύο υποσύνολα στο σημείο $X_1=t_1$. Μετά η περιοχή $X_1 \leq t_1$ χωρίζεται στο $X_2 = t_2$ και η περιοχή $X_1 > t_1$ στο σημείο $X_1 = t_3$. Τέλος η περιοχή $X_1 > t_3$ χωρίζεται στο σημείο $X_2 = t_4$. Η διαδικασία δίνεται σχηματικά στο παρακάτω σχήμα 5.1 Το αποτέλεσμα αυτής είναι ο διαχωρισμός του συνόλου σε πέντε περιοχές R_1 έως και R_5, όπως φαίνεται στο σχήμα 5.2. Η τιμή της απόκρισης είναι η μέση τιμή κάθε μιας από τις πέντε περιοχές $\bar{y}_1$ έως και $\bar{y}_5$.

Γενικά σε περιοχή R_m το μοντέλο παλινδρόμησης προβλέπει την έξοδο Y μέσω σταθεράς c_m σύμφωνα με την σχέση:

$$\hat{f}(x) = \sum_{m=1}^{5} c_m I\{(X_1, X_2) \in R_m\}$$



Εικόνα 4. 1 Διαχωρισμοί

Για το σχήμα 4.3 έχουμε τα εξής: Στο πάνω δεξιά σχήμα έχουμε διαχωρισμό με μεταβλητές δύο διαστάσεων με το binary recursive splitting, όπως και στο CART. Στο πάνω αριστερά σχήμα έχουμε γενικό διαχωρισμό όχι με recursive splitting. Στο κάτω αριστερά σχήμα φαίνεται το δένδρο που αντιστοιχεί στο πάνω δεξιά πρόβλημα και στο κάτω δεξιά βλέπουμε το διάγραμμα βλέπουμε την περιοχή πρόβλεψης.

4.3 Δένδρα παλινδρόμησης

Το ερώτημα που προκύπτει λοιπόν, είναι ο τρόπος να μεγαλώσουμε ένα δένδρο παλινδρόμησης. Έστω ότι τα δεδομένα αποτελούνται από p εισόδους και μία έξοδο για κάθε N παρατηρήσεις, δηλαδή (x_1,y_1) για i=1,2,..,N, με $x_i = (x_{i1},x_{i2},..,x_{ip})$. Ο αλγόριθμος πρέπει αυτόματα να αποφασίζει τις μεταβλητές διαχωρισμού (split) και τα σημεία διαχωρισμού. Επίσης πρέπει να αποφασίζει την τοπολογία, δηλαδή το σχήμα, που πρέπει να έχει το δένδρο. Θεωρούμε ότι αρχικά έχουμε να διαχωρίσουμε σε M περιοχές $R_1,\ldots,R_m$, και μοντελοποιούμε την έξοδο σαν μια σταθερά c_m σε κάθε περιοχή:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

Εάν χρησιμοποιήσουμε το κριτήριο ελαχιστοποίησης του αθροίσματος τετραγώνων $\sum(y_i-f(x_i))^2$ είναι εύκολο να δούμε ότι το καλύτερο c_m είναι ο μέσος όρος των y_i στην περιοχή R_m:

$$\hat{c}_m = avg(y_t \mid x_t \in R_m).$$

Τώρα για να βρούμε το καλύτερο binary partition σε σχέση με το άθροισμα ελαχίστων τετραγώνων είναι υπολογιστικά αδύνατο. Έτσι κάνουμε χρήση ενός άπληστου αλγορίθμου. Ξεκινώντας με όλα τα δεδομένα, λαμβάνοντας υπ΄ όψιν μια μεταβλητή διαχωρισμού j και σημείο διαχωρισμού s, ανδ ορίζουμε το ζευγάρι:

$$R_1(j,s) = \{X \mid X_j \le s\} \text{ and } R_2(j,s) = \{X \mid X_j > s\}.$$

Τότε ψάχνουμε μεταβλητή διαχωρισμού j και σημείο διαχωρισμού s που λύνει την παρακάτω εξίσωση:

$$\min_{j,s}[\min_{c_1} \sum_{x_t \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_t \in R_2(j,s)} (y_i - c_2)^2]$$

Για οποιαδήποτε επιλογή j και s η ελαχιστοποίηση λύνεται από το:

$$\hat{c}_1 = ave(y_i \mid x_i \in R_1(j,s)) \text{ and } \hat{c} = ave(y_i \mid x_i \in R_2(j,s)).$$

Για κάθε μεταβλητή διαχωρισμού ο ορισμός του σημείο διαχωρισμού s μπορεί να γίνει γρήγορα ελέγχοντας όλες τις εισόδους για ποιο ζευγάρι (j,s) είναι το κατάλληλο.

Έχοντας βρει τον καλύτερο διαχωρισμό, διαχωρίζουμε τα δεδομένα σε δύο περιοχές και επαναλαμβάνουμε τη διαδικασία για κάθε μια από τις δύο περιοχές.

Το ερώτημα που προκύπτει είναι πόσο πρέπει να μεγαλώσει ένα δένδρο. Προφανώς αν ένα δένδρο μεγαλώσει πολύ μπορεί να συμβεί overfit και αντίθετα αν μεγαλώσει λίγο δεν θα έχει μάθει όλες τις απαραίτητες πληροφορίες.

Το μέγεθος του δένδρου είναι μια παράμετρος για να βελτιστοποιηθεί το μοντέλο και το βέλτιστο μέγεθος πρέπει να βρεθεί με τα δεδομένα. Μία προσέγγιση θα ήταν να διαχωριστούν οι κόμβοι του δένδρου μόνο αν η μείωση στο άθροισμα τετραγώνων περνάει κάποιο κατώφλι. Αυτή η στρατηγική είναι κοντόφθαλμη, ωστόσο, εφόσον ένα άχρηστο σπάσιμο μπορεί στη συνέχεια να οδηγήσει σε ένα χρήσιμο.

Η προτιμούμενη στρατηγική είναι να μεγαλώσουμε ένα μεγάλο δένδρο T_0, σταματώντας τον διαχωρισμό μόνο όταν ένα ελάχιστο μέγεθος κόμβου βρεθεί. Τότε αυτό το μεγάλο δένδρο κλαδεύεται χρησιμοποιώντας cost-complexity pruning που περιγράφεται ακριβώς παρακάτω.

Ορίζουμε υποδένδρο T υποσύνολο του T_0, ως οποιοδήποτε δένδρο που προκύπτει από κλάδεμα του T_0. Δίνουμε δείκτη m στους τελικούς κόμβους, με το m να

συμβολίζει την περιοχή R_m. Με |T| να είναι ο αριθμός των τελικών κόμβων στο T. Έστω

$$N_m = \#\{x_i \in R_m\},$$
$$\hat{c}_m = \frac{1}{N_m} \sum_{x_t \in R_m} y_i,$$
$$Q_m(T) = \frac{1}{N_m} \sum_{x_t \in R_m} (y_i - \hat{c}_m)^2,$$

Ορίζουμε το cost complexity κριτήριο ως:

$$C_a(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + a\,|T|.$$

Η ιδέα είναι να βρούμε για κάθε α, το υποδένδρο T_α για να ελαχιστοποιηθεί το $C_a(T)$.

Η παράμετρος $\alpha \geq 0$ ορίζει την σχέση μεταξύ του μέγεθος δένδρου και τις ικανότητας του να ταιριάζει στα δεδομένα. Μεγάλες τιμές του α έχουν ως αποτέλεσμα μικρότερα δένδρα T_α. Για να βρούμε το T_α χρησιμοποιούμε το weakest link pruning: χωρίζουμε τον εσωτερικό κόμβο που παράγει το λιγότερο κέρδος πληροφορίας και συνεχίζουμε μέχρι να βρούμε δένδρο με μόνο έναν κόμβο. Αυτό δίνει μια σειρά από υποδένδρα που κάποιο από αυτά είναι το T_α.

4.4 Δένδρα κατηγοριοποίησης

Αν στόχος είναι η κατηγοριοποίηση δηλαδή να έχουμε έξοδο 1,2,...,K, οι μόνες αλλαγές που χρειάζονται στο αλγόριθμο είναι το κριτήριο για το διαχωρισμό τν κόμβων. Για παλινδρόμηση χρησιμοποιήσαμε το τετραγωνικό σφάλμα ως μέτρο αγνότητας των κόμβων $Q_m(T)$. Σε έναν κόμβο m, που συμβολίζει μια περιοχή R_m με N_m παρατηρήσεις, έστω:

$$p_{mk} = \frac{1}{N_m} \sum_{x_t \in R_m} I(y_i = k),$$

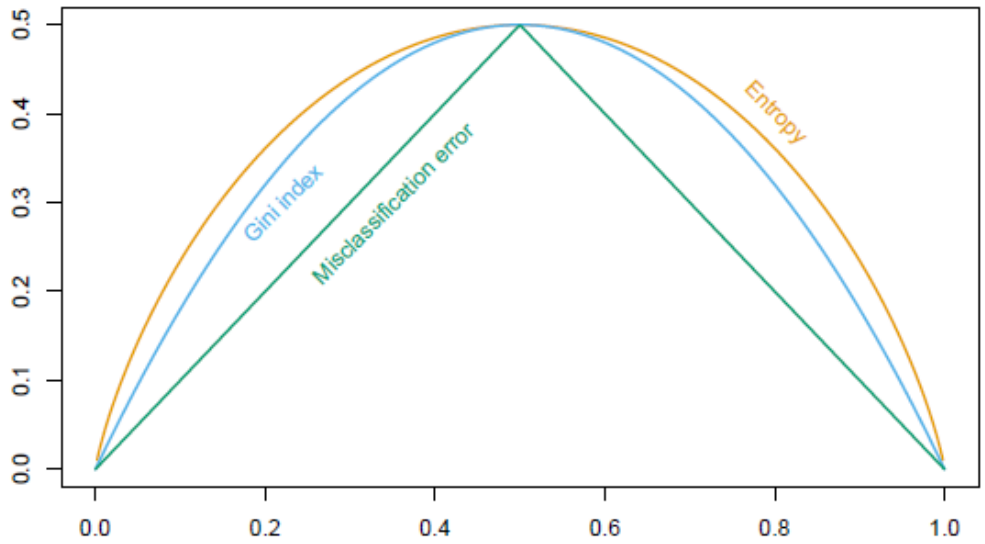
Το ποσοστό της κλάσης k παρατηρήσεων στον κόμβο m. Κατηγοριοποιούμε τις παρατηρήσεις στον κόμβο m στην κλάση $k(m) = \text{argmax}_k \hat{p}_{mk}$, η κλάση που πληοψηφεί στον κόμβο m. Άλλες μετρικές $Q_m(T)$ της λεγόμενης ακαθαρσίας κόμβου(node impurity) είναι οι παρακάτω:

Misclassification error: $\frac{1}{N_m}\sum_{i\in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}.$

Gini index: $\sum_{k\neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}).$

Cross-entropy or deviance: $-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$

Για παράδειγμα, αν έχουμε δύο κλάσεις και p το ποσοστό στην δεύτερη κλάση αυτές οι τρεις μετρικές είναι 1-max(p,1-p), 2p(1-p) και logp – 1(1-p)log(1-p) αντίστοιχα. Διαγραμματικά φαίνονται σοτ σχήμα 6.4.



Εικόνα 4. 2 Μετρικές Node Impurity

Οι συναρτήσεις entropy και Gini index είναι πιο ευαίσθητες σε αλλαγές στο κόμβο πιθανότητας παρά το σφάλμα κατηγοριοποίησης. Για αυτό τον λόγο χρησιμοποιούνται πολύ όταν μεγαλώνει ένα δένδρο.

4.5 Συμπεράσματα για την τεχνική CART

Τα δένδρα είναι χρήσιμα για μεγάλα σύνολα δεδομένων καθώς, για να επιτύχουμε το μοντέλο πρέπει να γίνουν ισχυρές παραδοχές ενώ δεν γίνονται υποθέσεις για τα δεδομένα μας. Έτσι λοιπόν σε μεγάλα σύνολα δεδομένων με πολλές μεταβλητές προς επιλογή είναι χρήσιμη η τεχνική CART καθώς σε άλλες μοντέλα παλινδρόμησης δεν έχουμε τόσο καλά αποτελέσματα. Συμπερασματικά μπορούμε να πάρουμε έτσι καλύτερες προβλέψεις. Σε σχέση με άλλα μοντέλα, τα δένδρα μπορούν να προτείνουν πιθανές αλληλεπιδράσεις μεταξύ των μεταβλητών και έτσι να ανακαλύψουν και μη γραμμική σχέση ανάμεσα στις μεταβλητές εισόδου. Τέλος πιο

προηγμένες τεχνικές, όπως τα random forests και η μέθοδος bagging βασίζονται στα δένδρα απόφασης.

4.6 Εισαγωγή στα Random Forests

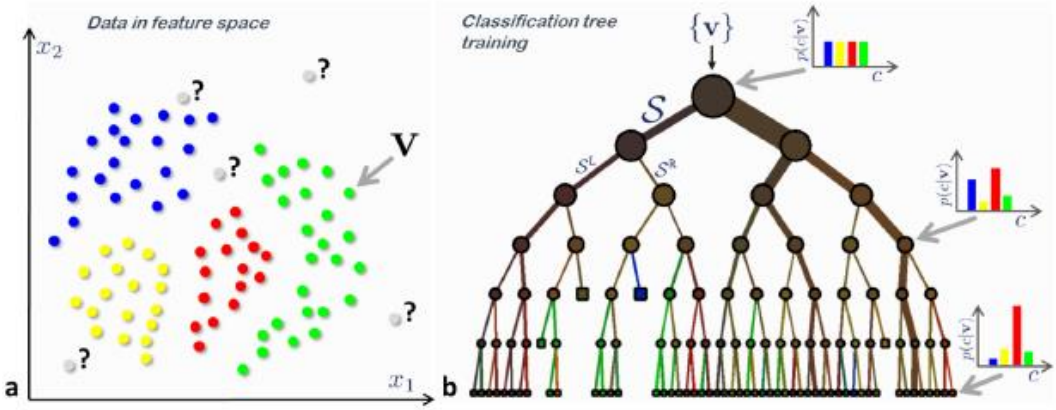
Τα Random Forests είναι μια μέθοδος μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί, όπως και οι υπόλοιπες της παρούσας διπλωματικής, τόσο σε προβλήματα ταξινόμησης όσο και παλινδρόμησης. Η πρώτη ανάπτυξη της μεθόδου προτάθηκε από τον Tin Kam Ho, το 1995. Κατά την λειτουργία αυτής της μεθόδου, κατασκευάζεται πλήθος από δένδρα απόφασης σε κάποιο δείγμα του συνόλου των δεδομένων, όταν εκπαιδεύεται το μοντέλο και ύστερα, προκειμένου να καταλήξει στην έξοδο συνυπολογίζονται όλα τα δένδρα. Τα Random Forests αποτελούν μια μέθοδο παρόμοια με την bagging, η οποία παίρνει πολλά αμερόληπτα μοντέλα με θόρυβο και βρίσκει την μέση τιμή αυτών, μειώνοντας την διακύμανση της εξόδου.

Τα δένδρα απόφασης έχουν τις ιδιότητες που έχουν περιγραφεί σε προηγούμενο κεφάλαιο και επομένως είναι ιδανικά για μεθόδους bagging καθώς μπορεί να συλλάβουν πολύπλοκες αλληλεπιδράσεις μεταξύ των δεδομένων. Επίσης αν αποκτήσουν το ιδανικό βάθος αποτελούν ένα πολύ αμερόληπτο και κατά συνέπεια ακριβές μοντέλο. Ακόμα, λόγω των αποκλίσεων των τιμών κάθε δένδρου, τελικώς είναι ένας καλός δείκτης της τελικής εξόδου. Το καθένα δένδρο που παράγεται είναι ανεξάρτητο από τα άλλο δένδρα και παρέχει την ίδια κατανομή πιθανότητας ως προς την τελική έξοδο. Έτσι, ο μέρος όρος των δένδρων περιέχει την ίδια μεροληψία με κάθε δένδρο ξεχωριστά πετυχαίνοντας βελτίωση μέσω της διακύμανσης. Ακόμα το υπολογιστικό κόστος εκπαίδευσης του Random Forest είναι σχετικά μικρό και η μάθηση μπορεί να καταστεί αρκετά πετυχημένη και με λίγα δεδομένα. Από την άλλη μπορεί εύκολα να συμβεί overfitting στο μοντέλο.

4.7 Περιγραφή της μάθησης Random Forest

Πριν συμβεί κάθε σπάσιμο (spit) γίνεται η τυχαία επιλογή κάποιων μεταβλητών εισόδου ως υποψήφιες προς διαχωρισμό. Ύστερα, από αυτές τις μταβλητές γίνεται επιλογή του σημείου σπασίματος (split point) στο οποία θα υπάρξει το μεγαλύτερο κέρδος πληροφορίας και την μεταβλητή στην οποία θα συμβεί το split. Έπειτα προκύπτουν δύο κόμβοι-απόγονοι οι οποίοι δίνουν την βέλτιστη πληροφορία. Η διαδικασία αυτή γίνεται αναδρομικά (σε αντιστοιχεία με τη μέθοδο CART) για κάθε δένδρο και για πολλά διαφορετικά δένδρα (οι παράμετροι ρυθμίζονται από τον χρήστη), που λαμβάνονται τυχαία. Τέλος, η έξοδος βασίζεται στο σύνολο τον δένδρο ως ο μέσος όρος.

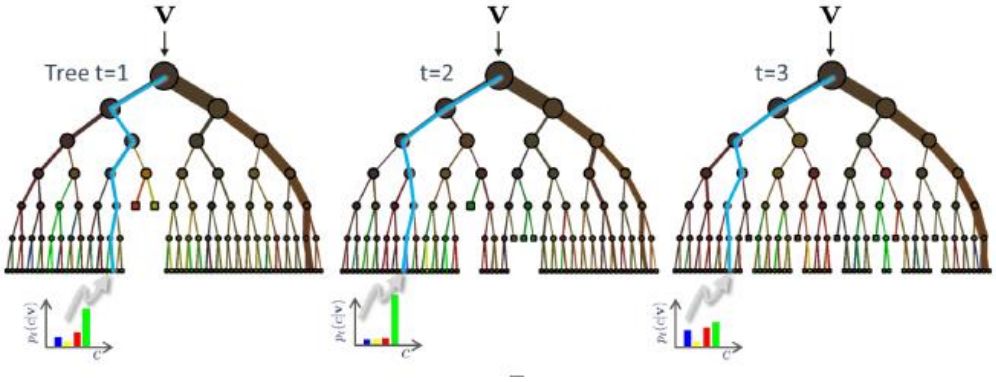
Στο παρακάτω σχήμα 5.1 φαίνεται ένα παράδειγμα μάθησης ενός classification tree.



Εικόνα 4. 3 Εκπαίδευση Δένδρου

Πρώτα διαλέγουμε τυχαία δεδομένα από το σύνολο εκπαίδευσης και ύστερα V-πλήθος δεδομένων χρησιμοποιούνται ώστε να βελτιστοποιηθούν οι παράμετροι του δένδρου.. Τα split points που μας δίνουν το μεγαλύτερο κέρδος πληροφορίας χρησιμοποιούνται για να σπάσουμε το δείγμα σε S_L και S_R κόμβους. Όσο πηγαίνουμε από ένα κόμβο, προς τα κάτω η συνάρτηση εντροπίας μειώνεται, δηλαδή αυξάνεται η ακρίβεια στην κατανομή των δεδομένων σε κλάσεις. Σε αυτό το παράδειγμα λοιπόν η μεταβλητή έχει πολύ μεγαλύτερη πιθανότητα να ανήκει στην κόκκκινη κλάση, στο φύλλο C. Παρατηρούμε ότι αρχικά οι μεταβλητές έχουν την ίδια κατανομή πιθανότητας

Στο σχήμα 5.2 φαίνεται ένα παράδειγμα δημιουργίας εξόδου για κάποιο από τα δεδομένα. Το κάθε τυχαίο δένδρο παράγει προβλέψεις οι οποίες έχουν προέλθει από την αναδρομική διαδικασία. Παρατηρείται ότι κάθε πρόβλεψη είναι διαφορετική για κάθε δένδρο και τα δένδρα εκπαιδεύονται ξεχωριστά. Κατά την δοκιμή του μοντέλου το σημείο των αρχικών δεδομένων v ωθείται σε όλα τα δένδρα μέχρι να φτάσει στο κατάλληλο φύλλο όπου θα ληφθεί η εκτίμηση. Στο παράδειγμα το δένδρο t = 2 παράγει την πιο σίγουρη πρόβλεψη ότι το σημείο θα ανήκει στην κλάση με πράσινο χρώμα, ενώ το t = 3 έχει μια πιο ομοιόμορφη κατανομή πιθανοτήτων. Η κάθε πρόβλεψη για το φύλλο είναι $p_t(c|v)$.



Εικόνα 4. 4 Έξοδος ως σύνολο της Εκπαίδευσης

Η τελική πρόβλεψη για το σημείο χρησιμοποιεί το σύνολο των τυχαίων δένδρων και όπως στα παραπάνω είναι απλά ο μέσος όρος των προβλέψεων για το σημείο αυτό δηλαδή:

$$p(c \mid v|) = \frac{1}{T}\sum_{t}^{T} p_t(c \mid v|)$$

4.8 Ορισμός των Random Forests

Για παλινδρόμηση ή κατηγοριοποίηση έχουμε το εξής αλγόριθμο:

1. Για b=1 έως B:
   (α) Διάλεξε ένα bootstrap δείγμα Z^* μεγέθους N από τα δεδομένα εκπαίδευσης.
   (β) Ανέπτυξε ένα τυχαίο δένδρο T_b στο δείγμα που πήρες, με το να επαναλαμβάνεις αναδρομικά τα παρακάτω βήματα για κάθε τελικό κόμβο του δένδρου, έως ότου φτάσεις σε κάποιο ελάχιστο μέγεθος κόμβων n_{min}.
   I. Διάλεξε m τυχαία χαρακτηριστικά από τα p-χαρακτηριστικά.
   II. Επέλεξε την καλύτερη μεταβλητή/σημείο διαχωρισμού από τα m.
   III. Διαίρεσε τον κόμβου σε κόμβους-απογόνους.

2. Η έξοδος είναι το σύνολο (ensemble) των δένδρων $\{T_b\}_1^B$.

Για να γίνει μια πρόβλεψη στο νέο σημείο x είναι:

Regression :

$$\widehat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b)$$

Classification:

: Έστω $\widehat{C}_b(x)$ η πρόβλεψη για την κλάση του b-δέντρου. Τότε:

$$\widehat{C}_{rf}^{B}(x_{}) = \text{majority vote}\left\{\widehat{C}_b(x)\right\}_1^B$$

Σε αντίθεση με το CART, λοιπόν, το split γίνεται σε μέρος τον δεδομένων κάθε φορά και η πρόβλεψη είναι ο μέσος όρος των εξόδων. Πιο συγκεκριμένα, τα τυχαία δένδρα είναι identically distributed. Ο μέσος όρος B identically distributed μεταβλητών έχει διακύμανση ίση με:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Καθώς το B αυξάνει, ο δεύτερος όρος εξαφανίζεται, επομένως όσο αυξάνει το σύνολο των δεδομένων τόσο λιγότερο μειώνεται η διακύμανση. Ακόμα, όσο πιο συσχετισμένα είναι τα δένδρα μεταξύ τους τόσο περιορίζονται τα οφέλη του μέσου όρου. Η ιδέα των random forests είναι να μειωθεί η συσχέτιση ανάμεσα στα δένδρα, χωρίς να μειωθεί πολύ η διακύμανση. Αυτό πετυχαίνεται μέσω της τυχαίας επιλογής των μεταβλητών. Ειδικότερα, πριν από κάθε σπάσιμο διαλέγουμε $m \leq p$ από τις μεταβλητές εισόδου τυχαία ως υποψήφιες προς διαχωρισμό. Τυπικές τιμές για m είναι από $\sqrt{p}$ έως και 1.

Όταν B τέτοια δένδρα $\{T(x;\Theta_b)\}_1^B$ δημιουργηθούν, τότε ο μέσος όρος αυτών θα είναι η έξοδος. Η μείωση του m θα μείωνε την συσχέτιση οποιουδήποτε ζεύγους δένδρων στον σύνολο των τυχαίων δένδρων, οπότε θα μείωνε και την διακύμανση στον μέσο όρο. Αυτός είναι ο λόγος που τα μοντέλα CART δεν παράγουν τόσο καλά αποτελέσματα όσο τα μοντέλα random forests. Λαμβάνουν υπ΄ όψιν ένα πολύ μεγαλύτερο σύνολο δεδομένων, οπότε έχουν και μεγαλύτερη διακύμανση στην έξοδο και τα αποτελέσματα είναι γενικά χειρότερα. Τελικώς στα random forests, λόγω της μικρής συσχέτισης των δένδρων μεταξύ τους, μέσω του μέσου όρου μπορούμε να πετύχουμε πολύ καλά αποτελέσματα λόγω μείωσης της διακύμανσης.

4.9 Δείγματα out of bag

Ένα σημαντικό χαρακτηριστικό του αλγορίθμου είναι ότι χρησιμοποιεί δείγματα out of bag (OOB), τα οποία ορίζονται ως εξής:

Για κάθε παρατήρηση $z_i = (x_i, y_i)$ δημιούργησε έναν *random forest predictor* μέσω του μέσου όρου των δέντρων που αντιστοιχούν στα δείγματα στα οποία δεν βρίσκεται ο z_i.

Μια εκτίμηση σφάλματος OOB είναι σχεδόν ταυτόσημη με αυτή που λαμβάνουμε με ένα N-fold Cross Validation. Σε αντίθεση όμως με άλλους μη γραμμικούς εκτιμητές, μπορούμε καθώς τροφοδοτούμε το δέντρο σε μία ακολουθία να εκτελούμε παράλληλα cross validation. Μόλις το σφάλμα OOB σταθεροποιηθεί, η εκπαίδευση μπορεί να σταματήσει.

4.10 Σημαντικότητα των Features

΄Ενα πολύ σημαντικό χαρακτηριστικό του αλγορίθμου Random Forest όπως και του αλγορίθμου CART είναι η σημαντικότητα των predictors. Σε κάθε διαχωρισμό και σε κάθε δέντρο ξεχωριστά, η επιπλέον βελτίωση στην απόδοση του αλγορίθμου μέσω του συγκεκριμένου predictor και σύμφωνα με το κριτήριο διαχωρισμού είναι το μέτρο σημαντικότητας της μεταβλητής, και συσσωρεύεται για όλα τα δέντρα στο δάσος και ξεχωριστά για κάθε μεταβλητή. Στον αλγόριθμο Random Forest λόγω του κριτηρίου διαχωρισμού, η πιθανότητα όλες οι μεταβλητές να έχουν ρόλο στο τελικό δέντρο, ακόμα και μικρό, είναι πολύ αυξημένη, ειδικά σε σχέση με άλλες μεθόδους όπως το gradient boosting.
Επιπρόσθετα, στον αλγόριθμο Random Forest μπορούμε μέσω των OOB samples να δημιουργήσουμε ένα μοντέλο σημαντικότητας που μετρά την προβλεπτική ικανότητα της κάθε μεταβλητής. ΄Οταν το δέντρο b μεγαλώνει, μέσω των OOB samples μπορούμε να μετρήσουμε την ακρίβεια πρόβλεψης. ΄Επειτα οι τιμές για την μεταβλητή j μετατίθενται τυχαία στα oob δείγματα και η ακρίβεια υπολογίζεται ξανά. Η μέση τιμή της μείωσης της ακρίβειας ως αποτέλεσμα των μεταθέσεων σε όλα τα δέντρα είναι ένας δείκτης της σημαντικότητας της μεταβλητής j στο random forest.

4.11 Random Forests και Overfitting

΄Οταν στο δείγμα έχουμε μικρό αριθμό σημαντικών μεταβλητών τότε ο αλγόριθμος Random Forest εμφανίζει χειρότερη επίδοση κάθε φορά που αυξάνουμε τον αριθμό των μεταβλητών που δίνουν θόρυβο στην έξοδο. ΄Οταν ο αριθμός των σημαντικών μεταβλητών αυξάνεται, έχουμε ισχυρή επίδοση ακόμα και με πολλές μεταβλητές με θόρυβο. Για παράδειγμα, με 6

ΚΕΦΑΛΑΙΟ 5 SUPPORT VECTOR MACHINES

5.1 Εισαγωγή

Η τεχνική Support Vector Machine (SVM) αποτελεί μια γενίκευση των τεχνικών διαχωρισμού, δηλαδή κατηγοριοποίησης. Σε αυτή την περίπτωση όμως, οι κατηγορίες δεν είναι απόλυτα διαχωρίσιμες αλλά περιέχουν επικάλυψη. Το σύνορο που προκύπτει από την τεχνική SVM μπορεί να είναι και μη γραμμικό. Επίσης η SVM μπορεί να χρησιμοποιηθεί και για την παλινδρόμηση. Η SVM ανήκει στην κατηγορία της επιβλεπόμενης μάθησης. Η πρώτη προσέγγιση έγινε από την Hava Siegelmann και τον Vladimir Vapnik.

5.2 Ο Support Vector Classifier

Τα δεδομένα εκπαίδευσης αποτελούνται από N ζευγάρια $(x_1 , y_1),\ldots, (x_N , y_N)$, με $x_i \in R^p$ και $y_i \in \{-1,1\}$. Ορίζεται το hyperplane ως:

$$\{x : f(x) = x^T\beta + \beta_0 = 0\}$$

Με β ένα μοναδιαίο διάνυσμα: ||β|| =1. Ένας κανόνας κατηγοριοποίησης που προκύπτει από την f(x) είναι:

$$G(x) = sign[x^T\beta + \beta_0]$$

Η f(x) δίνει μια προσημασμένη απόσταση ενός σημείου x από το hyperplane της f(x). Αφού οι κλάσεις είναι διαχωρίσιμες μπορούμε να βρούμε μια συνάρτηση : $f(x) = x^T\beta + \beta_0$ με

$y_i f(x_i) > 0$ για κάθε x_i. Έτσι μπορούμε να βρούμε το hyperplane που δημιουργεί το μεγαλύτερο «κενό» μεταξύ των σημείων εκπαίδευσης για τις κλάσεις 1 και -1 (σχήμα 6.1). το πρόβλημα βελτιστοποίησης:

$$\max_{\beta,\beta_0,\|\beta\|=1} M$$
$$subject : y_i(x_i^T\beta + b_0) \geq M$$

Ακόμα μπορεί να οριστεί ως πρόβλημα ελαχιστοποίησης:

$$\min_{b,b_0} \| \beta \|$$
$$subject: y_i(x_i\beta + \beta_0) \geq 1,$$
$$i = 1$$

Το φάσμα στο σχήμα είναι ότι Μ στοιχεία μακρυά από το hyperplane να βρίσκονται σε κάθε μεριά οπότε έχουμε απόσταση 2Μ στοιχεία. Αυτό λέγεται περιθώριο. Το $M = \frac{1}{||\beta||}$ είναι ο συνήθης τρόπος για γραφτεί το κριτήριο για τα διαχωρισμένα δεδομένα.

Στην περίπτωση που τα δεδομένα έχουν επικάλυψη στο χώρο ένας τρόπος να αντιμετωπιστεί το πρόβλημα είναι να μεγιστοποιηθεί το Μ, αλλά αυτό επιτρέπει σε κάποια σημεία να είναι στη λάθος μεριά του περιθωρίου. Ορίζουμε τις μεταβλητές ξ = (ξ_1, ξ_2, ..., ξ_N).

Υπάρχουν δύο τρόποι να αλλάξει το παραπάνω πρόβλημα μεγιστοποίησης

$$\min_{b,b_0} \| \beta \|$$
$$subject: y_i(x_i\beta + \beta_0) \geq 1,$$
$$i = 1$$

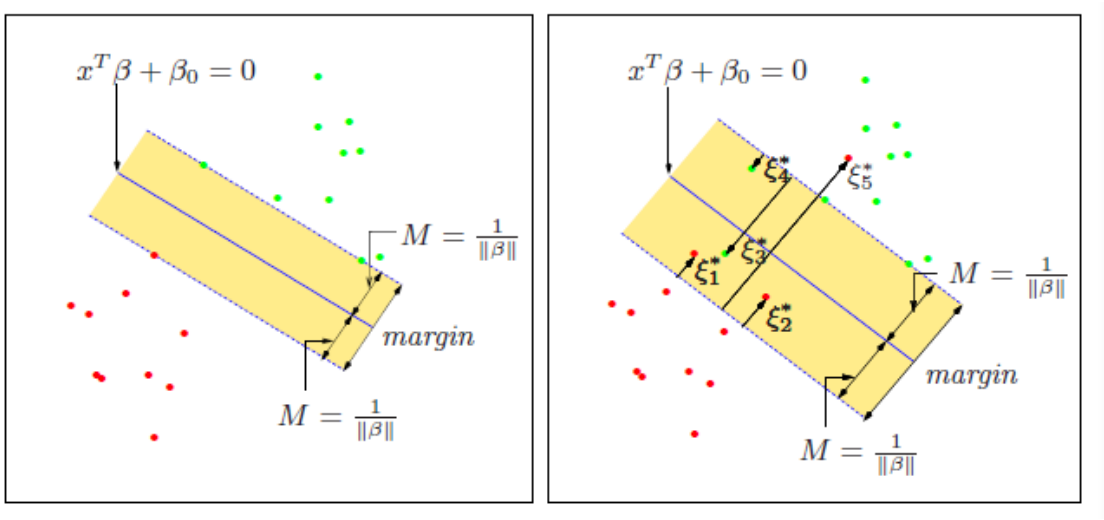
Or

$$y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i)$$

για κάθε i , $\xi_i \geq 0$, $\sum \xi i \leq constant.$

Οι δύο τρόπο οδηγούν σε διαφορετικές λύσεις.

Για την πρώτη περίπτωση έχουμε ότι η τιμή ξ_i στον περιορισμό είναι ο αναλογικός όρος με τον οποίο η πρόβλεψη είναι στη λάθος μεριά του περιθωρίου. Έτσι με το να οριοθετηθεί το άθροισμα των $\xi_ι$ οριοθετούμε το συνολικό αναλογικό σύνολο με το οποίο οι προβλέψεις πέφτουν στη λάθος μεριά του περιθωρίου. Λανθασμένες κατηγοριοποιήσεις συμβαίνουν όταν $\xi_i > 1$, έτσι η οριοθέτηση του αθροίσματος σε μια τιμή Κ, οριοθετεί τις συνολικές λανθασμένες κατηγοριοποιήσεις της μάθησης σε πλήθος Κ.



Εικόνα 5. 1 Support Vector Classifiers

Ορίζοντας M $= \frac{1}{||\beta||}$ τότε το πρόβλημα βελτιστοποίησης γίνεται: min||β|| με

$$y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \forall i,$$
$$\xi_i \geq 0, \sum \xi_i \leq \text{constant}$$

Αυτός είναι ο συνηθέστερος τρόπος που ορίζεται το Support Vector Classifier για την μη διαχωρίσιμη περίπτωση. Από τη φύση του κριτηρίου λοιπόν, βλέπουμε πως στο όριο κάθε κλάσης δεν λαμβάνονται πολύ υπ΄ όψιν τα σημεία εντός της κλάσης. Αυτή η διαφορά σε σχέση με άλλες γραμμικές μεθόδους είναι πολύ σημαντική.

Η λύση του προβλήματος βελτιστοποίησης λύνεται με την μέθοδο Lagrange που δεν αναλύεται περαιτέρω σε αυτήν την διπλωματική.

Το support vector classifier που αναπτύχθηκε παραπάνω βρίσκει μόνο γραμμικά σύνορα στον χώρο των εισόδων. Όπως άλλες γραμμικές μέθοδοι, μπορούμε να κάνουμε αυτή την διαδικασία πιο ευέλικτη μεγαλώνοντας τον χώρο των εισόδων με την χρήση επέκτασης βάσης. Γενικά τα γραμμικά σύνορα πετυχαίνουν καλύτερο διαχωρισμό στον μεγαλύτερο χώρο σε σχέση με τα μη γραμμικά σύνορα.

Η τεχνική support vector machine είναι μια επέκταση αυτής της ιδέας, όπου η διάσταση που μεγαλύτερου χώρου επιτρέπεται να είναι πολύ μεγάλη, άπειρη σε κάποιες περιπτώσεις. Ίσως φαίνεται ότι οι υπολογισμοί είναι απαγορευτικοί ή ότι με τις επεκτάσεις να παρουσιαστεί overfitting. Η SVM λύνει ένα πρόβλημα function-fitting μέσω ενός συγκεκριμένου κριτηρίου κανονικοποίησης.

5.3 Support Vector Machines για κατηγοριοποίηση και Kernels

Παραπάνω παρουσιάστηκε το πρόβλημα βελτιστοποίησης. Τώρα η συνάρτηση βελτιστοποίησης με μετασχηματισμένες τις μεταβλητές εισόδου σε $h(x_i)$, με την χρήση Lagrange είναι:

$$L_D = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} a_i a_{i'} y_i y_{i'} \langle h(x_i), h(x_{i'}) \rangle$$

Η λύση μπορεί να γραφτεί ως εξής:

$$f(x) = h(x)^T \beta + \beta_0 = \sum_{i=1}^{N} a_i y_i \langle h(x), h(x_i) \rangle + \beta_0$$

Η συνάρτηση h(x) εμπλέκεται μόνο ως εσωτερικό αποτέλεσμα οπότε χρειάζεται να γνωρίζουμε μόνο την συνάρτηση kernel K(x,x') και όχι την h(x). Στην βιβλιογραφία υπάρχουν τέσσερεις δημοφιλείς συνάρτησης kernel, οι εξής:

$$K(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} \mathbf{X}_i \cdot \mathbf{X}_j & \text{Linear} \\ (\gamma \mathbf{X}_i \cdot \mathbf{X}_j + C)^d & \text{Polynomial} \\ \exp(-\gamma |\mathbf{X}_i - \mathbf{X}_j|^2) & \text{RBF} \\ \tanh(\gamma \mathbf{X}_i \cdot \mathbf{X}_j + C) & \text{Sigmoid} \end{cases}$$

5.4 Support Vector Machines για Παλινδρόμηση

Αρχικά μελετούμε το γραμμικό μοντέλο

$$f(x) = x^T\beta + \beta_0,$$

και στην συνέχεια μελετάμε μη γραμμικές γενικεύσεις. Για να βρούμε το β, λαμβάνουμε υπ΄οψιν την ελαχιστοποίηση του:

$$H(\beta,\beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2}||\beta||^2,$$

Με $V_\varepsilon(r) = 0$, αν $|r| < \varepsilon$ ή $|r| - \varepsilon$, αλλιώς

Πρόκειται για «ε-insensitive» μέτρηση λάθους, αδιαφορώντας για σφάλματα με μέγεθος μικρότερα από ε. Μπορεί να παρομοιαστεί με το σύνορο στο support vector classification, όπου τα σημεία στην σωστή πλευρά και μακρυά από αυτό δεν λαμβάνονται υπ΄οψιν. Στην παλινδρόμιση, αυτά τα σημεία με μικρό σφάλμα μετρούν λιγότερο.

Αν $\hat{\beta}$, $\hat{\beta}_0$ ελαχιστοποιούν την H, η συνάρτηση λύση μπορεί να έχει την μορφή

$$\hat{\beta} = \sum_{i=1}^{N}(\hat{a}_i^* - \hat{a}_i)x_i$$

$$\hat{f}(x) = \sum_{i=1}^{N}(\hat{a}_i^* - \hat{a}_i)\langle x, x_i\rangle + \beta_0,$$

Με $\hat{a}_i$, $\hat{a}_i^*$ θετικοί και λύνουν το προγραμματιστικό πρόβλημα:

$$\min_{a_i, a_i^*} \sum_{i=1}^{N}(a_i^* + a_i) - \sum_{i=1}^{N} y_i(a_i^* + a_i) + \frac{1}{2}\sum_{i,i'=1}^{N}(a_i^* - a_i)(a_{i'}^* - a_{i'})\langle x_i, x_{i'}\rangle$$

Με δεδομένα τα:

$$0 \le a_i, a_i^* \le 1/\lambda,$$
$$\sum_{i=1}^{N}(a_i^* - a_i) = 0,$$
$$a_i a_i^* = 0.$$

Εξαιτίας της φύσης αυτών των περιορισμών, μόνο ένα υποσύνολο της λύσης είναι μη μηδενικό και οι σχετικές τιμές δεδομένων ονομάζονται «support vectors»

Η παράμετρο ε σχετίζεται με την συνάρτηση σφάλματος (loss function) και η λ είναι μια κλασική παράμετρο κανονικοποίησης και μπορεί βρεθεί π.χ. κατά την διαδικασία cross-validation.

<u>Παλινδρόμηση και Kernels</u>

Έστω ότι θεωρούμε την προσέγγιση της παλινδρόμησης με συνάρτηση βάσης $\{h_m(x)\}$, m=1,2…, M:

$$f(x) = \sum_{m=1}^{M}\beta_m h_m(x) + \beta_0$$

Για να βρούμε το β και το β_0, ελαχιστοποιούμε το

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2}\sum \beta_m{}^2$$

για κάποια γενική μέτρηση σφάλματος V(r). Για κάποιο επιλογή V(r), η λύση

$$f(x) = \sum \beta_m h_m(x) + \beta_0$$

έχει την μορφή

$$f(x) = \sum_{i=1}^{N} a_i K(x, x_i)$$

Με $K(x,y) = \sum_{m=1}^{M} h_m(x)h_m(y)$.

Εδώ δουλεύουμε με την περίπτωση όπου $V(r) = r^2$. Έστω **H** ο πίνακας NxM βάσης με το i-οστό στοιχείο $h_m(x_i)$ και M > N είναι μεγάλο. Για απλότητα θεωρούμε $\beta_0 = 0$ ή σταθερό που απορροφάται στο h. Προσεγγίζουμε το β ελαχιστοποιώντας το παρακάτω κριτήριο ελαχίστων τετραγώνων:

$$H(\beta) = (y - H\beta)^{\mathrm{T}}(y - H\beta) + \lambda \| \beta \|^2$$

Η λύση είναι

$$y = H\beta$$

Και το $\hat{\beta}$ ορίζεται από:

$$-H^{\mathrm{T}}(y - H\beta) + \lambda\beta = 0$$

Από αυτό φαίνεται ότι χρειάζεται να αξιολογηθεί ο πίνακας MxM στο μετασχηματισμένο χώρο. Ωστόσο μπορούμε να πολλαπλασιάσουμε από πριν με **H** και προκύπτει:

$$H\beta = (HH^T + \lambda I)^{-1} HH^T y$$

Ο πίνακας $\mathbf{HH}^{\mathrm{T}}$, που είναι διαστάσεων NxN αποτελείται από εσωτερικά αποτελέσματα μεταξύ ζευγαριών παρατηρήσεων i , i ‘. Πρόκειται δηλαδή, για την αξιολόγηση του «inner product kernel» $\{\mathbf{HH}^{\mathbf{T}}\}_{i,\,i'} = K(x_i, x_{i'})$. Είναι εύκολο να δείξουμε, σε αυτήν την περίπτωση ότι οι προβλεπόμενες τιμές για ένα αυθαίρετο x ικανοποιούν:

$$f(x) = h(x)^T \beta = \sum_{i=1}^{N} a_i K(x, x_i)$$

Με $\hat{\alpha} = (\mathbf{HH}^{\mathbf{T}} + \lambda\mathbf{I})^{-1}y$. Όπως και στα προηγούμενα του support vector machine, δεν χρειάζεται να συγκεκριμενοποιήσουμε ή να αξιολογήσουμε το μεγάλο σύνολο των συναρτήσεων $h_1(x),\ldots, h_M(x)$. Μόνο το εσωτερικό αποτέλεσμα του kernel $K(x_i, x_{i'})$ χρειάζεται να αξιολογηθεί, σε N σημεία εκπαίδευσης για κάθε i, i ‘ , και στα σημεία x για τις προβλέψεις εκεί. Σημείωση ότι η επιλογή του h_m σημαίνει ότι ο $\mathbf{HH}^{\mathrm{T}}$ μπορεί να υπολογιστεί με κόστος $N^2/2$ αξιολογήσεις του K και όχι με άμεσο κόστος N^2M. Επίσης η επιλογή τετραγωνικής νόρμας $\|\beta\|^2$ στο σφάλμα έχει εξάρτηση στην προηγούμενη ιδιότητα. Αν π.χ. έχουμε |β| ίσως προκύψει ανώτερο μοντέλο. Τα kernel στην παλινδρόμηση ορίζονται όπως και στην κατηγοριοποίηση.

Η τεχνική support vector machine για classification με πολλές κλάσεις μπορεί να προκύψει από την λύση πολλών προβλημάτων με δύο κλάσεις. Σε άλλη περίπτωση μπορεί να οριστεί πολλαπλή συνάρτηση σφάλματος, με κατάλληλο kernel.

ΚΕΦΑΛΑΙΟ 6 ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

6.1 Γενική περιγραφή της πειραματικής διαδικασίας

Αρχικά δόθηκε ένα αρχείο που περιέχει, για 410 μετοχές του δείκτη S&P 500 του χρηματιστηρίου της Νέας Υόρκης, 950 τιμές κλεισίματος. Από αυτά τα δεδομένα, οι πρώτες 600 μέρες χρησιμοποιήθηκαν για να πραγματοποιηθεί η διαδικάσια της μάθησης, τα επόμενα 100 για την διαδικασία «επικύρωσης» (validation) του καθενός μοντέλου και τα υπόλοιπα για την χάραξη στρατηγικών με βάση το ήδη επικυρωμένο μοντέλο.

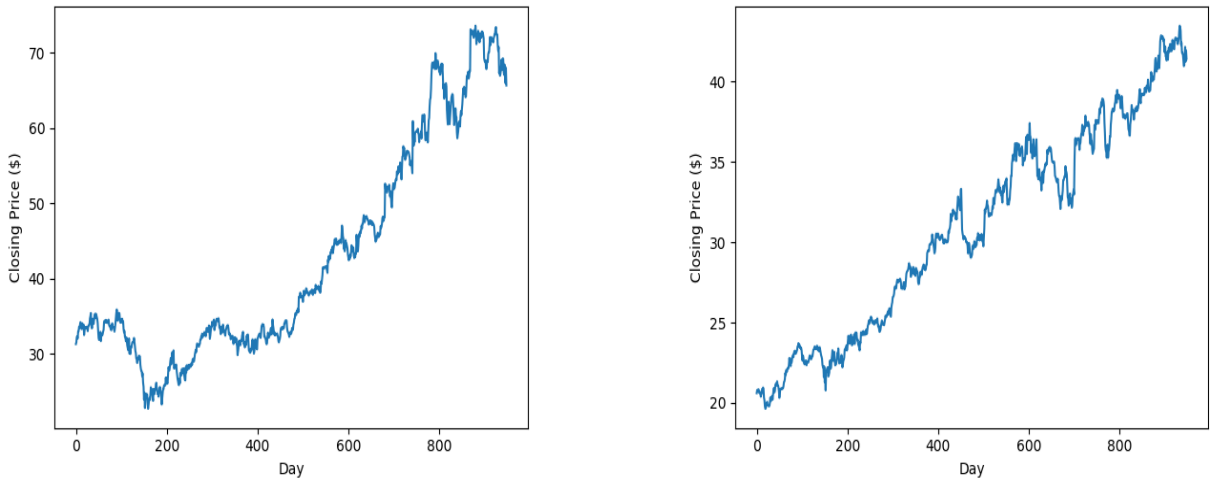
Στα πειράματα έγινε χρήση των παρακάτω αλγορίθμων

- Νευρωνικά Δίκτυα
- Δένδρα Απόφασης (CART)
- Support Vector Machine
- Random Forest

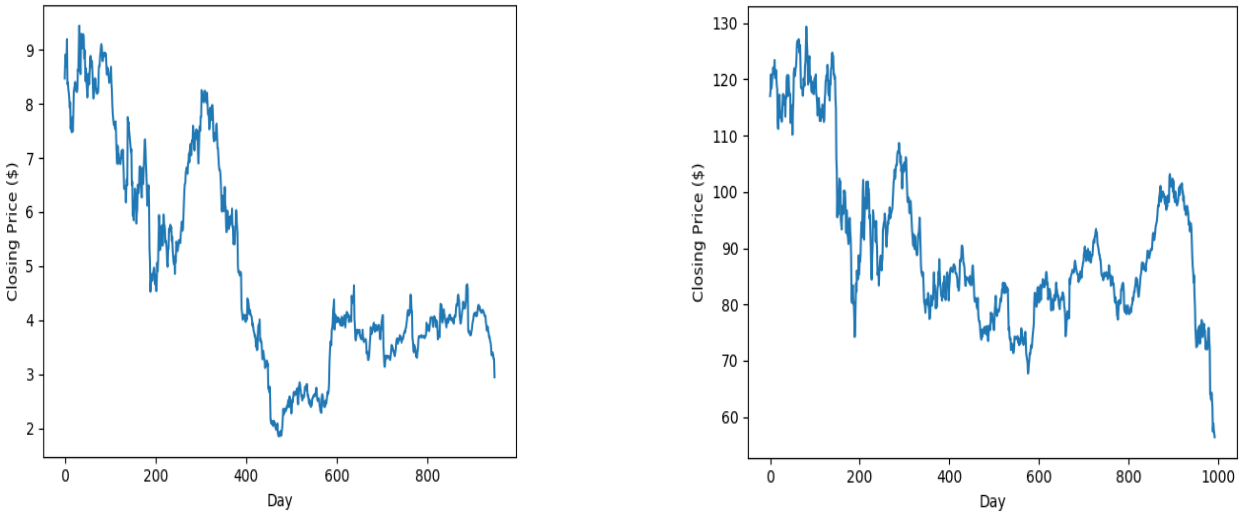
Οι υλοποιήσεις των παραπάνω τεχνικών μηχανικής μάθησης αποτελούν εργαλεία του scikit-learn μιας πλατφόρμας ανοιχτού κώδικα που σχετίζεται με την μηχανική μάθηση και γενικότερα την ανάλυση δεδομένων (data analysis). Επίσης το scikit-learn χρησιμοποιεί την γλώσσα προγραμματισμού Python και επομένως σε αυτή αναπτύχθηκε ο απαιτούμενος κώδικας για την πειραματική διαδικασία.

6.2 Περιγραφή των αρχικών δεδομένων dataset

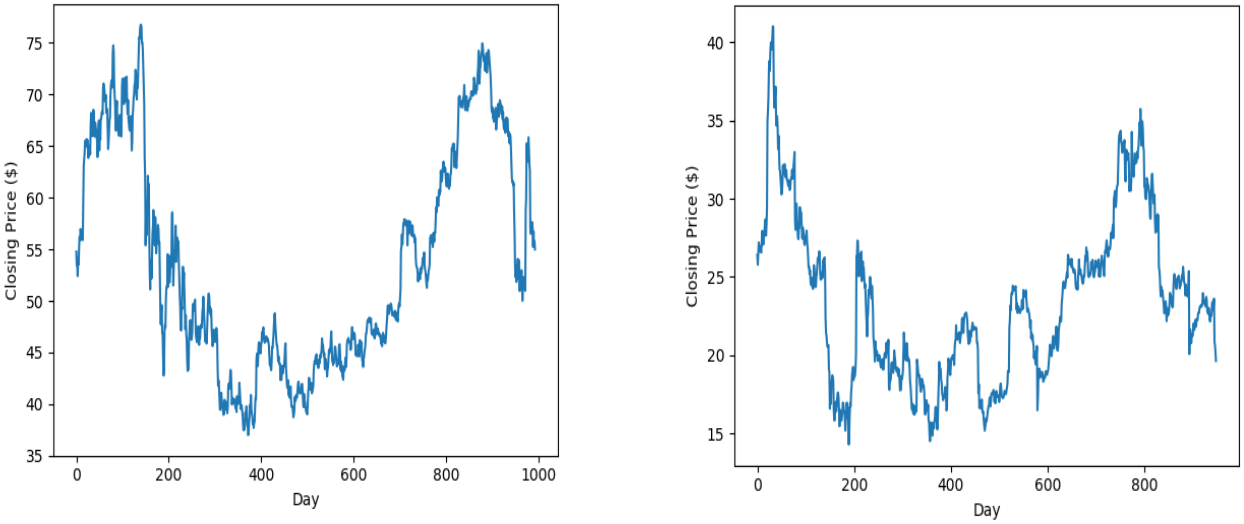
Αρχικά το dataset που δόθηκε περιέχει τιμές κλεισίματος 410 μετοχών του δείκτη S&P 500 του χρηματιστηρίου της Νέας Υόρκης, 950 τιμές κλεισίματος. Από αυτά τα δεδομένα, οι πρώτες 600 μέρες χρησιμοποιήθηκαν για να πραγματοποιηθεί η διαδικασία της μάθησης, τα επόμενα 100 για την διαδικασία «επικύρωσης» (validation) του καθενός μοντέλου και τα υπόλοιπα για την χάραξη στρατηγικών με βάση το ήδη επικυρωμένο μοντέλο.. Στο δείγμα υπήρχαν μετοχές με θόρυβο, με ανοδική τάση και με καθοδική τάση. Παρακάτω φαίνονται ενδεικτικά οι χρονοσειρές κάποιων μετοχών.



Εικόνα 6. 1 Μετοχές του dataset με ανοδική τάση



Εικόνα 6. 2 Μετοχές του dataset με καθοδική τάση



Εικόνα 6. 3 Μετοχές του dataset με θόρυβο

6.3 Τρόπος λήψης των μετρήσεων

Παρακάτω ακολουθούν τα αποτελέσματα των μετρήσεων για κάθε τεχνική μηχανικής μάθησης. Όλες οι μετρήσεις έγιναν με τον τρόπο της iterative prediction, δηλαδή γινόταν πρόβλεψη για μία ημέρα μπροστά, στην συνέχεια με την χρήση και αυτής της πρόβλεψης γινόταν πρόβλεψη για την επομένη ημέρα κ.ο.κ. . Επίσης για το validate του μοντέλου, δηλαδή για να επιλεχθούν οι κατάλληλες παράμετροι κάθε φορά ώστε να ελαχιστοποιείται το σφάλμα (sMAPE), έγιναν οι εξής περίοδοι προβλέψεων:

- 3 περίοδοι-3 ημέρες ορίζοντας
- 14 περίοδοι-7 ημέρες ορίζοντας
- 7 περίοδοι-14 ημέρες ορίζοντας
- 3 περίοδοι-30 ημέρες ορίζοντας

6.4 Μετρήσεις με Νευρωνικά Δίκτυα

Στον επόμενο πίνακα φαίνονται τα αποτελέσματα των μετρήσεων με τη χρήση νευρωνικών δικτύων, για διάφορες τιμές χρονικού ορίζονται πρόβλεψης και εισόδων του NN (παλαιότερες τιμές κλεισίματος).

ΟΡΙΖΟΝΤΑΣ ΠΡΟΒΛΕΨΗΣ	**3**	**7**	**14**	**30**
ΠΛΗΘΟΣ ΕΙΣΟΔΩΝ	***SMAPE***	***SMAPE***	***SMAPE***	***SMAPE***
1	1.98	2.89	3.99	7.17
3	1.98	2.91	4.00	7.22
7	1.97	2.89	3.98	6.93
14	1.98	2.93	4.00	7.13
30	1.98	2.98	4.01	6.79

Πίνακας 6. 1 Αποτελέσματα Μετρήσεων με ΤΝΔ

Σε αυτή την περίπτωση το NN είχε τα εξής χαρακτηριστικά:

- (hidden_layer_sizes=((past_used * 2). Το hiddev_layer_sizes είναι ο αριθμός των νευρώνων του «εσωτερικού στρώματος»
- activation='identity'. Στην επιλογή αυτή διαλέγουμε συνάρτηση ενεργοποίησης.
- solver='lbfgs'. Ο solver που χρησιμοποιείται για την βελτιστοποίηση των βαρών όπως αυτά ορίζονται στο κεφάλαιο ανάλυσης του ANN.
- max_iter=500. Εδώ διαλέγουμε τον μέγιστο αριθμό των επαναλήψεων μέχρι να επιτευχθεί η σύγκλιση.

Με την αλλαγή των επαναλήψεων (max_iter) δεν παρατηρήθηκε αξιοσημείωτη διαφορά. Επίσης ο solver adam δεν χρησιμοποιήθηκε σύμφωνα με τις οδηγίες του scikit-learn για datasets σαν και το προκείμενο.Στην συνέχεια έγινε διερεύνηση αν αξίζει να προστεθεί νέο layer στο NN, οπότε θα μας δινόταν η δυνατότητα και για χρήση μη γραμμικής συνάρτησης ενεργοποίησης. Έγινε έλεγχος για ορίζοντα 7 ημερών:

ΠΛΗΘΟΣ ΕΙΣΟΔΩΝ	***SMAPE***
1	3,88
3	3,96
7	4,02
14	4,18
30	4,44

Πίνακας 6. 2 Αποτελέσματα μετρήσεων με ΤΝΔ (deep learning)

Σε αυτήν την περίπτωση το NN είχε τα εξής χαρακτηριστικά:

- hidden_layer_sizes=((past_used * 2),3)

- activation='logistic'
- solver='lbfgs'
- max_iter=500

Καθώς το σφάλμα στην περίπτωση του deep learning βρέθηκε μεγαλύτερο από την απλή περίπτωση δεν έγινε περαιτέρω διερεύνηση για περισσότερα layers ή άλλες μη γραμμικές συναρτήσεις ενεργοποίησης.

6.5 Μετρήσεις με την μέθοδο CART

Στην μέθοδο CART έγιναν δοκιμές έχοντας ως εισόδους 5,9 και 12 προηγούμενες τιμές και παρατηρήθηκε ότι δεν έχει σημασία πόσες θα μπουν αρχικά. Οπότε έγινε ανάλυση με μόνη παράμετρο το βάθος του δένδρου, όπως αυτό ορίστηκε στο αντίστοιχο κεφάλαιο. Τα αποτελέσματα φαίνονται παρακάτω

Ορίζοντας πρόβλεψης		**3**	**7**	**14**	**30**	**Βάθος=3**
		2,92	3,54	4,40	6,42	
Ορίζοντας πρόβλεψης		**3**	**7**	**14**	**30**	**Βάθος=5**
		2,17	2,97	3,83	6,23	
Ορίζοντας πρόβλεψης		**3**	**7**	**14**	**30**	**Βάθος=7**
		2,39	3,22	4,07	6,34	
Ορίζοντας πρόβλεψης		**3**	**7**	**14**	**30**	**Βάθος=10**
		2,66	3,60	4,57	6,59	

Πίνακας 6. 3 Αποτελέσματα μετρήσεων με CART

6.6 Μετρήσεις με την μέθοδο SVM

Σε αυτήν την μέθοδο εξετάστηκε μόνο το γραμμικό kernel. Έγιναν πειράματα και με

kernel =’rbf’ ωστόσο τα σφάλματα ήταν πολύ μεγαλύτερα που είναι ανούσιο να αναφερθούν και να διερευνηθούν αυτές οι περιπτώσεις.

Παρακάτω φαίνονται τα αποτελέσματα για τις μετρήσεις που έγιναν:

Ορίζοντας πρόβλεψης	**3**	**7**	**14**	**30**
Είσοδοι				
1	1.98	2.86	3.92	6.89
3	1.99	2.87	3.93	6.92
7	1.97	2.86	3.93	6.72
14	2.00	2.87	3.97	6.88
30	2.01	2.88	3.97	6.88

Πίνακας 6. 4 Αποτελέσματα μετρήσεων με SVM

6.7 Μετρήσεις με την μέθοδο Random Forest

Σε αυτήν την μέθοδο, αφού πήραμε ως δεδομένο το μέγιστο βάθος των δένδρων από τα προηγούμενα πειράματα, ως παράμετροι ήταν το πλήθος των δασών, και η ρύθμιση max_features. Έγιναν μετρήσεις για 100,500,1000 δάση και για 3,6,17 features. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα:

Ορίζοντας πρόβλεψης	**3**	**7**	**14**	**30**		**estimators=100**
max_feat						
3	2.34	3.25	4.32	6.13		
6	2.23	3.10497	4.12536	6.12		
12	2.18	3.0	3.93	6.13		
						estimators=500
3	2.34	3.24	4.33	6.08		
6	2.23	3.11	4.14	6.12		
12	2.18	2.99	3.94	6.14		
3	2.34	3.24	4.33	6.08		**estimators =1000**
6	2.23	3.11	4.14	6.12		
12	2.18	2.99	3.94	6.14		

Πίνακας 6. 5 Αποτελέσματα μετρήσεων με Random Forest

6.8 Προτεινόμενες Εναλλακτικές από την Βιβλιογραφία

Στην διεθνή Βιβλιογραφία υπάρχει πληθώρα εναλλακτιών και διάφορων τεχνικών μηχανικής μάθησης σχετικά με την πρόβλεψη τιμής μετοχών και κατα συνέπεια με τις αγοραπωλησίες

στα χρηματιστήρια ανά τον κόσμο. Επίσης υπάρχουν τεχνικές που αναπτύσσουν οι επενδυτικοί οίκοι οι οποίες όμως δεν είναι προσβάσιμες από εξωτερικούς παράγοντες. Σε αυτήν την ενότητα σκοπός είναι να αναφερθούν παραδείγματα διαφορετικών τεχνικών, να γίνει περίληψη της λειτουργίας τους και να εντοπιστούν ομοιότητες και διαφορές με τις τεχνικές που αναπτύχθηκαν στην παρούσα διπλωματική εργασία.

- Εναλλακτική 1:

Αρχικά η χρονοσειρά της τιμής περνά από εκθετική εξομάλυνση. Στην συνέχεια δημιουργούνται διάφοροι τεχνικοί δείκτες ,όπως ο δείκτης «On balance volume» που εκτιμά την τιμή της μετοχής στο μέλλον με βάση τις αλλαγές στον όγκο συναλλαγών της μετοχής, και αυτοί οι δείκτες αποτελούν τις εισόδους σε έναν tree-based classifier. Ο classifier σε αυτήν την περίπτωση είναι ο Random-forest και Gradient boosted decision trees. Ο χρονικός ορίζοντας για την διαδικασία των συναλλαγών και κατά συνέπεια την αξιολόγηση του μοντέλου ποικίλει από 3 έως και 30 ημέρες. Γενικά τα αποτελέσματα αυτής της διαδικασίας δίνουν σε όλες τις περιπτώσεις καλύτερη ακρίβεια σε σχέση με την γραμμική παλινδρόμηση αλλά και τα τεχνητά νευρωνικά δίκτυα. Συμπερασματικά στην παραπάνω διαδικασία ο σκοπός δεν ήταν η τιμή στο μέλλον αλλά το αν θα αυξηθεί ή θα μειωθεί και επιπρόσθετα οι εισόδου του μοντέλου μηχανικής μάθησης ήταν διαφορετικές και πιο πολύπλοκες σε σχέση με το μοντέλο που χρησιμοποιήθηκε στην διπλωματική εργασία. (Suryoday Basak, Saibal Kar, Snehanshu Saha, Luckyson Khaidem, Sudeepa Roy Dey, 2018)

- Εναλλακτική 2:

Σε αυτή την εργασία γίνεται χρήση της τεχνικής SVM προκειμένου να προβλεφθεί την κατεύθυνση της τιμής στο μέλλον. Και σε αυτό το παράδειγμα ως είσοδοι στο μοντέλο δεν χρησιμοποιούνται οι τιμές των μετοχών αλλά διάφοροι δείκτες. Ορισμένοι απο αυτούς είναι το «Momentum» που ορίζεται ως η αλλαγή της τιμής σε βάθος ενός συγκεκριμένου χρονικού ορίζοντα, ο δείκτης «ROC» που ορίζεται ως η διαφορά της παρούσας τιμής με μία ορισμένη παλαιότερη τιμή, καθώς και άλλη πιο περίπλοκοι όπως ο δείκτης «%D» και «%K». Και σε αυτήν την περίπτωση οι συγγραφείς ορίζουν το πρόβλημα ως ένα πρόβλημα classification παρά regression. Έτσι οι έξοδοι του συστήματος είναι «0» ή «1» που το καθένα αντίστοιχα δείχνει αν θα πρέπει να αγοραστεί η μετοχή ή όχι. Αυτή είναι και η κύρια διαφορά με την οπτική που ακολουθεί η παρούσα διπλωματική. (Kyoung-jae Kim, 2013)

- Εναλλακτική 3:

Στην εργασία των Hiransha M. , Gopalakrishnan E.A. , Vijay Krishna Menon, Soman K.P βρίσκουμε μια εφαρμογή της μηχανικής μάθησης και συγκεκριμένα του deep-learning η οποία είναι παρόμοια με αυτή της παρούσας διπλωματικής σχετικά με τις εισόδους και τον τρόπο που γίνεται λήψη των μετρήσεων. Γίνεται χρήση των τεχνικών MLP, RNN, LSTM, CNN που είναι παραλλαγές των απλών νευρωνικών δικτύων που περιγράφηκαν στην διπλωματική. Η εκπαίδευση έγινε μόνο για μια

μετοχή του χρηματιστηριακού δείκτη της Ινδίας (NSE) αλλά το μοντέλο χρησιμοποιήθηκε για να προβλεφθούν οι τιμές των μετοχών πέντε διαφορετικών εταιρειών του NYSE και του NSE. Παρατηρήθηκε ότι καλύτερα αποτελέσματα είχε το μοντέλο CNN αλλά κυρίως πως και οι δύο αγορές μοιράζονται κάποιου είδους κοινή δυναμική. Τέλος τα αποτελέσματα συγκρίθηκαν με αυτά του ARIMA model και παρατηρήθηκε πώς τα νευρωνικά δίκτυα παράγουν μικρότερο σφάλμα. (Hiransha M. , Gopalakrishnan E.A. , Vijay Krishna Menon, Soman K.P , 2018)

- Εναλλακτική 4

Σε αυτήν την εργασία ως μεταβλητές εισόδου και εξόδου στο μοντέλο SVR τοποθετούνται δείκτες ΤΑ. Οι ΤΑ δείκτες είναι αποτέλεσμα επεξεργασίας προηγούμενων τιμών μετοχών αφού έχουν περάσει από κάποια συνάρτηση. Ο απλούστερος δείκτες τέτοιου είδους είναι και ο κινούμενος μέσος πλήθους Τ παλιών τιμών κλεισίματος. Η ανάλυση έγινε με αυτόν τον δείκτη ανάλογα το χρονικό πλαίσιο που γινόταν κάθε φορά η πρόβλεψη, από λεπτά έως μέρες. Τα αποτελέσματα έδειξαν πως το SVR έχει προβλεπτική ικανότητα σε σχέση με απλούστερα μοντέλα. Τέλος αξίζει να σημειωθεί πως οι μετρήσεις έδειξαν πως το γραμμικό kernel δίνει την καλύτερη δυνατή ακρίβεια, κάτι που παρατηρήθηκε και στην διπλωματική εργασία. (Bruno Miranda Henrique, Vinicius Amorim Sobreiro, Herbert Kimura, 2018)

- Εναλλακτική 5

Και σε αυτήν την περίπτωση γίνεται χρήση των απλών τεχνικών δεικτών που χρησιμοποιήθηκαν και παραπάνω. Ωστόσο η διαφορά της εργασίας αυτής έγκειται στο ότι τα δεδομένα «περνούν» από δύο μοντέλα μηχανικής μάθησης ώστε στο τέλος να προκύψει η τελική πρόβλεψη. Το πρώτο είναι ένα SVR όπου γίνεται προετοιμασία για την είσοδο στο δεύτερο στάδιο, το οποίο είναι είτε SVR, είτε Random Forest, είτε ANN. Αρχικά προβλέπονται οι τεχνικοί δείκτες για την ημέρα που επιθυμείται να παραχθεί πρόβλεψη και στην συνέχεια αυτός ο προβλεπόμενος δείκτης εισέρχεται στο δεύτερο στάδιο και εν τέλει προκύπτει η τελική πρόβλεψη για την τιμή κλεισίματος. Με αυτόν τον τρόπο το μοντέλο πρόβλεψης στο δεύτερο στάδιο έχει ως είσοδο δεδομένα που αφορούν την t+n (n ο ορίζοντας) ημέρα και όχι την t που συμβαίνει στο πρώτο στάδιο και εν γένει στα μοντέλα που περιγράφηκαν παραπάνω. Αυτή είναι και η ουσιώδης διαφορά από όλες τις προαναφερθείσες εργασίες.(Jigar Patel, Sahil Shah, Priyank Thakkar, K Kotecha, 2015)

ΚΕΦΑΛΑΙΟ 7 ΧΑΡΤΟΦΥΛΑΚΙΑ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

7.1 Χαρτοφυλάκια

Με βάση τα προηγούμενα ευρήματα και χρησιμοποιώντας πάντα την βέλτιστη τεχνική μηχανικής μάθησης κάθε φορά αναπτυχθεί διάφορα χαρτοφυλάκια και διερευνήθηκε η κερδοφορία καθενός από αυτά. Έγιναν πειράματα με ορίζοντα επτά ημέρες και τριάντα ημέρες.

Η διαδικασία είναι η εξής:

Για τις μέρες 700 έως την τελευταία ημέρα που έχουμε στα χέρια μας κατασκευάστηκαν χαρτοφυλάκια όπου οι γίνονταν, με βάση τις προβλεπόμενες αποδόσεις που προέκυψαν από το SVM (kernel=linear, no_features=7). Συγκεκριμένα το SVM έδινε τις προβλεπόμενες τιμές στο τέλος της περιόδου 7 ημερών και το εργαλείο αγόραζε τις μετοχές με την μεγαλύτερη απόδοση στο διάστημα αυτό. Όταν ερχόταν η επόμενη περίοδος όλα τα χρήματα επανεπενδύονταν με τον ίδιο τρόπο. Το αρχικό κεφάλαιο θεωρήθηκε 100000$.

Παρακάτω δίνεται ένα παράδειγμα για τις πρώτες δύο περιόδους αγοράζοντας κάθε φορά τις δέκα μετοχές με την μεγαλύτερη προβλεπόμενη απόδοση.

Πρώτη περίοδος

TOP 10	Αρχική τιμή	Προβλεπόμενη	Πραγματική	Πλήθος μετοχών	Αξία	Υπόλοιπο	Τελική αξία
4.053	58.54	60.91	56.07	170	9951.8	48.2	9531.9
3.214	17.07	17.62	16.55	585	9985.95	14.05	9681.75
1.975	22.37	22.81	21.62	447	9999.39	0.61	9664.14
1.829	40.86	41.61	42.14	244	9969.84	30.16	10282.16
1.670	48	48.8	48.68	208	9984	16	10125.44
1.523	12.21	12.39	12.29	819	9999.99	0.01	10065.51
1.408	34	34.48	33.1	294	9996	4	9731.4
1.354	7.64	7.74	7.68	1308	9993.12	6.88	10045.44
1.297	8.28	8.38	8.25	1207	9993.96	6.04	9957.75

1.291	46.28	46.87	46.29	216	9996.48	3.52	9998.64

Starting Cash = $ 100000

Ending Cash = $ 99213.6

Δεύτερη περίοδος

TOP 10	Αρχική Τιμή	Προβλεπόμενη	Πραγματική	Πλήθος μετοχών	Αξία	Υπόλοιπο	Τελική αξία
3.377	46.64	48.23	48.44	212	9887.68	33.62	10269.28
3.102	56.07	57.81	56.82	176	9868.32	52.98	10000.32
2.464	16.55	16.96	17.58	599	9913.45	7.85	10530.42
2.161	42.14	43.05	41.92	235	9902.9	18.4	9851.2
2.080	8.25	8.42	8.48	1202	9916.5	4.8	10192.96
1.940	31.52	32.13	32.37	314	9897.28	24.02	10164.18
1.750	43.42	44.18	44	228	9899.76	21.54	10032
1.746	88.73	90.28	89	111	9849.03	72.27	9879
1.698	33.1	33.66	33.18	299	9902.88	24.4	9920.82
1.693	73.86	75.11	78.34	134	9873.12	24.06	10497.56

Πίνακας 7. 1 Παράδειγμα Διαδικασίας Συναλλαγής

Πίνακας 7. 2 Παράδειγμα Διαδικασίας Συναλλαγής

Starting Cash = $ 99213.6

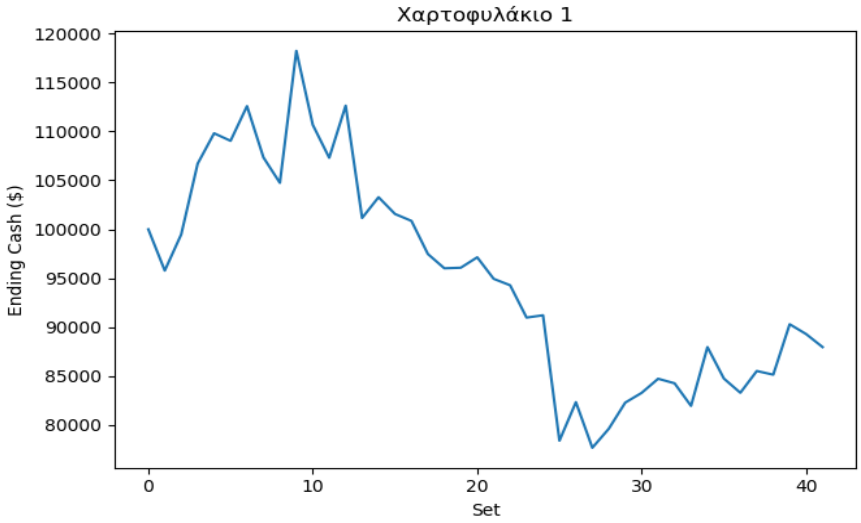
Ending Cash = $ 101622.27

7.2 Χαρτοφυλάκια για ορίζοντα 7 ημερών

Η παραπάνω διαδικασία έγινε για 41 περιόδους των 7 ημερών, όσο δηλαδή μας επέτρεπαν τα δεδομένα μας. Παρακάτω φαίνονται οι αποτιμήσεις των χαρτοφυλακίων στην αρχή και στο τέλος των 41 περιόδων καθώς και η τελική απόδοση.

1ο Χαρτοφυλάκιο

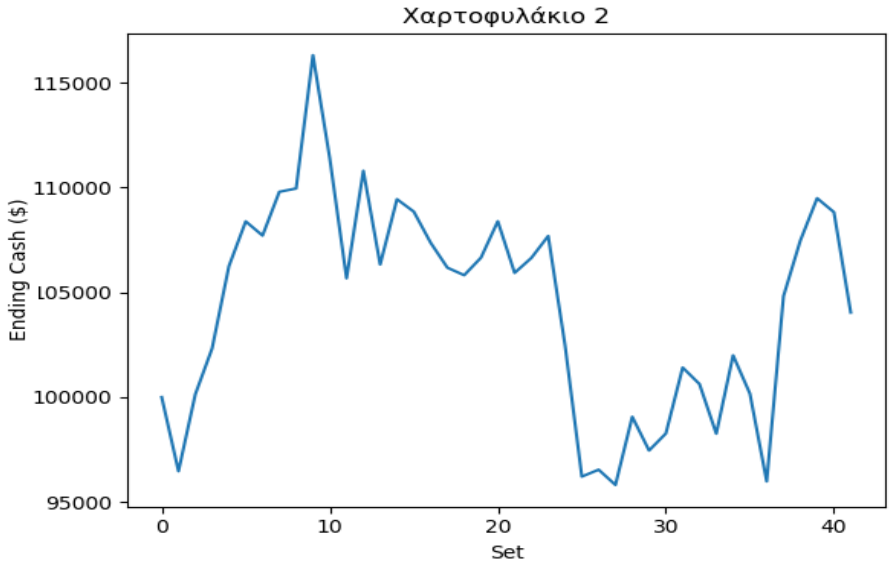
Σε αυτό το χαρτοφυλάκιο κάθε φορά αγοράζουμε την μετοχή με την μεγαλύτερη προβλεπόμενη απόδοση. Προέκυψαν, στο τέλος των περιόδων, $ 87996,24 δηλαδή η απόδοση του χαρτοφυλακίου ήταν -12,03%.



Εικόνα 7. 1 Τελικά μετρητά σε κάθε περίοδο(7 ημ.) μετρήσεων για το Χαρτ. 1

2ο Χαρτοφυλάκιο

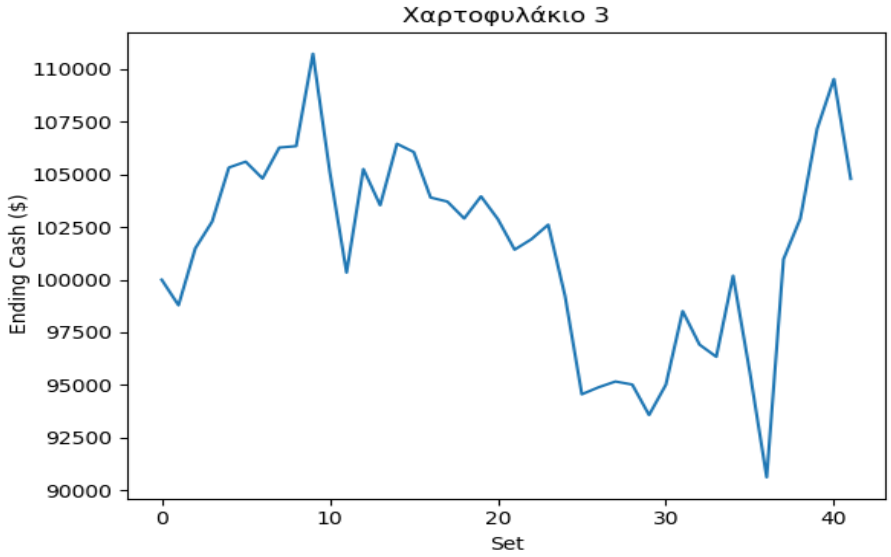
Σε αυτό το χαρτοφυλάκιο κάθε φορά αγοράζουμε τις τρεις μετοχές με τις μεγαλύτερες προβλεπόμενες αποδόσεις, αντίστοιχα. Προέκυψαν, στο τέλος των περιόδων, $ 104035,75 δηλαδή η απόδοση ήταν 4.04%.



Εικόνα 7. 2 Τελικά μετρητά σε κάθε περίοδο(7 ημ.) μετρήσεων για το Χαρτ. 2

3ο Χαρτοφυλάκιο

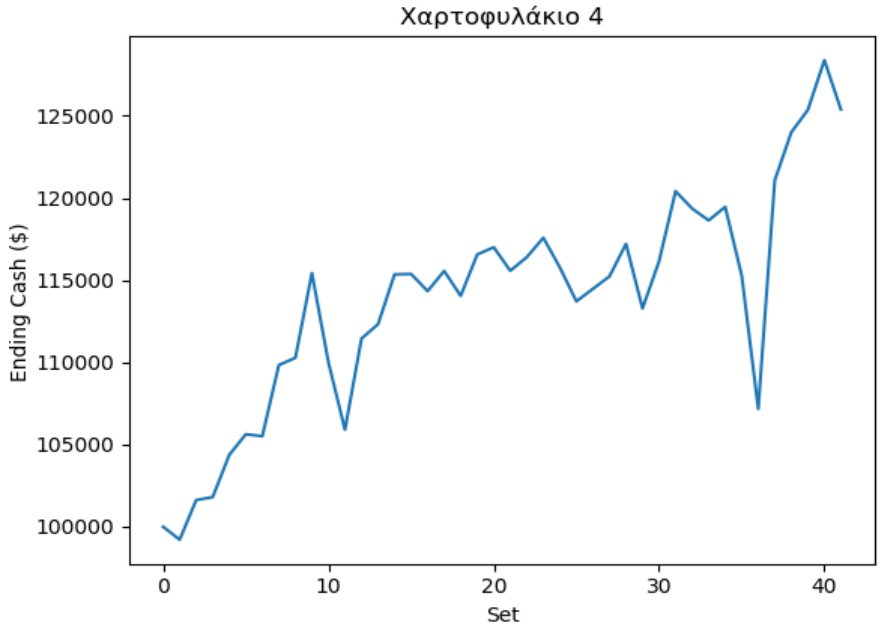
Σε αυτό το χαρτοφυλάκιο κάθε φορά αγοράζουμε τις πέντε μετοχές με τις μεγαλύτερες προβλεπόμενες αποδόσεις, αντίστοιχα. Προέκυψαν, στο τέλος των περιόδων, $ 104787,35 δηλαδή η απόδοση ήταν 4,87%.



Εικόνα 7. 3 Τελικά μετρητά σε κάθε περίοδο(7 ημ.) μετρήσεων για το Χαρτ. 3

4ο Χαρτοφυλάκιο

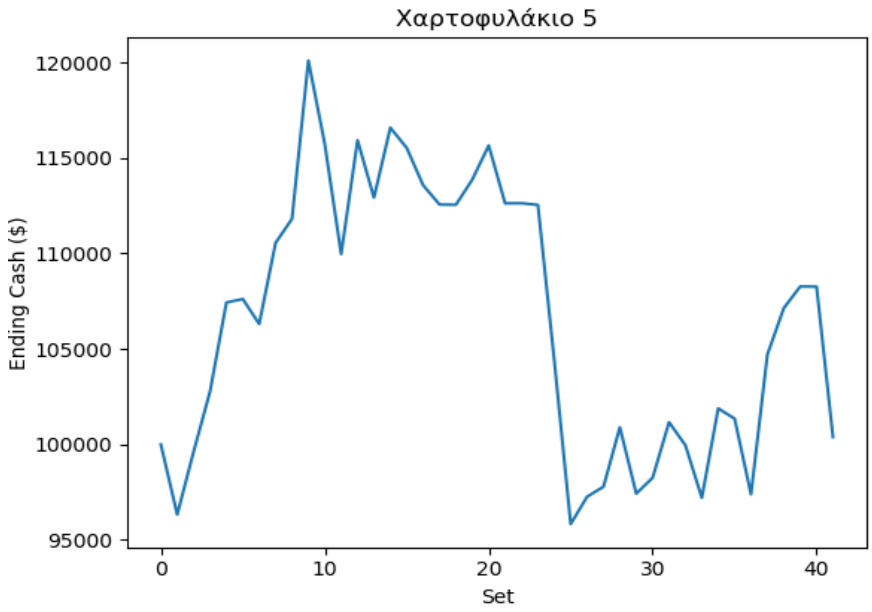
Σε αυτό το χαρτοφυλάκιο κάθε φορά αγοράζουμε τις δέκα μετοχές με τις μεγαλύτερες προβλεπόμενες αποδόσεις, αντίστοιχα. Προέκυψαν, στο τέλος των περιόδων, $ 125385,41 δηλαδή η απόδοση ήταν 25,39%.



Εικόνα 7. 4 Τελικά μετρητά σε κάθε περίοδο(7 ημ.) μετρήσεων για το Χαρτ. 4

5ο Χαρτοφυλάκιο

Σε αυτό το χαρτοφυλάκιο κάθε φορά αγοράζουμε τις τρεις μετοχές με τις μεγαλύτερες προβλεπόμενες αποδόσεις ωστόσο χρησιμοποιούμε και βάρη για το πόσο που επενδύεται στην κάθε μία. Το βάρος για την μετοχή με την μεγαλύτερη απόδοση είναι: w_1 = 55%, για την μετοχή με την αμέσως μεγαλύτερη απόδοση w_2 = 30% ενώ για την τρίτη μετοχή είναι w_3 = 15%. Προέκυψαν, στο τέλος των περιόδων, $ 100383,31 δηλαδή η απόδοση ήταν 0,383%.



Εικόνα 7. 5 Τελικά μετρητά σε κάθε περίοδο(7 ημ.) μετρήσεων για το Χαρτ. 5

Συγκεντρωτικά τα αποτελέσματα για τα διάφορα χαρτοφυλάκια φαίνονται στον παρακάτω πίνακα.

Χαρτοφυλάκιο No.	Τελικά Μετρητά ($)	Απόδοση(%)
1	87966.24	-12.03
2	104035.75	4.04
3	104787.35	4.87
4	125385.41	25.39
5	100383.31	0.383

Πίνακας 7. 3 Απόδοση Χαρτοφυλακίων με ορίζοντα 7 ημερών

Στα προηγούμενα το κόστος συναλλαγής δεν λήφθηκε υπ΄ όψιν. Ωστόσο πρέπει να συμπεριληφθεί. Σύμφωνα με την ιστοσελίδα της αμερικάνικης εταιρείας που δραστηριοποιείται στο χώρο των χρηματοπιστωτικών υπηρεσιών προκύπτει ότι η συγκεκριμένη εταιρεία χρεώνει τους πελάτες της για κάθε συναλλαγή μετοχών το

ποσό των $ 4,95 ανά συναλλαγή. Με βάση αυτή την χρέωση προκύπτει ο παρακάτω επικαιροποιημένος πίνακας για τα τελικά χρήματα και την τελική απόδοση.

Χαρτοφυλάκιο No.	Τελικά Μετρητά ($)	Απόδοση(%)
1	87763.29	-12.24
2	103426.9	3.43
3	103772.6	3.73
4	123335.91	23.33
5	99774.46	-0,998

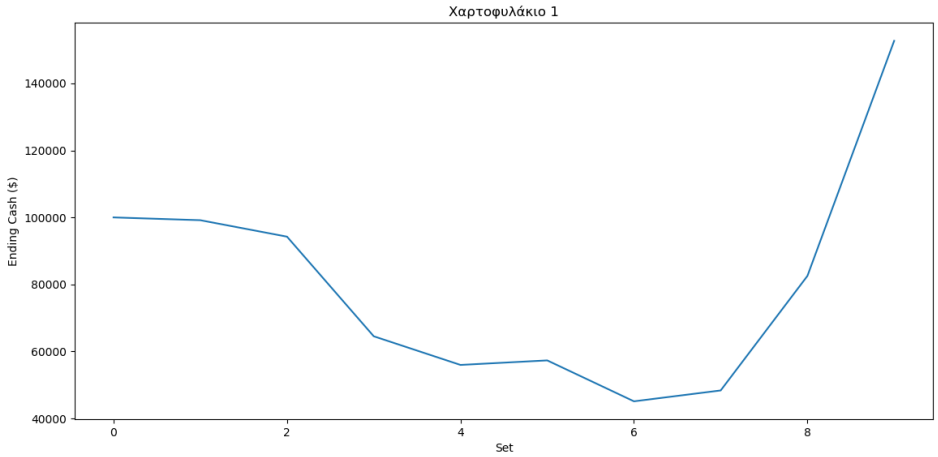
Πίνακας 7. 4 Απόδοση Χαρτοφυλακίων 7 ημερών μετά την αφαίρεση του κόστους συναλλαγής

7.3 Χαρτοφυλάκια με ορίζοντα 30 ημερών

Για τις μέρες 700 έως την τελευταία ημέρα που έχουμε στα χέρια μας κατασκευάστηκαν χαρτοφυλάκια όπου οι γίνονταν, με βάση τις προβλεπόμενες αποδόσεις που προέκυψαν από την τεχνική Random Forest (max_depth=5,max_features=3,n_estimators=500). Αντίστοιχα με την στρατηγική για 7 ημέρες ορίζοντα κατασκευάστηκαν τα εξής χαρτοφυλάκια, με ορίζοντα 30 ημέρες. Όλες οι υπόλοιπες παράμετροι παρέμειναν ίδιες, δηλαδή τα αρχικά μετρητά και το ποσό επανεπένδυσης. Έγιναν συνολικά 9 set μετρήσεων και κατασκευάστηκαν χαρτοφυλάκια με το ίδιο τρόπο με πριν.

1ο Χαρτοφυλάκιο

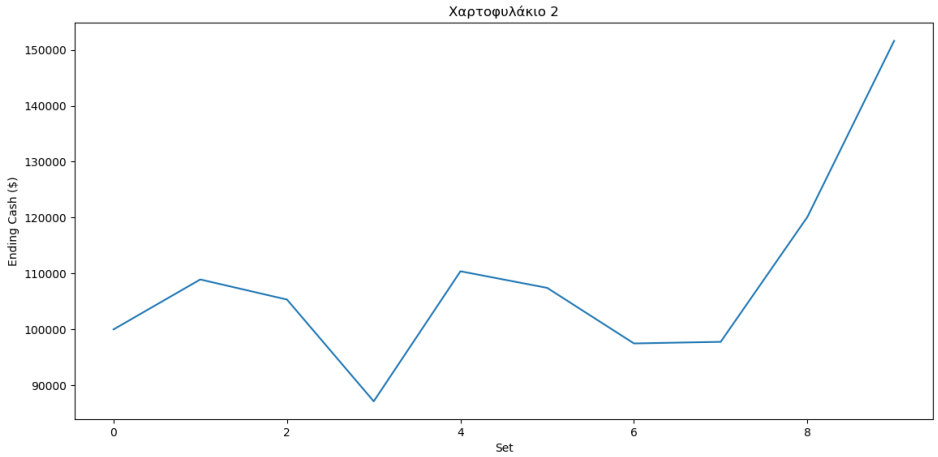
Σε αυτό το χαρτοφυλάκιο κάθε φορά αγοράζουμε την μετοχή με την μεγαλύτερη προβλεπόμενη απόδοση. Προέκυψαν, στο τέλος των περιόδων, $ 152704,3 δηλαδή η απόδοση του χαρτοφυλακίου ήταν 52,7%.



Εικόνα 7. 6 Τελικά μετρητά σε κάθε περίοδο(30 ημ.) μετρήσεων για το Χαρτ. 1

2ο Χαρτοφυλάκιο

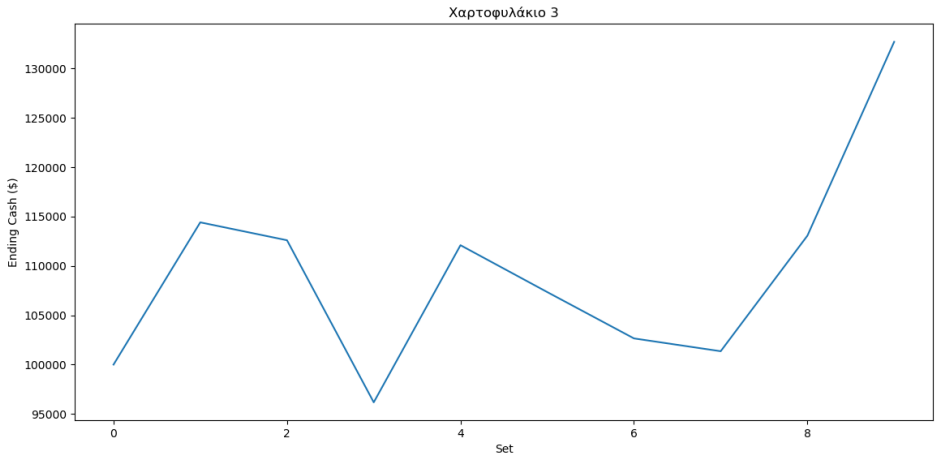
Σε αυτό το χαρτοφυλάκιο κάθε φορά αγοράζουμε τις τρεις μετοχές με τις μεγαλύτερες προβλεπόμενες αποδόσεις, αντίστοιχα. Προέκυψαν, στο τέλος των περιόδων, $ 151052,48 δηλαδή η απόδοση ήταν 51,05%.



Εικόνα 7. 7 Τελικά μετρητά σε κάθε περίοδο(30 ημ.) μετρήσεων για το Χαρτ. 2

3ο Χαρτοφυλάκιο

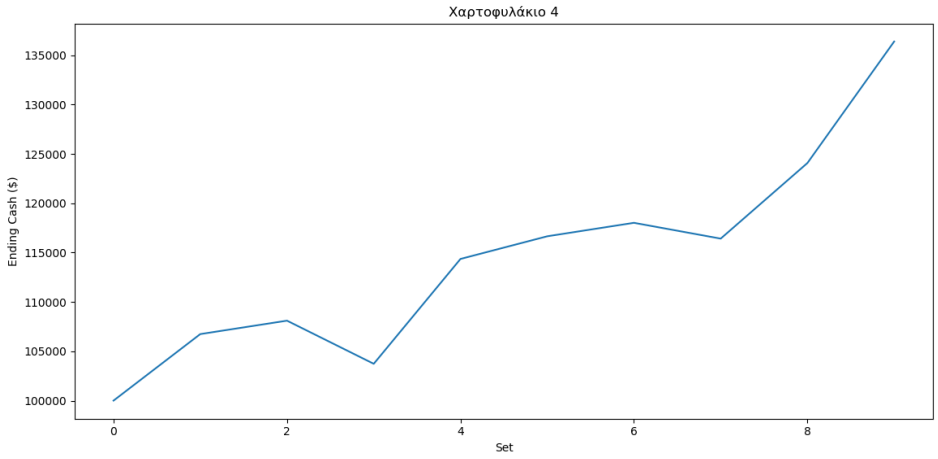
Σε αυτό το χαρτοφυλάκιο κάθε φορά αγοράζουμε τις πέντε μετοχές με τις μεγαλύτερες προβλεπόμενες αποδόσεις, αντίστοιχα. Προέκυψαν, στο τέλος των περιόδων, $ 132712.41 δηλαδή η απόδοση ήταν 32,27%.



Εικόνα 7. 8 Τελικά μετρητά σε κάθε περίοδο(30 ημ.) μετρήσεων για το Χαρτ. 3

4ο Χαρτοφυλάκιο

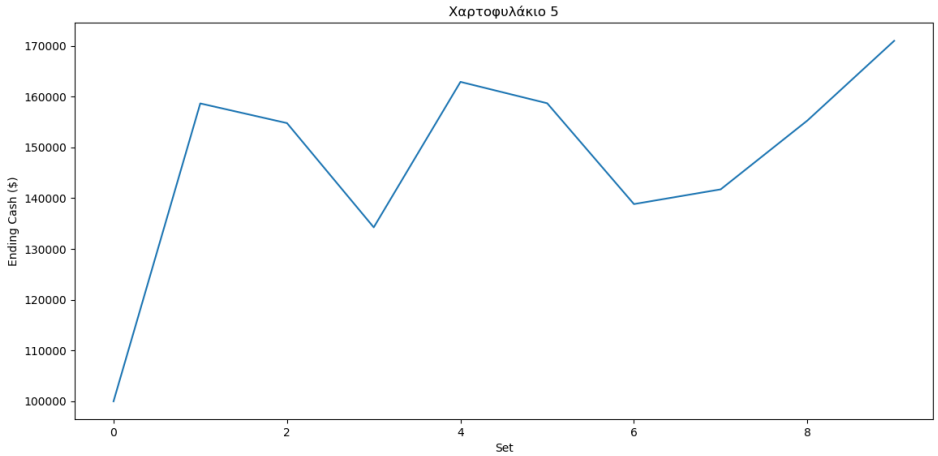
Σε αυτό το χαρτοφυλάκιο κάθε φορά αγοράζουμε τις δέκα μετοχές με τις μεγαλύτερες προβλεπόμενες αποδόσεις, αντίστοιχα. Προέκυψαν, στο τέλος των περιόδων, $ 136355,36 δηλαδή η απόδοση ήταν 36,36%.



Εικόνα 7. 9 Τελικά μετρητά σε κάθε περίοδο(30 ημ.) μετρήσεων για το Χαρτ. 4

5ο Χαρτοφυλάκιο

Σε αυτό το χαρτοφυλάκιο κάθε φορά αγοράζουμε τις τρεις μετοχές με τις μεγαλύτερες προβλεπόμενες αποδόσεις ωστόσο χρησιμοποιούμε και βάρη για το πόσο που επενδύεται στην κάθε μία. Το βάρος για την μετοχή με την μεγαλύτερη απόδοση είναι: w_1 = 55%, για την μετοχή με την αμέσως μεγαλύτερη απόδοση w_2 = 30% ενώ για την τρίτη μετοχή είναι w_3 = 15%. Προέκυψαν, στο τέλος των περιόδων, $ 170983.79 δηλαδή η απόδοση ήταν 70,98%



Εικόνα 7. 10 Τελικά μετρητά σε κάθε περίοδο(30 ημ.) μετρήσεων για το Χαρτ. 5

.

Συγκεντρωτικά τα αποτελέσματα για τα διάφορα χαρτοφυλάκια φαίνονται στον παρακάτω πίνακα.

Χαρτοφυλάκιο No.	Τελικά Μετρητά ($)	Απόδοση(%)
1	152704.3	52.7
2	151052,48	51,02
3	132712.41	32.71
4	136355.36	36.36
5	170.353.57	70.98

Πίνακας 7. 5 Απόδοση Χαρτοφυλακίων με ορίζοντα 30 ημερών

Όπως και στην στρατηγική των 7 ημερών έτσι και εδώ θα πρέπει να αφαιρέσουμε την προμήθεια, δηλαδή το κόστος συναλλαγής. Για την τιμολόγηση ισχύουν ότι και στην προηγούμενη περίπτωση. Έτσι έχουμε τον παρακάτω πίνακα για Τελικά Μετρητά και Απόδοση:

Χαρτοφυλάκιο No.	Τελικά Μετρητά ($)	Απόδοση(%)
1	150529.3	50.53
2	149365.48	49.37
3	131578.41	31.57
4	134987.36	34.99
5	170019.92	70.19

Πίνακας 7. 6 Απόδοση Χαρτοφυλακίων με ορίζοντα 30 ημερών μετά την αφαίρεση του κόστους συναλλαγής

7.4 Συμπεράσματα Διπλωματικής με βάση τα Χαρτοφυλάκια

Αρχικά παρατηρούμε πως στα Χαρτοφυλάκια με ορίζοντα τις 30 ημέρες. τα κέρδη ειναι υψηλότερα από αυτά με ορίζοντα 7 ημέρες. Επίσης παρατηρούμε πως στα περισσότερα από αυτά τα χαρτοφυλάκια 30 ημερών, στο όγδοο και στο ένατο σετ μετρήσεων έχουμε μεγάλη αύξηση του κέρδους. Αυτό οφείλεται στις δύο τελευταίες επιλογές μετοχών που έγιναν. Για παράδειγμα, για το Χαρτοφυλάκιο 1 παρατίθενται οι παρακάτω αγοραπωλησίες που φαίνεται το μεγάλο κέρδος (σχεδόν διπλασιασμός) στις τελευταίες δύο περιόδους.

Αρχική Τιμή($)	Τελική Τιμή ($)
25.28	25.07
25.28	25.07
17.96	17.07
7.64	5.23
2.49	2.16
33.07	33.87
1.22	0.96
39.99	42.87
0.58	0.99
23.14	42.81

Πίνακας 7. 7 Παράδειγμα αρχικής και τελικής τιμής σε αγοραπωλησία μετοχών

Σε αυτό το σημείο πρέπει να παρατηρήσουμε πως για αυτές τις δύο μετοχές (όπως και σε κάθε παρόμοια περίπτωση με την μέθοδο Random Forest) το σύστημα μας προέβλεπε μεγάλη αύξηση καθώς σε όλη την προηγούμενη περίοδο (δηλαδή αυτή που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου) η τιμή της μετοχής ήταν σημαντικά υψηλότερη. Δηλαδή το Random Forest έδινε μεγάλη προβλεπόμενη απόδοση σε μια μετοχή που ήταν χαμηλότερα σε σχέση με το προηγούμενο

διάστημα. Με αυτόν τον τρόπο τα χαρτοφυλάκια χτίστηκαν, κατά κάποιον τρόπο ποντάροντας σε πιθανό «rebound».

Εν συνεχεία πρέπει να παρατηρήσουμε πως σε ένα χαρτοφυλάκιο με ορίζοντα τις 30 ημέρες υπήρξε περίοδος που το μετρητά είχαν φτάσει τις 40000$, δηλαδή ζημία 60000$. Αυτό δεν συνέβει σε κανένα χαρτοφυλάκιο με ορίζοντα τις 7 ημέρες. Κατά συνέπεια μπορούμε να πούμε ότι ο μικρότερος κίνδυνος βρίσκεται στα χαρτοφυλάκια με ορίζοντα τις 7 ημέρες όπου δεν παρατηρήθηκε ζημιά μεγαλύτερη από τις περίπου 20000$.

Κανένα χαρτοφυλάκιο με ορίζοντα 30 ημερών δεν βγήκε ζημιογόνο και σε όλα προέκυψαν υψηλές αποδόσεις. Ακόμα, στην περίοδο που έγιναν οι μετρήσεις οι περισσότερες μετοχές είχαν ανοδική πορεία. Έτσι μπορούμε να πούμε πως θα συνέφερε κάποιον επενδυτή να δημιουργεί χαρτοφυλάκια με μεγαλύτερο ορίζοντα επένδυσης. Επίσης με αυτόν τον τρόπο τα κόστη συναλλαγής μειώνονται. Αυτό το γεγονός θα έχει μεγαλύτερη σημασία για χαρτοφυλάκια λίγων χιλιάδων δολαρίων.

Τέλος, παρατηρούμε ότι στο Χαρτοφυλάκιο 1 της στρατηγικής των 7 ημερών, όπου διαλέγουμε μόνο μία μετοχή για να γίνει επένδυση παρατηρείται ζημία. Επιπρόσθετα και στο Χαρτοφυλάκιο 1 της στρατηγικής των 30 ημερών, όπου και πάλι διαλέγουμε μία μετοχή για επένδυση, αν εξαιρεθεί η τελευταία περίοδος έχουμε και εκεί ζημία σε όλες τις προηγούμενες περιόδους. Συμπερασματικά η στρατηγική να γίνεται επένδυση σε μία μόνο μετοχή πρέπει να αποφεύγεται από τους επενδυτές.

7.5 Ευκαιρίες για επιπρόσθετη βελτίωση

Τα κέρδη με βάση το σύστημα που αναπτύχθηκε είναι μεγάλο ωστόσο το γεγονός αυτό δεν πρέπει να θεωρηθεί ως δεδομένο μιας και με αναλυτική παρατήρηση των συναλλαγών φάνηκε η τύχη να έπαιξε μεγάλο ρόλο (όπως εξηγείται στο προηγούμενο υποκεφάλαιο). Είναι αναγκαίο λοιπόν να αναφερθούν νέες ιδέες για ανάπτυξη και περεταίρω βελτίωση του αλγορίθμου. Μοντέλα σαν αυτό που κατασκευάστηκε μπορούν να χρησιμοποιηθούν από επαγγελματίες του χρηματοπιστωτικού χώρου, αλλά και ερασιτέχνες που κατέχουν όμως ένα επαρκές επίπεδο κατανόησης των τεχνικών προβλέψεων και της μηχανικής μάθησης ώστε να κάνουν μια γρήγορη ανάλυση και πρόβλεψη με βάση τις προηγούμενες τιμές των μετοχών. Με αυτόν τον τρόπο έχουν την δυνατότητα να υποστηρίξουν καλύτερα τις αποφάσεις τους στο χρηματιστήριο λαμβάνοντας υπ΄ όψιν μόνο μαθηματικοποιημένες τεχνικές, δηλαδή χωρίς ανθρώπινη παρέμβαση. Επίσης το σύστημα αυτό αποτελεί μια πρώτη προσέγγιση σε αυτοματοποιημένες αγοραπωλησίες δίνοντας στον χρήστη μια γεύση του πως κάποιος μπορεί να χτίσει τέτοιους αλγορίθμους, που χρησιμοποιούν όλοι οι επενδυτικοί οίκοι. Ωστόσο το σύστημα χρήζει βελτίωσης.

Αρχικά, όπως αναφέρθηκε και στο σημείο που παρουσιάστηκαν παραδείγματα από την βιβλιογραφία, υπάρχει η δυνατότητα στα δεδομένα προς εκπαίδευση να μην περιλαμβάνονται μόνο οι τιμές κλεισίματος αλλά και άλλα δεδομένα. Τα δεδομένα αυτά μπορούν να είναι αρκετά απλά έως και πολύπλοκοι τεχνικοί δείκτες που χρησιμοποιούνται από τους αναλυτές. Για παράδειγμα, ο όγκος συναλλαγών μια μετοχής παίζει πολύ σημαντικό ρόλο στην πρόβλεψη της μελλοντικής τιμής της από τους αναλυτές. Στην παρούσα μετοχή, το μέγεθος αυτό δεν λήφθηκε υπ΄ όψιν. Σε μια απλή προσπάθεια βελτίωσης, θα μπορούσαμε να ορίσουμε ως είσοδο στο σύστημα μας όχι μόνο την τιμή κλεισίματος αλλά και τον όγκο συναλλαγών, δηλαδή το διάνυσμα τους. Αντίστοιχα θα μπορούν να κατασκευαστούν παρόμοια διανύσματα που εμπλέκουν, επιπρόσθετα, την μέση τιμή της μετοχής στην διάρκεια της ημέρας και την τιμή ανοίγματος. Αρκετοί τεχνικοί δείκτες θα μπορούσαν να αποτελέσουν, επίσης την είσοδο της διαδικασίας εκπαίδευσης, οι οποίοι απεικονίζουν διάφορες καταστάσεις της μετοχής όπως π.χ. δείκτες που δείχνουν την ελάχιστη τιμή, το ρίσκο κ.ο.κ. . Υπάρχει πληθώρα τέτοιων δεικτών οι οποίοι μπορούν να βρεθούν εύκολα στο Διαδίκτυο και κάποιοι αναφέρθηκαν στο υποκεφάλαιο που έγινε περιγραφή παραδειγμάτων από την Βιβλιογραφία. Αυτό που αξίζει να τονιστεί είναι το γεγονός ότι ένα σύστημα μηχανικής μάθησης που θα έχει ως έξοδο όχι την προβλεπόμενη τιμή αλλά σύσταση προς τον επενδυτή για αγορά, πώληση ή αμετάβλητη θέση θα ήταν πιο ρεαλιστικό καθώς έτσι λειτουργούν τα περισσότερα παρόμοια σύστημα στον χρηματοπιστωτικό χώρο.

Ακόμα, υπάρχει η δυνατότητα να γίνουν δοκιμές και για άλλα είδη χαρτοφυλακίων και η κατασκευή αυτών των χαρτοφυλακίων να γίνει με συνδυασμό της πρόβλεψης και άλλων κριτηρίων. Για παράδειγμα, στην κατασκευή των χαρτοφυλακίων δεν λήφθηκε υπ΄ όψιν ο τομέας της οικονομίας στον οποίο δραστηριοποιείται κάθε εταιρεία (industry). Θα μπορούσαμε λοιπόν να διαρθρώσουμε χαρτοφυλάκια που αφορούν έναν τομέα της οικονομίας ή συνδυασμούς αυτών. Προσθέτοντας μετοχές πολλών τομέων της οικονομίας το ρίσκο της επένδυσης μειώνεται καθώς υπάρχει διαφοροποίηση. Ακόμα, θα μπορούσαμε να υπολογίσουμε τις συσχετίσεις μετοχών και να προσθέσουμε στο χαρτοφυλάκια ασυσχέτιστες μεταξύ τους μετοχές. Επίσης, έχουμε την δυνατότητα, αν σκοπός είναι η κατασκευή ενός συντηρητικού χαρτοφυλακίου, να υπολογίσουμε την τυπική απόκλιση κάθε μετοχής με τα ιστορικά στοιχεία που έχουμε και στην συνέχεια να αποκλείσουμε από το σύστημα εκείνες τις μετοχές που παρουσιάζουν μεγάλες διακυμάνσεις στην τιμή τους (δηλαδή υψηλό ρίσκο). Για περισσότερη ανάπτυξη, όσον αφορά τα χαρτοφυλάκια, πρέπει να ανατρέξουμε σε προχωρημένες τεχνικές κατασκευής χαρτοφυλακίων συνυπολογίζοντας τα αποτελέσματα του προβλεπτικού μοντέλου που κατασκευάσαμε.

Τέλος, σε κάθε περίπτωση θα ήταν χρήσιμο να γίνουν δοκιμές με περισσότερους ορίζοντες πρόβλεψης. Θα πρέπει να προσέξουμε εδώ καθώς το κόστος συναλλαγής μπορεί να μεγαλώσει αρκετά αν γίνονται αγορές και πωλήσεις συχνά ενώ αν ο ορίζοντας επένδυσης είναι πολύ μεγάλος τότε ουσιαστικά το σύστημα δεν είναι τόσο παρεμβατικό οπότε σε περίοδο πτώσης δεν θα γίνουν διορθωτικές κινήσεις.

Βιβλιογραφία

[1] I . Vlahavas , P . Kefalas , N . Bassiliades , F . Kokkoras , I. Sakellariou. Artificial Intelligence – 3rd Edition. University of Macedonia Press, 2011.

[2] Φ. Πετρόπουλος, Β. Ασημακόπουλος. Επιχειρησιακές προβλέψεις. Συμμετρία, 2013.

[3] T. Hastie, R. Tibshirani, Jerome Friedman. The Elements of Statistical Learning – 2nd Edition. Springer, 2008.

[4] Suryoday Basak, Saibal Kar, Snehanshu Saha, Luckyson Khaidem, Sudeepa Roy Dey. Predicting the direction of stock market prices using tree-based classifiers, The North American Journal of Economics and Finance, Volume 47, Pages 552-567, 2019.

[5] Kyoung-jae Kim. Financial time series forecasting using support vector machines, Neurocomputing,Volume 55, Issues 1–2,Pages 307-319, 2013.

[6] Hiransha M. , Gopalakrishnan E.A. , Vijay Krishna Menon, Soman K.P . NSE Stock Market Prediction Using Deep-Learning Models, International Conference on Computional Intelligence and Data Science (ICCIDS 2018), 2018.

[7] Bruno Miranda Henrique, Vinicius Amorim Sobreiro, Herbert Kimura. Stock price prediction using support vector regression on daily and up to the minute prices, The Journal of Finance and Data Science 4 (2018) 183-201, 2018.

[8] Jigar Patel, Sahil Shah, Priyank Thakkar, K Kotecha. Predicting stock market index fusion of machine learning techniques, Expert Systems with Applications Volume 42, Issue 4, Pages (2162-2172), 2015