



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΙΩΑΝΝΙΝΩΝ

ΣΧΟΛΗ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΠΣ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ Τ.Ε.

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Η ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ
ΓΙΑ ΑΒΕΒΑΙΑ ΔΕΔΟΜΕΝΑ

Σπυρίδων Σιουράς

Επιβλέπων: Σταύρος Αδάμ
Επίκουρος Καθηγητής

Τόπος έκδοσης, Ιανουάριος, 2022

REGRESSION IN MACHINE LEARNING BASED ON UNCERTAIN DATA

Εγκρίθηκε από τριμελή εξεταστική επιτροπή

Τόπος, Ημερομηνία

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Επιβλέπων καθηγητής

Σταύρος Αδάμ,

Επίκουρος Καθηγητής

2. Μέλος επιτροπής

Γρηγόριος Δουμένης,

Επίκουρος Καθηγητής

3. Μέλος επιτροπής

Πέτρος Καρβέλης,

Επίκουρος Καθηγητής

© Σιουράς, Σπυρίδων, 2022.

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα πτυχιακή εργασία είναι εκ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Σιουράς, Σπυρίδων

Υπογραφή

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της πτυχιακής μου εργασίας θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που με στήριξαν καθ' όλη τη διάρκεια των σπουδών μου.

Πρώτα θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Σταύρο Αδάμ για την πολύτιμη βοήθεια και καθοδήγηση που μου παρείχε σε όλη την διάρκεια εκπόνησης της πτυχιακής εργασίας, αλλά και για την εμπιστοσύνη που μου έδειξε.

Επιπλέον, θα ήθελα να ευχαριστήσω τους γονείς μου και τον αδερφό μου για την βοήθεια και την συμπαράσταση που μου έδιναν, όχι μόνο κατά την διάρκεια των σπουδών, αλλά και γενικότερα.

Τέλος, θα ήθελα να ευχαριστήσω τους φίλους μου, που με στήριζαν σε όλη την διάρκεια των σπουδών.

ΠΕΡΙΛΗΨΗ

Τα αβέβαια δεδομένα είναι μεγάλο πρόβλημα στον κλάδο της μηχανικής μάθησης, καθώς κλασικές τεχνικές στατιστικής και μηχανικής μάθησης δεν μπορούν να διαχειριστούν τέτοιου είδους δεδομένα. Αβέβαια δεδομένα μπορούν να προκύψουν για διάφορους λόγους και υπάρχουν διαφορετικά είδη αβέβαιων δεδομένων. Η τεχνική αναπαράστασης και αντιμετώπισης των αβέβαιων δεδομένων, εξαρτάται από την μορφή της αβεβαιότητας. Στην παρούσα πτυχιακή εργασία γίνεται ανάλυση τεχνικών παλινδρόμησης σε αβέβαια δεδομένα, όπου θεωρείται γνωστό το εύρος στο οποίο βρίσκεται η πραγματική τιμή, αλλά όχι η μεμονωμένη τιμή. Η αβεβαιότητα θα αναπαρασταθεί με διαστήματα.

Σε αυτήν την εργασία αναλύεται αρχικά τι είναι η παλινδρόμηση, τι είναι η αβεβαιότητα στην μηχανική μάθηση και με τι τεχνικές μπορούν να αντιμετωπιστούν τα αβέβαια δεδομένα και, στο τέλος του πρώτου κεφαλαίου, γίνεται μια αναφορά στο τι είναι η ανάλυση διαστημάτων. Στη συνέχεια, γίνεται ανάλυση διάσημων κλασικών τεχνικών που χρησιμοποιούνται για την παλινδρόμηση και τεχνικών παλινδρόμησης που εφαρμόζονται σε δεδομένα σε μορφή διαστημάτων. Τέλος, χρησιμοποιούνται δύο τεχνικές παλινδρόμησης, η Lassoir και η τεχνική των k -Πλησιέστερων Γειτόνων (k -Nearest Neighbors), σε ένα ένα σύνολο δεδομένων που δείχνει την ζήτηση ηλεκτρικού ρεύματος στο δίκτυο διανομής ηλεκτρικής ενέργειας της Ισπανίας.

Λέξεις-κλειδιά: Παλινδρόμηση, Αβεβαιότητα, Αβέβαια δεδομένα, Διαστήματα.

ABSTRACT

Uncertain data is a major problem in Machine Learning, as classical statistical and machine learning techniques cannot manage such data. Uncertain data can occur for a variety of reasons and there are different types of uncertainties. Techniques and modelling of uncertain data depends on the type of uncertainty. In this dissertation, regression techniques for uncertain data are analyzed, where the range of the value is known, but the individual value is not known. Uncertain data will be modelled with intervals.

The dissertation starts with analyzing Regression as a term, what uncertainty in Machine Learning is, how uncertain data can be managed and an introduction to Interval Analysis. Moreover, classical regression techniques and interval regression techniques are analyzed. Finally, two regression techniques, Lassoir and K-Nearest Neighbors, are used on a dataset, that shows the demand for electricity in Spain.

Keywords: Regression, Uncertainty, Uncertain data, Intervals.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΕΥΧΑΡΙΣΤΙΕΣ	1
ΠΕΡΙΛΗΨΗ	2
ABSTRACT	3
ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ	4
ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ	6
ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ/ΕΙΚΟΝΩΝ	7
ΠΙΝΑΚΑΣ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ	9
1.Βασικές έννοιες	10
1.1 Παλινδρόμηση.....	10
1.2 Αβεβαιότητα στη Μηχανική Μάθηση	13
1.3 Αναπαράσταση αβέβαιων δεδομένων	14
1.3.1 Συμβολικά δεδομένα	14
1.3.2 Ασαφής προσέγγιση.....	15
1.3.3 Πιθανολογική προσέγγιση	16
1.4 Διαστήματα (Intervals).....	18
2.Τεχνικές Παλινδρόμησης	20
2.1 Κλασικές τεχνικές Παλινδρόμησης.....	20
2.1.1 Γραμμική Παλινδρόμηση	20
2.1.2 k-Πλησιέστεροι Γείτονες	21
2.1.3 Μηχανές Διανυσμάτων Υποστήριξης.....	22
2.1.4 Δέντρα απόφασης και Random Forests.....	22
2.1.5 Τεχνητά Νευρωνικά δίκτυα.....	24
2.2 Τεχνικές Παλινδρόμησης σε δεδομένα μορφής διαστημάτων	28
2.2.1 Μέθοδος κέντρου	28
2.2.2 Μέθοδος κέντρου και εύρους.....	29

2.2.3	Διμεταβλητή μέθοδος κέντρου και εύρους.....	30
2.2.4	Περιορισμένη μέθοδος	31
3.	Υλοποίηση.....	32
3.1	Σύνολο δεδομένων	32
3.2	Lassoir	37
3.3	Πλησιέστεροι Γείτονες.....	41
4.	Συμπεράσματα και Μελλοντική Εργασία	47
	BIBΛΙΟΓΡΑΦΙΑ	49

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 3.1 Ελάχιστες τιμές ζήτησης ανά ημέρα.....	32
Πίνακας 3.2 Μέγιστες τιμές ζήτησης ανά ημέρα.....	33
Πίνακας 3.3 Διαστήματα ζήτησης ανά ημέρα.....	34

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ/ΕΙΚΟΝΩΝ

Εικόνα1.1.Ευθεία παλινδρόμησης.....	1
Εικόνα1.2.Απεικόνιση σφάλματος σε απλό παράδειγμα Παλινδρόμησης.....	3
Εικόνα1.3. Δείγμα συνόλου δεδομένων απογραφής.....	5
Εικόνα2.1. Δομή απόφασης δέντρου.....	12
Εικόνα2.2. Παράδειγμα απόφασης δέντρου.....	13
Εικόνα2.3. Ένα πολυεπίπεδο Τεχνητό Νευρωνικό Δίκτυο.....	14
Εικόνα2.4.Ένα απλό Τεχνητό Νευρωνικό Δίκτυο.....	15
Εικόνα2.5. Λογιστική συνάρτηση.....	16
Εικόνα3.1. Δείγμα του συνόλου δεδομένων.....	32
Εικόνα 3.3. Διάγραμμα διασποράς των κάτω ορίων.....	36
Εικόνα3.3. Διάγραμμα διασποράς των άνω ορίων.....	36
Εικόνα3.4 Εκτιμήτριες για τα κέντρα.....	39
Εικόνα3.5. Εκτιμήτριες για τις ακτίνες.....	40
Εικόνα3.6 Συνολικά σφάλματα.....	40
Εικόνα 3.7 Οι προβλέψεις (κόκκινο χρώμα) σε σύγκριση με τα πραγματικά διαστήματα (μπλε χρώμα).....	41
Εικόνα 3.8. Εκτιμήτριες για τα κέντρα, με βάση την τεχνική πλησιέστερων γειτόνων.....	43
Εικόνα 3.9. Εκτιμήτριες για τις ακτίνες, με βάση την τεχνική πλησιέστερων γειτόνων.....	43
Εικόνα 3.10. Η πρόβλεψη (κόκκινο) του πρώτου στοιχείου σε σχέση με το πραγματικό διάστημα (μπλε).....	44

Εικόνα 3.11. Η πρόβλεψη (κόκκινο) του δεύτερου στοιχείου σε σχέση με το πραγματικό διάστημα

(μπλε).....44

Εικόνα 3.12. Η πρόβλεψη (κόκκινο) του τρίτου στοιχείου σε σχέση με το πραγματικό διάστημα

(μπλε).....45

Εικόνα 3.13. Η πρόβλεψη (κόκκινο) του τέταρτου στοιχείου σε σχέση με το πραγματικό διάστημα

(μπλε).....45

Εικόνα 3.14 Η πρόβλεψη (κόκκινο) του πέμπτου στοιχείου σε σχέση με το πραγματικό διάστημα

(μπλε).....46

Εικόνα 3.15. Η πρόβλεψη (κόκκινο) του έκτου στοιχείου σε σχέση με το πραγματικό διάστημα

(μπλε).....46

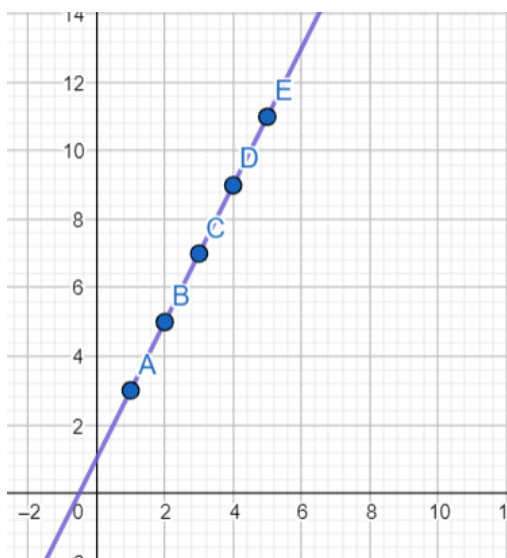
ΠΙΝΑΚΑΣ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ

MET.....	Μέθοδος Ελάχιστων Τετραγώνων
TNΔ.....	Τεχνητά Νευρωνικά Δίκτυα
CM.....	Center Method
CRM.....	Center and Range Method
BCRM.....	Bivariate Center and Range Method
CONCM.....	Constrained Center Method
CONCRM.....	Constrained Center and Range Method
MCS.....	Monte Carlo Simulation
PDF.....	Probability Density Function

1.Βασικές έννοιες

1.1 Παλινδρόμηση

Η ανάλυση παλινδρόμησης είναι μια στατιστική τεχνική του που ερευνά και μοντελοποιεί την σχέση μεταξύ μεταβλητών εισόδων (ανεξάρτητες τιμές) και μιας τιμής εξόδου (εξαρτημένη τιμή). (Montgomery, Peck and Vining, 2012)



Εικόνα 1.1. Ευθεία παλινδρόμησης.

Όπως φαίνεται στην παραπάνω Εικόνα 1.1 το μοντέλο που δημιουργεί η παλινδρόμηση είναι μια συνάρτηση, με την οποία μπορούμε να προβλέψουμε την τιμή της εξαρτημένης μεταβλητής, μέσω των τιμών της ανεξάρτητης μεταβλητής ή των ανεξάρτητων μεταβλητών. Η μορφή της συνάρτησης αλλάζει, με βάση ένα δείγμα των δεδομένων, προσαρμόζοντας στους συντελεστές της αναλυτικής μορφής της, μέχρις ότου η συνάρτηση να προβλέπει με ικανοποιητική ακρίβεια την εξαρτημένη τιμή. Η πιο διαδεδομένη μέθοδος που χρησιμοποιείται για την προσαρμογή των συντελεστών είναι η μέθοδος των ελάχιστων τετραγώνων (MET).

Όπως αναφέρθηκε παραπάνω, η παλινδρόμηση είναι η εύρεση μιας συνάρτησης

$$y \approx f(x_1, x_2, \dots, x_k; \theta),$$

η οποία να επεξηγεί την σχέση μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών. Σκοπός είναι η εύρεση των κατάλληλων συντελεστών, ώστε να προβλέπει με μεγάλη ακρίβεια την πραγματική τιμή του y . Μια συνάρτηση παλινδρόμησης είναι γραμμική όταν υπάρχει γραμμική σχέση μεταξύ των συντελεστών της συνάρτησης, χωρίς να επηρεάζεται από την μορφή των ανεξάρτητων μεταβλητών. Για παράδειγμα, η συνάρτηση $f(x, \beta) = \beta_1 + \beta_2 * \sin(x)$, ενώ είναι ημιτονική ως προς το x , είναι γραμμική ως προς τους συντελεστές β και, συνεπώς, η συνάρτηση παλινδρόμησης είναι γραμμική. Γενικά, μια συνάρτηση παλινδρόμησης είναι μη-γραμμική, όταν η μερική παράγωγος για οποιοδήποτε συντελεστή β_j δεν είναι σταθερή συνάρτηση. Παράδειγμα μη-γραμμικής συνάρτησης είναι η ακόλουθη:

$$f(x, \beta) = \beta_0 + \beta_1^{\beta_2 x}.$$

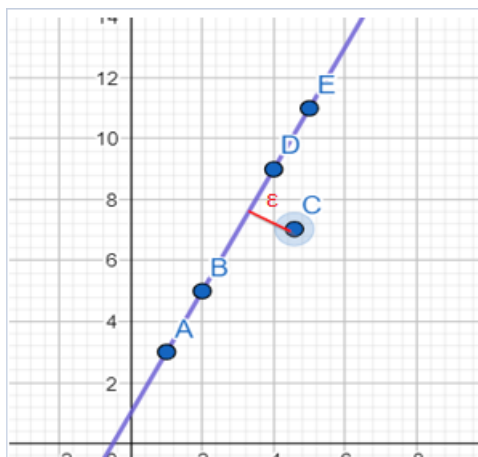
Τα μη-γραμμικά μοντέλα μπορούν να χωριστούν σε 3 κατηγορίες:

- Μη-διαχωρίσιμα μοντέλα, όπου η μερική παράγωγος δεν είναι σταθερή για κανέναν συντελεστή.
- Διαχωρίσιμα μοντέλα, όπου η μερική παράγωγος ενός συντελεστή είναι σταθερή συνάρτηση.
- Εγγενώς γραμμικά μοντέλα, τα οποία είναι μη-γραμμικά μοντέλα που μπορούν να μετασχηματιστούν σε γραμμικά μοντέλα.

Η πρώτη μορφή παλινδρόμησης που χρησιμοποιήθηκε για την εύρεση εξίσωσης μιας ευθείας είναι η μέθοδος ελαχίστων τετραγώνων, που χρησιμοποίησαν οι Legendre (Legendre, 1805) και Gauss (Gauss, 1809) για τον προσδιορισμό της τροχιάς αντικειμένων γύρω από τον ήλιο. Όμως, ο όρος της παλινδρόμησης επινοήθηκε από τον Galton το 1886 (Galton, 1886), όπου περιέγραψε το φαινόμενο ότι το ύψος απογόνων, οι οποίοι είχαν ψηλούς προγόνους τείνει να πέφτει προς τον μέσο όρο. Για τον Galton, ο όρος παλινδρόμηση ήταν μόνο βιολογικής σημασίας. Αργότερα, οι Yule (Yule, 1897) και Pearson (Pearson, 1903) γενίκευσαν τον όρο Παλινδρόμηση και τον συνδύασαν με την στατιστική. Στις δημοσιεύσεις τους, η από κοινού κατανομή (joint distribution) των εξαρτημένων και ανεξάρτητων μεταβλητών θεωρείται ότι είναι κανονική (γκαουσιανή). Ο

Fisher (Fisher, 1922-1925), αντιθέτως, θεώρησε ότι η υποθετική κατανομή είναι κανονική, αλλά η από κοινού κατανομή δεν χρειάζεται να είναι απαραίτητα κανονική.

Υπάρχει (συνήθως) διαφορά μεταξύ των εξαρτημένων μεταβλητών των δεδομένων και των εκτιμήσεων της συνάρτησης, που ονομάζεται σφάλμα. Το σφάλμα είναι στοχαστικό, κάτι που εξηγεί και το γιατί η παλινδρόμηση είναι στατιστική συνάρτηση και όχι ντετερμινιστική. Σύμφωνα με την Εικόνα 1.2 το σφάλμα μπορεί να οριστεί ως η απόσταση του σημείου από την ευθεία παλινδρόμησης και συνήθως συμβολίζεται με ε . το ε δε φαίνεται στο σχήμα



Εικόνα 1.2. Απεικόνιση σφάλματος σε απλό παράδειγμα Παλινδρόμησης.

Η μελέτη της σχέσης ανάμεσα σε μια εξαρτημένη και σε μια ανεξάρτητη μεταβλητή ονομάζεται απλή παλινδρόμηση, ενώ η μελέτη ανάμεσα σε μια εξαρτημένη και δύο ή περισσότερες ανεξάρτητες μεταβλητές, ονομάζεται πολλαπλή παλινδρόμηση. Συνήθως, μεγαλύτερος αριθμός ανεξάρτητων μεταβλητών συνεπάγεται μεγαλύτερη ακρίβεια πρόβλεψης της εξαρτημένης μεταβλητής, καθώς υπάρχουν περισσότεροι όροι που εξηγούν το φαινόμενο.

Εκτός από τις κλασσικές προσεγγίσεις κατασκευής του μοντέλου παλινδρόμησης, όπως οι προαναφερόμενες, οι πιο σύγχρονες προσεγγίσεις κάνουν χρήση τεχνικών της μηχανικής μάθησης. Γνωστοί αλγόριθμοι παλινδρόμησης αποτελούν η γραμμική παλινδρόμηση, τα

Δέντρα απόφασης, ο αλγόριθμος K-Πλησιέστερων Γειτόνων, Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM), παλινδρόμηση LASSO, κλπ.

Η παλινδρόμηση έχει εφαρμογές σε πολλούς κλάδους της επιστήμης για την εκτίμηση σχέσεων μεταξύ μεταβλητών, που έχουν ως σκοπό κυρίως την πρόβλεψη και βελτιστοποίηση. Μερικά παραδείγματα είναι η πρόβλεψη κατανάλωσης ηλεκτρικής ενέργειας, η πρόβλεψη της αξίας μετοχών στα χρηματιστήρια, η κατανόηση της σχέσης μεταξύ διαφημιστικών δαπανών και εσόδων μιας εταιρείας, η κατανόηση της σχέσης μεταξύ διαφορετικών ειδών προπόνησης και της απόδοσης ενός αθλητή, η κατανόηση σχέσης μεταξύ δόσης φαρμάκων και την αρτηριακή πίεση ενός ασθενή, η μέτρηση της απόδοσης λιπάσματος και νερού σε καλλιέργειες και άλλες.

1.2 Αβεβαιότητα στη Μηχανική Μάθηση

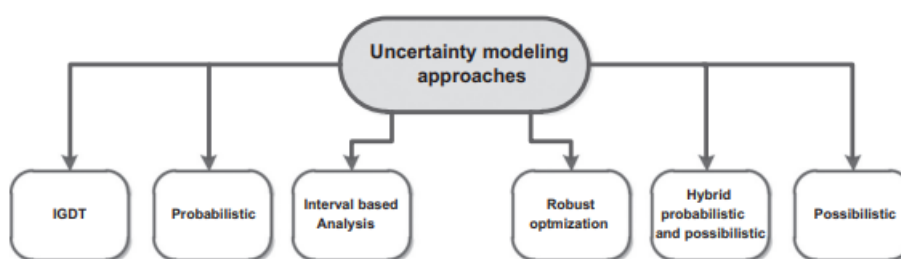
Η μηχανική μάθηση ασχολείται, ουσιαστικά, με τη γενίκευση σχέσεων μεταξύ δεδομένων, με σκοπό την δημιουργία μοντέλων που είναι σε θέση να περιγράφουν και να προβλέπουν ένα φαινόμενο ή μια διεργασία. Ως εκ τούτου, η μηχανική μάθηση είναι στενά συνδεδεμένη με την έννοια της αβεβαιότητας, αφού η διαδικασία δημιουργίας μοντέλων βασίζεται σε μια διαδικασία επαγωγής, δηλαδή αντικατάστασης των παρατηρήσεων από γενικευμένα μοντέλα. Τα μοντέλα αυτά κάνουν υποθέσεις, με αποτέλεσμα να είναι και αβέβαια ως προς την πρόβλεψη τους. Πέρα από την αβεβαιότητα στην υπόθεση του μοντέλου, μπορεί να υπάρχει αβεβαιότητα και στα δεδομένα, τα οποία μπορεί να είναι ασαφή ή θορυβώδη. (Hüllermeier and Waegeman, 2021). Η αβεβαιότητα στα δεδομένα μπορεί να οφείλεται σε διάφορους λόγους, όπως λάθος μετρήσεις αισθητήρων, ή γνωρίζοντας ότι υπάρχει απόκλιση στις μετρήσεις, ή μπορεί να οφείλεται σε παράγοντες που δεν μπορούν να προβλεφθούν με βεβαιότητα, όπως οι τιμές καυσίμων, οι μετοχές και ο καιρός.

Σε κάθε περίπτωση η αξιόπιστη αναπαράσταση και των δύο ειδών αβεβαιότητας είναι επιθυμητή. Στην πτυχιακή θα ασχοληθούμε με την αβεβαιότητα στα δεδομένα. Συγκεκριμένα, θα ασχοληθούμε με εκείνη την περίπτωση, όπου το εύρος των τιμών που μπορεί να πάρει μια τιμή εισόδου είναι γνωστή, χωρίς όμως να είναι γνωστή η ακριβής μεμονωμένη τιμή (crisp value). Στη συνέχεια θα αναλυθούν οι τρόποι με τους οποίους μπορούν να αναπαρασταθούν αβέβαια δεδομένα.

1.3 Αναπαράσταση αβέβαιων δεδομένων

Οι ήδη υπάρχουσες τεχνικές αναπαράστασης αβέβαιων δεδομένων καλύπτουν ένα ευρύ φάσμα αβεβαιότητας, καθώς υπάρχουν προσεγγίσεις που είναι πιθανοτικές, ασαφείς, υβριδικές ασαφή-πιθανοτικές, θεωρίας απόφασης κενών πληροφοριών (IGDT) και ισχυρής βελτιστοποίησης. Οι διαφορετικές τεχνικές απεικονίζονται συγκεντρωτικά στην Εικόνα 1.3.

Ο σκοπός των τεχνικών αυτών είναι να υπολογίσουν την επίδραση των αβέβαιων δεδομένων στις τελικές τιμές εξόδου. Η κάθε μια από αυτές περιγράφει με διαφορετικό τρόπο την αβεβαιότητα στα δεδομένα και η επιλογή της καλύτερης αναπαράστασης των δεδομένων εξαρτάται από το πρόβλημα.



Εικόνα 1.3. Προσεγγίσεις αναπαράστασης αβέβαιων δεδομένων. (Aien, Hajebrahimi and Fotuhi-Firuzabad, 2016)

1.3.1 Συμβολικά δεδομένα

Η στατιστική είναι η επιστήμη που ασχολείται με τα δεδομένα. (Wei Xu, 2020) Τα δεδομένα συνήθως είναι μεμονωμένες τιμές και μπορούν να αναπαρασταθούν με ένα σημείο σε ένα καρτεσιανό γινόμενο. Όμως, στην περίπτωση που ένα χαρακτηριστικό δεν παίρνει μια τιμή, αλλά ένα εύρος τιμών, το χαρακτηριστικό δεν αναπαριστάται με ένα σημείο, αλλά με υπερκύβους ή άλλα πολύεδρα στον χώρο. Τέτοια δεδομένα, που μπορεί να είναι μια λίστα, ένα διάστημα, ένα ιστόγραμμα κ.ο.κ, ονομάζονται συμβολικά δεδομένα (Symbolic data), και αποτελούν ένα είδος στο οποίο αναφέρθηκε για πρώτη φορά ο Diday (Diday, 1987).

Για αυτού του είδους δεδομένα, οι κλασικές τεχνικές της στατιστικής δεν μπορούν να εφαρμοστούν όπως έχουν. Οι Diday (Diday, 1995-1996) και Emilion (Emilion, 1996-1998) είναι οι πρώτοι που ασχολήθηκαν με αυτό το πρόβλημα και δημιούργησαν το μαθηματικό υπόβαθρο για την μελέτη και ανάλυση αυτών των δεδομένων. Οι κύριες κατηγορίες συμβολικών δεδομένων είναι τα δεδομένα πολλών τιμών (multi-valued data - π.χ. λίστες), τα δεδομένα διαστημάτων (interval-valued data) και τα τροπικά δεδομένα (modal-valued data - π.χ. ιστογράμματα). Οι μεμονωμένες τιμές μπορούν και αυτές να θεωρηθούν ειδική περίπτωση των συμβολικών δεδομένων, όπου κάθε τιμή έχει πιθανότητα 1 να συμβεί σε μια κατανομή.

ID	Race	Gender	Age	Marital Status	Parents Living	Weight	Income ^a
1	White	Male	[70,76]	{D,M}	{0,1}	[150,230]	[60,80]
2	Black	Female	[20,30]	{S,M}	{1,2}	[158,188]	[30,50]
3	Asian	Male	[30,36]	{S,M}	{1,2}	[120,200]	[50,80]
4	Hispanic	Female	[11,16]	{S}	{0,1,2}	[73,150]	[20,40]
...	...	...	...	...	...	...	...
<i>n</i>	Hispanic	Female	[20,26]	{M,S}	{0,1,2}	[120,196]	[30,60]

^a Income in \$1000's

Εικόνα 1.3. Δείγμα συνόλου δεδομένων απογραφής (WEI XU, 2020)

Στην παραπάνω Εικόνα 1.3 απεικονίζεται ένα δείγμα συνόλου δεδομένων απογραφής. Μπορεί εύκολα να διαπιστώσει κάποιος, ότι κάποια δεδομένα για πολλούς λόγους είναι αδύνατον να είναι μεμονωμένες τιμές, είτε γιατί δεν μας ενδιαφέρουν οι μεμονωμένες τιμές είτε διότι κάποιες πληροφορίες είναι απόρρητες και είναι πιθανόν πολίτες να μην θέλουν να τις μοιραστούν, όπως το εισόδημα και το βάρος, και άλλες που δεν έχει νόημα η τιμή από μόνη της, όπως η ηλικία.

1.3.2 Ασαφής προσέγγιση

- Ένας άλλος τρόπος για να αναπαρασταθούν τα αβέβαια δεδομένα είναι με ασαφή σύνολα. Η έννοια των ασαφών συνόλων προτάθηκε για πρώτη φορά από τον Lotfi Zadeh (Zadeh, 1965) και αποτελεί ένα τρόπο μίμησης της ανθρώπινης αντίληψης, για έννοιες που δεν είναι αντικειμενικές αλλά καθαρά υποκειμενικές (π.χ. μεγάλος-

μικρός, πολύ-λίγο). Ένα ασαφές σύνολο αποτελείται από ένα ζεύγος (A, m) , όπου το A είναι ένα σύνολο και για κάθε στοιχείο $x \in A$, υπάρχει η τιμή της συνάρτησης m στο x , $m(x)$, που δείχνει τον βαθμό συμμετοχής δηλ. το βαθμό με τον οποίο κάθε στοιχείο ανήκει στο ασαφές σύνολο A . Ο βαθμός συμμετοχής παίρνει τιμές μεταξύ 0 (δεν ανήκει ή δεν υπάρχει συσχέτιση) και 1 (ανήκει ή υπάρχει πλήρης συσχέτιση). Τα ασαφή δεδομένα δεν είναι μέρος των συμβολικών δεδομένων. Οι λόγοι είναι οι εξής:

- Τα ασαφή σύνολα υπάρχουν για να μιμηθούν την ανθρώπινη σκέψη, ενώ τα συμβολικά δεδομένα είναι ένας τρόπος αναπαράστασης δεδομένων για να αναλυθούν σωστά με τεχνικές όπως είναι η συσταδοποίηση, η ανάλυση κύριων συνιστωσών (PCA) και η παλινδρόμηση κ.λπ. Τα ασαφή δεδομένα αποτελούνται από ασαφείς αριθμούς και ασαφή διαστήματα, ενώ τα συμβολικά δεδομένα από μεμονωμένες τιμές, πολλαπλές τιμές, διαστήματα, ιστογράμματα και κατανομές.
- Η έννοια του βαθμού συμμετοχής είναι διαφορετική από την πιθανότητα που χρησιμοποιείται στα συμβολικά δεδομένα. (Zadeh, 1965)
- Τα συμβολικά δεδομένα βοηθούν στο να κατασκευαστούν στατιστικά μοντέλα και να πραγματοποιήσουν ανάλυση των δεδομένων, τα ασαφή δεδομένα προσπαθούν απλώς να λύσουν το πρόβλημα, χωρίς να κατανοήσουν τον τρόπο λειτουργίας.

1.3.3 Πιθανολογική προσέγγιση

Σε αυτήν την προσέγγιση η συνάρτηση πυκνότητας πιθανότητας (Probability Density Function – PDF) των τιμών εισόδου θεωρείται γνωστή. Για μια συνάρτηση $y = f(x)$, όπου x το διάνυσμα εισόδων των αβέβαιων δεδομένων, σκοποός είναι να βρεθεί η συνάρτηση πυκνότητας πιθανότητας για το y . Για την εύρεση της PDF του y υπάρχουν δύο ειδών μέθοδοι, η αριθμητική προσέγγιση και η αναλυτική προσέγγιση. Μια από τις πιο διάσημες τεχνικές αριθμητικής προσέγγισης είναι η προσομοίωση Monte Carlo (Monte Carlo Simulation – MCS), η οποία χρησιμοποιείται κυρίως για μη-γραμμικές συναρτήσεις με σύνθετα ή και αβέβαια δεδομένα. Η MCS είναι μια στοχαστική και επαναληπτική μέθοδος και λειτουργεί ως εξής:

Βήμα 1. MCS counter $C = 1$

Βήμα 2. Δημιουργία ενός δείγματος με τυχαία επιλογή από το διάνυσμα X ,

χρησιμοποιώντας την PDF για κάθε στοιχείο x_j .

Βήμα 3. Υπολογισμός της y_c , θεωρώντας ότι $X = X_c$ και $y_c = f(X_c)$

Βήμα 4. Υπολογισμός της προβλεπόμενης τιμής y , με τον τύπο $E(y) = \frac{\sum y_c}{c}$

Βήμα 5. Υπολογισμός της διακύμανσης του y , $\sigma(y) = E(Y^2) - E^2(Y)$

Βήμα 6. Εάν πληρείται το κριτήριο τερματισμού: πήγαινε *Βήμα 7*,
αλλιώς $C = C + 1$ και πήγαινε στο *Βήμα 2*

Βήμα 7. Τερματισμός.

(Aien, Hajebrahimi and Fotuhi-Firuzabad, 2016)

Υπάρχουν διάφορες παραλλαγές της τεχνικής MCS για λόγους προσαρμογής σε προβλήματα, όπως η ακολουθιακή προσομοίωση Monte Carlo (sequential Monte Carlo) και η ψευδο-ακολουθιακή προσομοίωση Monte Carlo (pseudo-sequential Monte Carlo simulation). Στα πλεονεκτήματα της μεθόδου MCS είναι ότι λειτουργεί σε σύνθετα προβλήματα και η τεχνική είναι διαισθητικά εύκολο να κατανοηθεί και να υλοποιηθεί. Από την άλλη, είναι μια αργή τεχνική, καθώς χρειάζεται πολλαπλές επαναλήψεις για τον υπολογισμό των συναρτήσεων, ενώ παράλληλα για μεγάλη ακρίβεια στα αποτελέσματα χρειάζεται ακόμα μεγαλύτερος αριθμός επαναλήψεων.

Η αναλυτική προσέγγιση χρησιμοποιεί μαθηματικές εκφράσεις, για να αναλύσει το σύστημα και τις εισόδους του. Η βασική ιδέα αυτής της προσέγγισης είναι να κάνει αριθμητική σε PDFs από τυχαίες μεταβλητές εισόδου. Η αναλυτική προσέγγιση μπορεί να χωριστεί σε δύο κύριες κατηγορίες:

- Μεθόδους γραμμικοποίησης, όπου οι μέθοδοι προσπαθούν να μετατρέψουν μια μη-γραμμική συνάρτηση σε μια γραμμική κατά προσέγγιση. Γνωστές τεχνικές αυτής της κατηγορίας είναι οι σειρές Taylor, η επέκταση Cornish-Fisher και Edgeworth, η σωρευτική μέθοδος και η μέθοδος συνέλιξης.
- Μέθοδοι προσέγγισης της συνάρτησης πυκνότητας πιθανότητας, η οποία έχει το πλεονέκτημα ότι είναι πιο εύκολο να προσεγγιστεί μια PDF, παρά μια μη-γραμμική συνάρτηση. Το κύριο χαρακτηριστικό αυτής της κατηγορίας είναι ο τρόπος με τον οποίο δημιουργούνται κατάλληλα δείγματα από τις τιμές εισόδων που μπορούν να κρατούν αποτελεσματικά πληροφορίες για την PDF τους. Αντιπροσωπευτικές τεχνικές αυτής της κατηγορίας είναι η μέθοδος εκτιμής σημείου (point

estimation method) και ο Unscented μετασχηματισμός. (Aien, Hajebrahimi and Fotuhi-Firuzabad, 2016)

Άλλες προσεγγίσεις για την αντιμετώπιση αβέβαιων δεδομένων είναι συνδυασμός ασαφούς και πιθανολογικής προσέγγισης, για δεδομένα τα οποία μερικά μοντελοποιούνται με την μια προσέγγιση και άλλα με την άλλη προσέγγιση, η μέθοδος IGDT (Information Gap Decision Theory), για αβέβαια δεδομένα που δεν μπορούν να μοντελοποιηθούν με συνάρτηση πυκνότητας πιθανότητας ή βαθμό συμμετοχής, λόγω έλλειψης πληροφοριών και, τέλος η εύρωστη βελτιστοποίηση (robust optimization), που χρησιμοποιείται σε αβέβαια δεδομένα, για τα οποία δεν είναι γνωστή η προέλευση της αβεβαιότητας. (Aien, Hajebrahimi and Fotuhi-Firuzabad, 2016)

Στην πτυχιακή θα αναλυθεί η περίπτωση των διαστημάτων των συμβολικών δεδομένων, καθώς για την παλινδρόμηση σε δεδομένα με μορφή διαστημάτων υπάρχουν πολλές διάσημες και αποτελεσματικές τεχνικές.

1.4 Διαστήματα (Intervals)

Η Ανάλυση Διαστημάτων είναι κλάδος των υπολογιστικών μαθηματικών και συγκεκριμένα της αριθμητικής ανάλυσης, ο οποίος έχει ως στόχο να αντιμετωπίσει το πρόβλημα της ανακρίβειας σε αποτελέσματα, που προκύπτουν από κλασικές αριθμητικές μεθόδους. Οι αλγόριθμοι και οι τεχνικές που έχουν αναπτυχθεί παράγουν πιο ακριβή φράγματα, και μπορούν με απόλυτη σιγουριά να υπολογίσουν ένα διάστημα που να περιέχει τις επιθυμητές τιμές των υπολογισμών, το οποίο έχει ως αποτέλεσμα αξιόπιστους και επιβεβαιωμένους (verified) μαθηματικούς υπολογισμούς. Η αντιμετώπιση και η διαχείριση των διαστημάτων οδήγησε στην δημιουργία της αριθμητικής διαστημάτων, όπου οι πράξεις γίνονται με διαστήματα και όχι με μεμονωμένες πραγματικές τιμές. Τομείς στους οποίους έχει πρακτική εφαρμογή η ανάλυση διαστημάτων είναι η χημεία, η δομική μηχανική, τα οικονομικά, τα προβλήματα βελτιστοποίησης, η ικανοποίηση περιορισμών, ο υπολογισμός τροχίας αστεροειδών και τα γραφικά. Επιπλέον, η ανάλυση διαστημάτων έχει χρησιμοποιηθεί για αποδείξεις θεωρημάτων με την βοήθεια

υπολογιστών, με χαρακτηριστικό παράδειγμα να είναι η υπόθεση του Kepler (Kepler conjecture). (Moore, Kearfott and Cloud, 2009)

Ως διάστημα ορίζεται ένα ζεύγος τιμών $[a, b]$, το οποίο αντιπροσωπεύει όλες τις πραγματικές τιμές x , για τις οποίες ισχύει $a \leq x \leq b$. Αν και υπάρχουν πολλές μορφές διαστημάτων, στην πτυχιακή όταν θα αναφέρεται η έννοια του διαστήματος θα εννοείται το κλειστό διάστημα που ορίζεται ως εξής:

$$[a, b] = \{x \in \mathbb{R}: a \leq x \leq b\}.$$

Έστω δύο διαστήματα $X = [\alpha, \beta]$ και $Y = [\gamma, \delta]$. Μερικές βασικές πράξεις διαστημάτων που θα χρειαστούν για τις τεχνικές Παλινδρόμησης είναι:

- Το πλάτος

$$w(X) = \beta - \alpha$$

- Η ακτίνα

$$\alpha(X) = w(X) / 2 = (\beta - \alpha) / 2$$

- Το κέντρο (midpoint)

$$m(X) = 1/2 \cdot (\alpha + \beta)$$

- Πρόσθεση

$$X + Y = \{x + y: x \in X, y \in Y\}$$

- Αφαίρεση

$$X - Y = \{x - y: x \in X, y \in Y\}$$

- Πολλαπλασιασμός

$$X \cdot Y = \{x \cdot y: x \in X, y \in Y\}$$

- Διάρθρωση (το διάστημα Y δεν πρέπει να περιέχει το μηδέν 0)

$$X / Y = \{x / y: x \in X, y \in Y\}$$

Αν τα διαστήματα X και Y ορίζονται ως $[\underline{x}, \bar{x}]$ και $[\underline{y}, \bar{y}]$ τότε οι αριθμητικές πράξεις δίνονται από τις ακόλουθες εκφράσεις:

$$[\underline{x}, \bar{x}] + [\underline{y}, \bar{y}] = [\underline{x} + \underline{y}, \bar{x} + \bar{y}]$$

$$[\underline{x}, \bar{x}] - [\underline{y}, \bar{y}] = [\underline{x} - \bar{y}, \bar{x} - \underline{y}]$$

$$[\underline{x}, \bar{x}] \cdot [\underline{y}, \bar{y}] = [\min(\underline{x}\underline{y}, \underline{x}\bar{y}, \bar{x}\underline{y}, \bar{x}\bar{y}), \max(\underline{x}\underline{y}, \underline{x}\bar{y}, \bar{x}\underline{y}, \bar{x}\bar{y})]$$

$$[\underline{x}, \bar{x}] \div [\underline{y}, \bar{y}] = [\underline{x}, \bar{x}] \cdot \frac{1}{[\underline{y}, \bar{y}]}, \text{ όπου}$$

$$\frac{1}{[\underline{y}, \bar{y}]} = \left[\frac{1}{\bar{y}}, \frac{1}{\underline{y}} \right], \text{ δοθέντος ότι } 0 \notin [\underline{y}, \bar{y}]$$

2.Τεχνικές Παλινδρόμησης

2.1Κλασικές τεχνικές Παλινδρόμησης

2.1.1 Γραμμική Παλινδρόμηση

Η γραμμική παλινδρόμηση είναι μια από τις πιο διαδεδομένες τεχνικές πρόβλεψης στην στατιστική και την μηχανική μάθηση. Έστω, ένα διάνυσμα $x^t = (x_1, x_2, \dots, x_p)$, όπου p ο αριθμός των μεταβλητών και x^t το ανάστροφο διάνυσμα. Η πρόβλεψη της εξαρτημένης τιμής y μπορεί να υπολογιστεί με τον εξής τύπο:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^p x_j \hat{\beta}_j, \quad (1)$$

Στην περίπτωση που στο διάνυσμα x προστεθεί και ένας άσος και ο συντελεστής β_0 στο διάνυσμα συντελεστών, ο τύπος μπορεί να γραφεί σαν γινόμενο 2 διανυσμάτων ως εξής:

$$\hat{y} = x^t \hat{\beta}, \quad (2)$$

Η πιο διαδεδομένη τεχνική για την κατασκευή του γραμμικού μοντέλου πάνω σε ένα σύνολο εκπαίδευσης, που περιγράφει ένα φαινόμενο, είναι η μέθοδος των ελάχιστων τετραγώνων (MET-RSS). Σε αυτήν την μέθοδο επιλέγεται κάθε συντελεστής και σκοπός είναι η ελαχιστοποίηση της σχέσης. Ο τύπος είναι:

$$RSS(\beta) = \sum_{i=1}^n (y_i - x^t \beta)^2, \quad (3)$$

όπου n ο αριθμός παρατηρήσεων του συνόλου δεδομένων. Η συνάρτηση 3 είναι τετραγωνική, επομένως υπάρχει πάντα ελάχιστη τιμή. Η εξίσωση μπορεί να γραφεί:

$$RSS(\beta) = (y - X\beta)^t (y - X\beta), \quad (4)$$

όπου X ένας πίνακας $n \times p$, με κάθε γραμμή του πίνακα να είναι ένα διάνυσμα του συνόλου εκπαίδευσης και y να είναι ένα διάνυσμα n μεγέθους που περιέχει τις τιμές εξόδου. Τέλος, αν το γινόμενο $X^t X$ είναι αντιστρέψιμος πίνακας, η εξίσωση μπορεί να γραφεί:

$$\hat{\beta} = (X^t X)^{-1} X^t y. \quad (5)$$

Άλλα γνωστά γραμμικά μοντέλα είναι οι αλγόριθμοι παλινδρόμησης Ridge και Lasso. Αυτοί οι αλγόριθμοι κανονικοποιούν τις παραμέτρους, και, στην περίπτωση του Lasso μπορεί να εξαλείψει και εντελώς κάποιους συντελεστές. Η κανονικοποίηση που κάνει ο Ridge, είναι και γνωστή σαν L2, ενώ η κανονικοποίηση του Lasso ως L1. (Hastie, Tibshirani and Friedman, 2009)

2.1.2 k-Πλησιέστεροι Γείτονες

Δοθέντος έναν αριθμού k και ένα παράδειγμα x_0 , βρίσκεται ένα σύνολο των K πλησιέστερων στοιχείων στο x_0 . Συνήθως, η πρόβλεψη $\hat{y}_0$ για το x_0 είναι ο μέσος όρος των τιμών εξόδου του συνόλου των πλησιέστερων γειτόνων. Ο τύπος είναι:

$$\hat{y}_0 = \frac{1}{K} \sum_{i \in N_0} y_i,$$

όπου N_0 το σύνολο των πλησιέστερων στοιχείων. Συνήθως, για την μέτρηση της απόστασης χρησιμοποιείται η MET (Μέθοδος Ελάχιστων Τετραγώνων), αλλά συχνά εφαρμόζονται κι άλλες τεχνικές μέτρησης απόστασης. Για την εύρεση κατάλληλης τιμής k , συνίσταται η χρήση της μεθόδου cross-validation για όλες τιμές $k = 1, 2, \dots, \sqrt{n}$, όπου n πληθάριμος του συνόλου N_0 . (Hastie, Tibshirani and Friedman, 2009)

2.1.3 Μηχανές Διανυσμάτων Υποστήριξης

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines – SVMs) είναι ένα βελτιωμένο γραμμικό μοντέλο, καθώς είναι λιγότερη ευαίσθητη σε ακραίες τιμές. Αυτό το καταφέρνει θέτοντας ένα κατώφλι ϵ , ώστε σφάλματα που βρίσκονται μέσα στο ορισμένο περιθώριο να μην λαμβάνονται υπόψιν. Έτσι, ακραίες τιμές έχουν μικρή έως και καθόλου επίδραση πάνω στο μοντέλο. Η συνάρτηση είναι:

$$LF(\beta) = C \sum_{i=1}^n L_{\epsilon}(y_i - \hat{y}_i) + \sum_{j=1}^p \beta_j^2,$$

όπου C ο συντελεστής ποινής (cost penalty) και L_{ϵ} η συνάρτηση που ορίζεται ως:

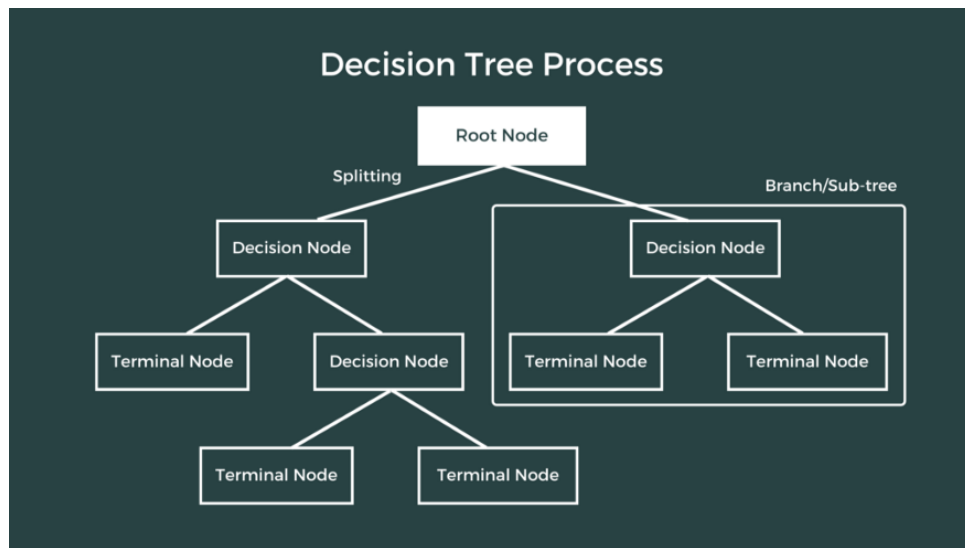
$$L_{\epsilon}(\xi) = \begin{cases} |\xi| - \epsilon & \text{si } |\xi| > \epsilon \\ 0 & \text{si } |\xi| \leq \epsilon \end{cases}$$

Σκοπός είναι να βρεθούν συντελεστές $\hat{\beta}_j$ που να ελαχιστοποιούν την $LF(\beta)$.

2.1.4 Δέντρα απόφασης και Random Forests

Τα δέντρα απόφασης στην παλινδρόμηση χρησιμοποιούν μοντέλα σε μορφή δέντρων. Ένα δέντρο απόφασης αποτελείται από:

- Ρίζα: Είναι η αρχή του δέντρου που αντιπροσωπεύει όλα τα δεδομένα.
- Κόμβους απόφασης: Είναι οι κόμβοι που διαχωρίζονται σε δύο ή περισσότερους κόμβους. Ο διαχωρισμός γίνεται με βάση ένα χαρακτηριστικό.
- Φύλλο/Τερματικός κόμβος: Κόμβοι, οι οποίοι δεν διαχωρίζονται άλλο. Συνήθως, αυτοί οι κόμβοι είναι το τελικό αποτέλεσμα (πρόβλεψη).



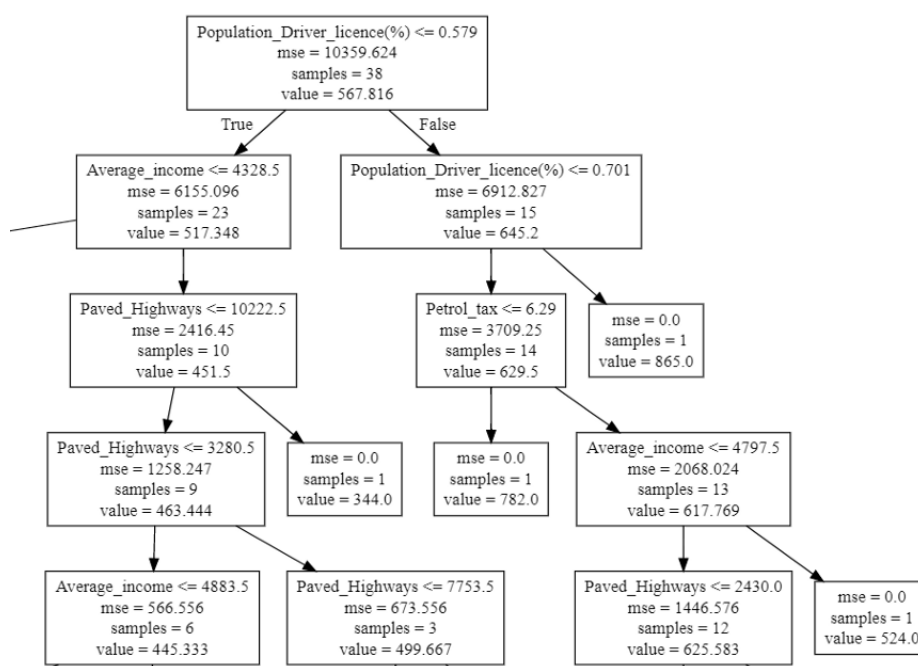
Εικόνα 2.1. Δομή απόφασης δέντρου. (The Click Reader, 2017)

Η διαδικασία διαχωρισμού των κόμβων ξεκινά από την ρίζα και ακολουθείται από ένα υποδέντρο που οδηγεί σε τερματικό κόμβο, που περιέχει την τελική πρόβλεψη. Η κατασκευή του δέντρου ξεκινά από πάνω προς τα κάτω, διαλέγοντας σε κάθε κόμβο μια μεταβλητή (χαρακτηριστικό) που διαχωρίζει το σύνολο δεδομένων κατά τον βέλτιστο τρόπο. Τα δέντρα που στον τερματικό κόμβο έχουν συνεχείς τιμές, ονομάζονται και δέντρα παλινδρόμησης .

Η ακρίβεια του μοντέλου υπολογίζεται με βάση τον κάθε κόμβο ξεχωριστά. Η συνάρτηση κόστους που χρησιμοποιείται πιο συχνά είναι ο τύπος RMSE (Root Mean Square Error) και δίνεται από τη σχέση:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$

όπου n το πλήθος των στοιχείων στον κόμβο, y_i η πραγματική τιμή της εξαρτημένης τιμής και $\hat{y}_i$ η πρόβλεψη του μοντέλου.



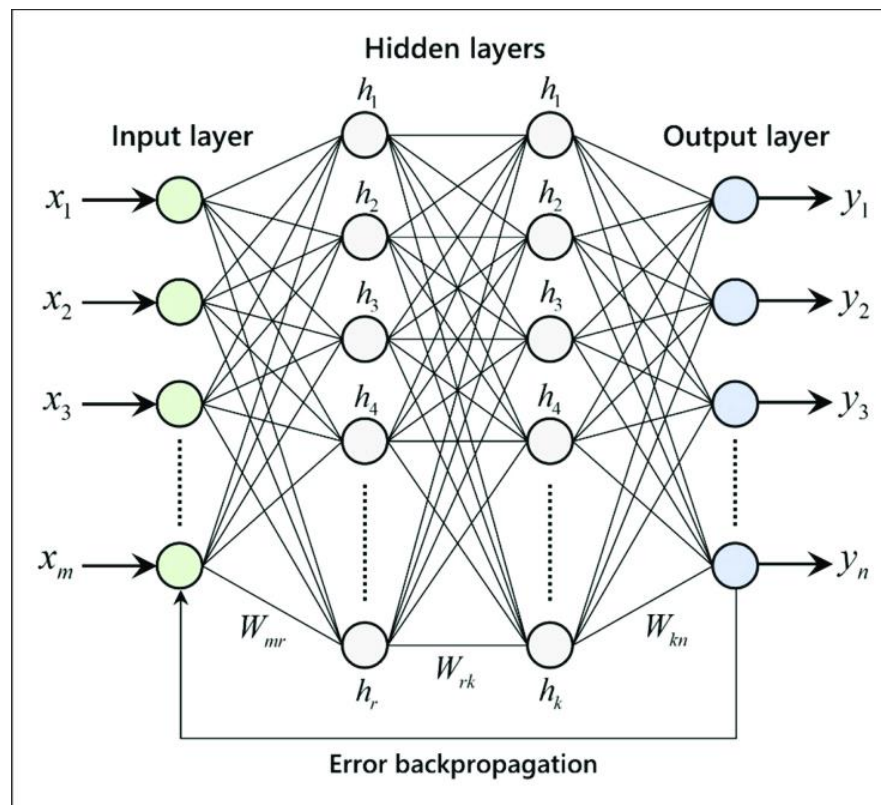
Εικόνα 2.2. Παράδειγμα απόφασης δέντρου. (The Click Reader, 2017)

Τα Random Forests είναι ένα σύνολο δέντρων απόφασης, όπου κάθε δέντρο διαλέγει τυχαία κάποια χαρακτηριστικά. Αυτή η τεχνική βοηθά στον περιορισμό του σφάλματος, λόγω μεροληψίας και διακύμανσης, κάνοντας τα Random Forests μια πιο εύρωστη τεχνική σε σχέση με τα δέντρα απόφασης. (Liberman, 2017)

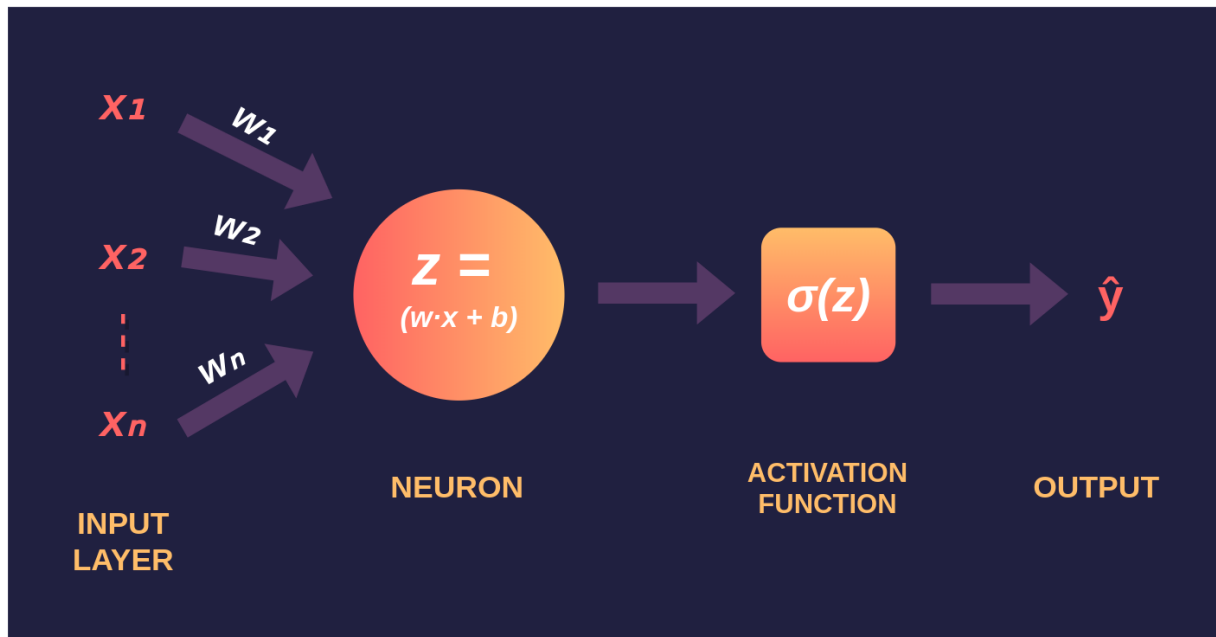
2.1.5 Τεχνητά Νευρωνικά δίκτυα

Στα μοντέλα τεχνητών νευρωνικών δικτύων (ΤΝΔ) η συνάρτηση που προσομοιώνει την σχέση ανάμεσα σε δεδομένα είναι η $y = f^*(x)$, όπου η $f(x)$ αποτελεί την σύνθεση περισσότερων συναρτήσεων και είναι της μορφής $f(x) = (f^{(n)} \circ f^{(n-1)} \circ \dots \circ f^{(1)})$, όπου $f^{(i)}$ το i – οστό επίπεδο του δικτύου, n το βάθος του δικτύου και $f^{(n)}$ η έξοδος του δικτύου. Τα υπόλοιπα επίπεδα ονομάζονται κρυφά επίπεδα. Ένα ΤΝΔ μπορεί να αναπαρασταθεί με ένα κατευθυνόμενο ακυκλικό γράφο, που περιγράφει την σύνθεση μεταξύ των συναρτήσεων. Τα κρυφά επίπεδα έχουν την ευθύνη να αναγνωρίσουν τα διάφορα μοτίβα που μπορεί να υπάρχουν στα δεδομένα. Τα δεδομένα εκπαίδευσης κατευθύνουν το επίπεδο εξόδου, για να βγάλει όσο το δυνατόν πιο κοντινά αποτελέσματα στην πραγματική έξοδο. Ο αλγόριθμος εκπαίδευσης είναι υπεύθυνος να αξιοποιήσει τα κρυφά επίπεδα με τέτοιο τρόπο ώστε να βγάλουν αποδεκτές λύσεις.

Το Perceptron είναι η πιο απλή μορφή ενός νευρωνικού δικτύου και εφευρέθηκε από τον Frank Rosenblatt το 1958. Αποτελούνται από n εισόδους, έναν νευρώνα και μια έξοδο, όπου n ο αριθμός των χαρακτηριστικών ενός συνόλου δεδομένων. Ένα ΤΝΔ έχει δύο φάσεις: την πρόσθια τροφοδότηση (forward propagation) και την ανατροφοδότηση (backpropagation). Η διαδικασία περάσματος δεδομένων μέσα στο ΤΝΔ είναι η πρόσθια τροφοδότηση, ενώ ανατροφοδότηση είναι η διαδικασία ενημέρωσης των βαρών.



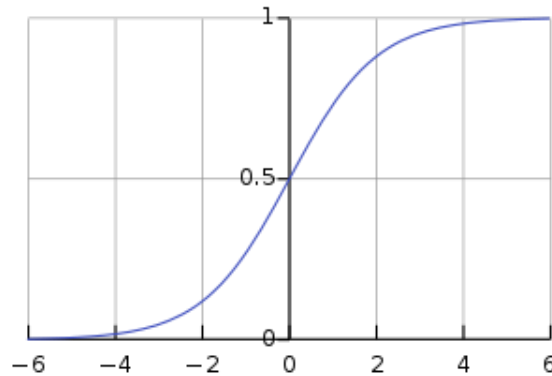
Εικόνα 2.3. Ένα Τεχνητό Νευρωνικό Δίκτυο πολλαπλών στρωμάτων (Education, 2022)



Εικόνα 2.4. Ένα απλό τεχνητό νευρωνικό δίκτυο. (A Gentle Introduction To Math Behind Neural Networks, 2022)

Κάθε είσοδος x_i πολλαπλασιάζεται με το αντίστοιχο βάρος του w_i και τα γινόμενα αθροίζονται. Επομένως, η σχέση είναι $\Sigma = (x_1 w_1) + (x_2 w_2) + \dots + (x_n w_n)$. Αν x και w δύο διανύσματα, όπου $x = (x_1, x_2, \dots, x_n)$ και $w = (w_1, w_2, \dots, w_n)$, τότε το άθροισμα των γινομένων μπορεί να γραφεί $\Sigma = x \cdot w$. Τα βάρη αντιπροσωπούν την ισχύς μεταξύ των νευρώνων και αποφασίζουν κατά πόσο επηρεάζει κάθε χαρακτηριστικού την έξοδο του νευρώνα. Δηλαδή, ένα χαρακτηριστικό με μεγαλύτερη τιμή βάρους από ένα άλλο, επηρεάζει περισσότερο την έξοδο του νευρώνα από το χαρακτηριστικό με την μικρότερη τιμή βάρους. Επίσης, στο τελικό άθροισμα προστίθεται μια τιμή πόλωσης (bias), που αντισταθμίζει την συνάρτηση ενεργοποίησης και την μετακινεί προς τα δεξιά ή τα αριστερά για να παραχθούν επιθυμητές τιμές εξόδου. Η σχέση που δημιουργείται είναι η εξής $Z = x \cdot w + b$, όπου b η τιμή πόλωσης. Η τιμή του Z εισάγεται σε μια συνάρτηση ενεργοποίησης στην περίπτωση του Perceptron, ή περισσότερες όταν έχει περισσότερα κρυφά επίπεδα (μια συνάρτηση ενεργοποίησης για κάθε κρυφό επίπεδο). Οι συναρτήσεις ενεργοποίησης συμβάλλουν στον χρόνο εκπαίδευσης του ΤΝΔ. Η πιο συνηθισμένη συνάρτηση που χρησιμοποιείται σαν συνάρτηση ενεργοποίησης είναι η λογιστική (Εικόνα 2.4), γνωστή και ως σιγμοειδής. Χρησιμοποιώντας σαν συνάρτηση ενεργοποίησης την λογιστική, η πρόβλεψη του ΤΝΔ υπολογίζεται από τον τύπο:

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^z}$$



Εικόνα 2.5. Λογιστική συνάρτηση (Logistic function - Wikipedia, 2022)

Εφόσον, έχει υπολογιστεί η τιμή πρόβλεψης από το ΤΝΔ, το νευρωνικό δίκτυο ξεκινά την διαδικασία της ανατροφοδότησης, ξεκινώντας από την εύρεση του σφάλματος. Η MET είναι η πιο διαδεδομένη μέθοδος εύρεσης σφαλμάτων και η διαδικασία είναι ίδια με αυτήν που αναφέρθηκε στην γραμμική παλινδρόμηση. Για την βελτιστοποίηση της εκπαίδευσης των ΤΝΔ χρησιμοποιείται συνήθως ο αλγόριθμος κατάβασης δυναμικού (gradient descent). Η διαδικασία της ανατροφοδότησης και ο αλγόριθμος κατάβασης δυναμικού επαναλαμβάνεται μέχρι να συγκλίνει και ενημέρωση των βαρών γίνεται ως εξής:

$$w_i = w_i - (a \cdot \frac{\partial C}{\partial w_i})$$

$$b = b - (a \cdot \frac{\partial C}{\partial b})$$

όπου $C = \text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$

και

$$\frac{\partial C}{\partial w_i} = \frac{2}{n} \cdot \text{sum}(y - \hat{y}) \cdot \sigma(z) \cdot (1 - \sigma(z)) \cdot x_i$$

$$\frac{\partial C}{\partial b} = \frac{2}{n} \cdot \text{sum}(y - \hat{y}) \cdot \sigma(z) \cdot (1 - \sigma(z))$$

2.2 Τεχνικές Παλινδρόμησης σε δεδομένα μορφής διαστημάτων

2.2.1 Μέθοδος κέντρου

Η μέθοδος κέντρου (Center Method) προτάθηκε το 2000 (Billiard, Diday, 2000) και είναι η πρώτη προσέγγιση για παλινδρόμηση ενός γραμμικού μοντέλου σε δεδομένα υπό μορφή διαστημάτων. Οι Billiard και Diday πρότειναν ένα μοντέλο όπου οι συντελεστές βγαίνουν με παλινδρόμηση πάνω στα κέντρα των διαστημάτων. Αφού ολοκληρωθεί η εκπαίδευση, η εφαρμογή του μοντέλου στα καινούργια δεδομένα πραγματοποιείται στα άνω και κάτω όρια των διαστημάτων. Στην ερευνά τους, οι Billiard και Diday σύγκριναν την δικιά τους τεχνική με δύο απλά μοντέλα παλινδρόμησης στα άνω και κάτω όρια και συμπέραναν ότι η προτεινόμενη τεχνική τους είναι πιο αποδοτική. Πρακτικά, αν X^c είναι ο πίνακας με τα κέντρα των διαστημάτων του πίνακα X , y^c το διάνυσμα των κέντρων του διανύσματος Y και αν το γινόμενο $((X^c)^t X^c)$ είναι αντιστρέψιμος πίνακας, τότε οι συντελεστές μπορούν να υπολογιστούν ως εξής:

$$\hat{\beta} = ((X^c)^t X^c)^{-1} (X^c)^t y^c. \quad (6)$$

Η πρόβλεψη του διαστήματος $y = [y_L, y_U]$ για ένα καινούργιο $x = (x_1, x_2, \dots, x_p)$, όπου $x_j = [a_j, b_j]$ είναι:

$$\hat{y}_L = (x_L)^t \hat{\beta} \text{ και } \hat{y}_U = (x_U)^t \hat{\beta} \quad (7)$$

όπου $(x_L)^t = (1, a_1, \dots, a_p)$ και $(x_U)^t = (1, b_1, \dots, b_p)$.

Αυτή η τεχνική, αν και πρωτότυπη για να αντιμετωπιστεί η παλινδρόμηση σε δεδομένα υπό μορφή διαστημάτων, λαμβάνει υπόψιν μόνο τα κέντρα των διαστημάτων και όχι άλλες σημαντικές πληροφορίες των διαστημάτων όπως το μήκος του διαστήματος που σχετίζεται με τη διασπορά των τιμών γύρω από το μέσο.

2.2.2 Μέθοδος κέντρου και εύρους

Η μέθοδος κέντρου και εύρους (Center and Range Method - CRM) (Neto and Carvahlo, 204) προτάθηκε από τους Lima Neto και Carvalho το 2004. Η CRM υπολογίζει το διάνυσμα συντελεστών β χρησιμοποιώντας και τα κέντρα και τα εύρη των διαστημάτων, έχοντας ως αποτέλεσμα μια πιο ακριβή πρόβλεψη από την CM μέθοδο. Η CRM χρησιμοποιεί δύο γραμμικά μοντέλα παλινδρόμησης, ένα μοντέλο για τα κέντρα και ένα για τα εύρη. Το γραμμικό μοντέλο για τα κέντρα είναι ακριβώς το ίδιο με το γραμμικό μοντέλο της CM. Για το δεύτερο γραμμικό μοντέλο χρησιμοποιείται η ακτίνα του κάθε διαστήματος. Επομένως, συμβολίζεται με X^r ο πίνακας με τις ακτίνες των διαστημάτων του πίνακα X και με y^r το διάνυσμα των ακτίνων του διανύσματος Y . Έτσι, αν το γινόμενο $((X^r)^t X^r)$ είναι αντιστρέψιμο, οι τύποι για την εύρεση των συντελεστών είναι:

$$\hat{\beta}^c = ((X^c)^t X^c)^{-1} (X^c)^t y^c, \quad (8)$$

όπου $\hat{\beta}^c$ οι συντελεστές που προκύπτουν από τα κέντρα και

$$\hat{\beta}^r = ((X^r)^t X^r)^{-1} (X^r)^t y^r, \quad (9)$$

όπου $\hat{\beta}^r$ οι συντελεστές που προκύπτουν από τις ακτίνες.

Τα όρια του διαστήματος της πρόβλεψης υπολογίζονται ως εξής:

$$\hat{y}_L = \hat{y}^c - \hat{y}^r \text{ και } \hat{y}_U = \hat{y}^c + \hat{y}^r \quad (10)$$

όπου:

$$\hat{y}^c = (x^c)^t \hat{\beta}^c \text{ και } \hat{y}^r = (x^r)^t \hat{\beta}^r \quad (11)$$

και $(x^c)^t = (1, x^c_1, \dots, x^c_p)$, $(x^r)^t = (1, x^r_1, \dots, x^r_p)$.

Το πρόβλημα αυτής της μεθόδου είναι ότι δεν μπορεί να εγγυηθεί στην πρόβλεψη το κάτω όριο να είναι μικρότερο ή ίσο από το πάνω όριο.

Τόσο η CMR όσο και η CM έχουν επεκταθεί και σε άλλες τεχνικές με μερικές να είναι μια γενίκευση της τεχνικής Lasso (Giordani, 2014), τα δέντρα απόφασης, η τεχνική Boosting, ο αλγόριθμος k-πλησιέστεροι γείτονες, οι μηχανές διανυσμάτων υποστήριξης και τα νευρωνικά δίκτυα (Chacón and Rodríguez, 2021).

2.2.3 Διμεταβλητή μέθοδος κέντρου και εύρους

Η Διμεταβλητή μέθοδος κέντρου και εύρους (Bivariate Center And Range Method - BCRM – Billiard and Diday, 2007) είναι μια μέθοδος που βασίζεται στην τεχνική CRM. Στην μέθοδο BCRM, σε αντίθεση με την CRM, δεν θεωρείται ότι οι μεταβλητές κέντρου και εύρους είναι ανεξάρτητες. Το μοντέλο φτιάχνεται κάνοντας χρήση και των δύο μεταβλητών ταυτόχρονα. Το μοντέλο παίρνει την εξής μορφή:

$$Y = X\beta + \epsilon$$

όπου $Y = (Y^c, Y^r)$, $Y^c = (Y_1^c, Y_2^c, \dots, Y_n^c)'$, $Y^r = (Y_1^r, Y_2^r, \dots, Y_n^r)'$ (διάνυσμα ευρών), $X_i = (1, X_{i1}^c, \dots, X_{ip}^c, X_{i1}^r, X_{ip}^r)$ και $\beta = (\beta_0, \beta_1^c, \dots, \beta_p^c, \beta_1^r, \dots, \beta_p^r)'$.

Τέλος, αν το γινόμενο $X^t X$ είναι αντιστρέψιμος πίνακας, η εξίσωση μπορεί να γραφεί:

$$\hat{\beta} = (X^t X)^{-1} X^t Y.$$

Η τιμή πρόβλεψης για μια καινούργια παρατήρηση X^{new} προκύπτει από τον τύπο:

$$\hat{Y}_L = \hat{Y}^c - \hat{Y}^R/2 \text{ και } \hat{Y}_U = \hat{Y}^c + \hat{Y}^R/2$$

όπου $(\hat{Y}^c, \hat{Y}^R) = X^{new} \hat{\beta}$.

Είναι φανερό ότι τόσο η CRM, όσο και η BCRM είναι καλύτερες σε σχέση με την μέθοδο κέντρου, αφού λαμβάνουν υπόψιν και το κέντρο και το εύρος των δεδομένων σε μορφή διαστημάτων. Ωστόσο, δεν εγγυώνται καμία από τις δύο μεθόδους ότι το κάτω όριο θα είναι μικρότερο από το άνω όριο του διαστήματος.

2.2.4 Περιορισμένη μέθοδος

Η περιορισμένη μέθοδος (Constrained Method- Lima Neto, 2005,2010) προτάθηκε από τον Lima Neto για να λύσει το πρόβλημα που αναφέρθηκε παραπάνω, δηλαδή το κάτω όριο να μην έχει υψηλότερη τιμή από το άνω όριο. Από την τεχνική αυτή προκύπτει ένα μοντέλο παλινδρόμησης με την προϋπόθεση ότι όλες οι παράμετροι να είναι θετικές τιμές. Η μέθοδος μπορεί να συνδυαστεί και με την μέθοδο κέντρου (CONCM method) και με την μέθοδο κέντρου και εύρους (CONCRM method). Επομένως, τα μοντέλα είναι ίδια με αυτά που αναφέρθηκαν πάνω για τις CM και CRM μεθόδους, με μόνη διαφορά να υπάρχει ο περιορισμός να είναι θετικές όλες οι τιμές του διανύσματος β .

Για να εξασφαλιστεί ότι οι τιμές κατά τον υπολογισμό των ελαχίστων τετραγώνων του διανύσματος β είναι θετικές, ο Neto χρησιμοποίησε έναν αλγόριθμο που πρότειναν οι Lawson και Hanson (Lawson και Hanson, 1974). Ο αλγόριθμος αναγνωρίζει τις τιμές που δεν είναι συμβατές με τους περιορισμούς και τις μετατρέπει σε μη-αρνητικές τιμές, μέσω μιας διαδικασίας επαναυπολογισμού των βαρών. Η σύγκλιση του αλγορίθμου αυτού έχει αποδειχθεί από τους Lawson και Hanson. Το μειονέκτημα της μεθόδου είναι ότι υποχρεώνει τους συντελεστές να είναι μη-αρνητικοί, που σε κάποιες περιπτώσεις δεν μπορεί να αντιπροσωπεύσει την σχέση μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών.

3.Υλοποίηση

Σε αυτό το κεφάλαιο παρουσιάζονται οι δύο τεχνικές παλινδρόμησης που χρησιμοποιήθηκαν σε δεδομένα υπό μορφή διαστημάτων. Το σύνολο δεδομένων (dataset) περιγράφει την εξέλιξη της ζήτησης των τιμών ηλεκτρικής ενέργειας για τον μήνα Νοέμβρη στο δίκτυο διανομής ηλεκτρικής ενέργειας στην Ισπανία, σε MW. Τα χαρακτηριστικά του συνόλου είναι η πραγματική, η αναμενόμενη και η προγραμματισμένη ζήτηση. Τα δεδομένα δίνονται με τη μορφή ενός πίνακα όπως φαίνεται στην Εικόνα 3.1. Κάθε γραμμή του πίνακα αντιπροσωπεύει ένα δεκάλεπτο της ημέρας και αποτελείται από 4320 γραμμές. Η πρώτη τεχνική που εφαρμόσαμε στο σύνολο δεδομένων είναι η τεχνική Lasso-IR (Giordani, 2014) του Paolo Giordani, ενώ η δεύτερη είναι η τεχνική των k-Πλησιέστερων Γειτόνων.

3.1 Σύνολο δεδομένων

Ωρα	Πραγματικός	Αναμενόμενος	προγραμματισμένος
2021-11-01 21:00	26.355	26.567	26.279
2021-11-01 21:10	26.623	26.477	26.279
2021-11-01 21:20	26.387	26.373	26.279
2021-11-01 21:30	26.171	26.256	26.279
2021-11-01 21:40	25.978	26.125	26.279
2021-11-01 21:50	25.566	25.981	26.279
2021-11-01 22:00	25.164	25.823	25.003
2021-11-01 22:10	24.875	25.630	25.003
2021-11-01 22:20	24.821	25.379	25.003
2021-11-01 22:30	24.352	25.070	25.003
2021-11-01 22:40	24.084	24.703	25.003
2021-11-01 22:50	23.928	24.279	25.003
2021-11-01 23:00	23.650	23.795	22.643
2021-11-01 23:10	23.246	23.327	22.643
2021-11-01 23:20	23.246	22.945	22.643
2021-11-01 23:30	22.864	22.649	22.643
2021-11-01 23:40	22.415	22.440	22.643
2021-11-01 23:50	22.399	22.317	22.643
2021-11-02 00:00	22.104	22.280	22.013

Εικόνα 3.1. Δείγμα του συνόλου δεδομένων. (Seguimiento de la demanda de energía eléctrica, 2022)

Για την παλινδρόμηση χρησιμοποιήσαμε τις αναμενόμενες και τις πραγματικές τιμές. Η ανεξάρτητη μεταβλητή είναι η αναμενόμενη ζήτηση, ενώ εξαρτημένη η πραγματική ζήτηση. Μια μέρα αποτελείται από 144 γραμμές, καθώς μια γραμμή αντιπροσωπεύει ένα δεκάλεπτο μιας ημέρας. Τα διαστήματα προκύπτουν θέτοντας σαν κάτω όριο του διαστήματος την ελάχιστη τιμή ζήτησης μιας ημέρας, ενώ σαν άνω όριο την μέγιστη τιμή μιας ημέρας. Το νέο σύνολο δεδομένων θα αποτελείται από 2 χαρακτηριστικά και 30 γραμμές.

Ημέρα	Αναμενόμενη Ζήτηση	Πραγματική Ζήτηση
1/11/2021	19042	18912
2/11/2021	19787	19704
3/11/2021	21797	21689
4/11/2021	22172	21958
5/11/2021	22267	22286
6/11/2021	21716	21637
7/11/2021	20181	20213
8/11/2021	21049	20776
9/11/2021	22882	22700
10/11/2021	22890	22689
11/11/2021	22989	22873
12/11/2021	23212	23000
13/11/2021	20181	20213
14/11/2021	20249	20156
15/11/2021	21279	21166
16/11/2021	23219	22815
17/11/2021	23052	22783
18/11/2021	23082	23166
19/11/2021	23499	23426
20/11/2021	22379	22138
21/11/2021	20559	20362
22/11/2021	21391	21062
23/11/2021	23137	22938
24/11/2021	23492	23450
25/11/2021	23654	23361
26/11/2021	23897	23689
27/11/2021	23050	22792
28/11/2021	21420	21266
29/11/2021	22136	22062
30/11/2021	24044	24133

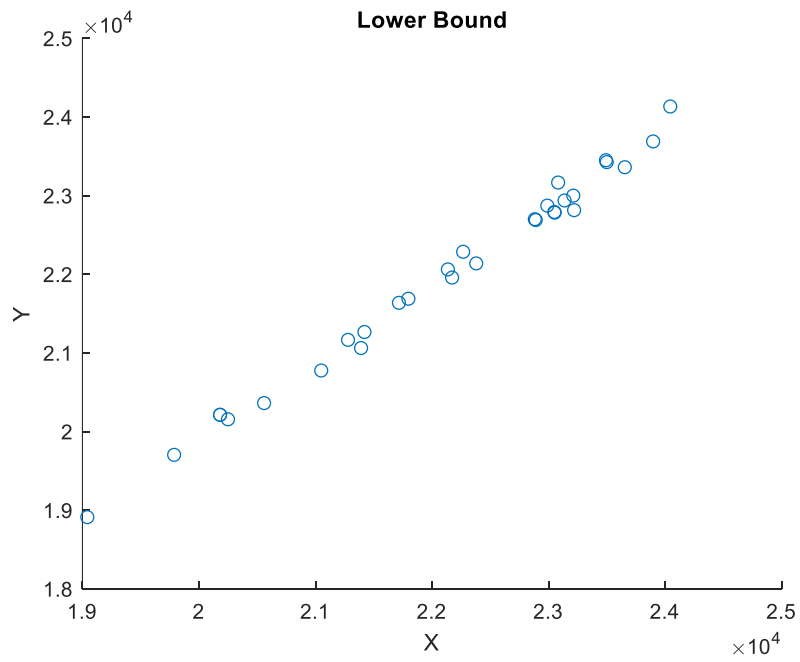
Πίνακας 3.1 Ελάχιστες τιμές ζήτησης ανά ημέρα

Ημέρα	Αναμενόμενη Ζήτηση	Πραγματική Ζήτηση
1/11/2021	26665	26740
2/11/2021	31295	31579
3/11/2021	32069	32224
4/11/2021	32449	32639
5/11/2021	32146	32066
6/11/2021	29352	29657
7/11/2021	29369	29522
8/11/2021	33428	33471
9/11/2021	33454	33492
10/11/2021	33719	33783
11/11/2021	33658	33695
12/11/2021	32347	32541
13/11/2021	33719	33783
14/11/2021	29289	29528
15/11/2021	33749	33793
16/11/2021	34363	34584
17/11/2021	34579	34526
18/11/2021	34448	34611
19/11/2021	33532	33490
20/11/2021	30163	30306
21/11/2021	29604	29800
22/11/2021	34756	34836
23/11/2021	35588	35726
24/11/2021	35854	35909
25/11/2021	36101	35982
26/11/2021	34887	34998
27/11/2021	32300	32500
28/11/2021	33120	33029
29/11/2021	36294	36455
30/11/2021	36446	36348

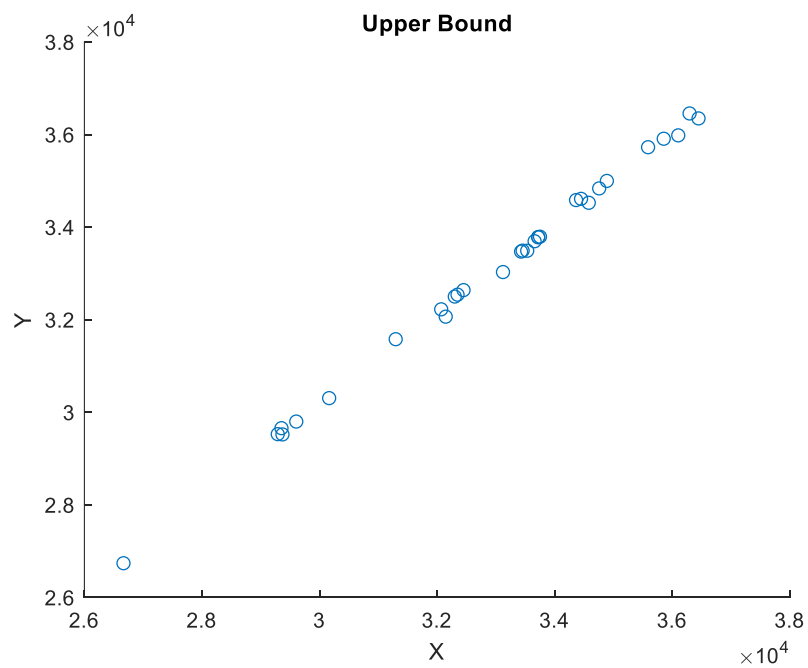
Πίνακας 3.2 Μέγιστες τιμές ζήτησης ανά ημέρα

Ημέρα	Αναμενόμενη Ζήτηση	Πραγματική Ζήτηση
1/11/2021	[19042, 26665]	[18912, 26740]
2/11/2021	[19787, 31295]	[19704, 31579]
3/11/2021	[21797, 32069]	[21689, 32224]
4/11/2021	[22172, 32449]	[21958, 32639]
5/11/2021	[22267, 32146]	[22286, 32066]
6/11/2021	[21716, 29352]	[21637, 29657]
7/11/2021	[20181, 29369]	[20213, 29522]
8/11/2021	[21049, 33428]	[20776, 33471]
9/11/2021	[22882, 33454]	[22700, 33492]
10/11/2021	[22890, 33719]	[22689, 33783]
11/11/2021	[22989, 33658]	[22873, 33695]
12/11/2021	[23212, 32347]	[23000, 32541]
13/11/2021	[20181, 33719]	[20213, 33783]
14/11/2021	[20249, 29289]	[20156, 29528]
15/11/2021	[21279, 33749]	[21166, 33793]
16/11/2021	[23219, 34363]	[22815, 34584]
17/11/2021	[23052, 34579]	[22783, 34526]
18/11/2021	[23082, 34448]	[23166, 34611]
19/11/2021	[23499, 33532]	[23426, 33490]
20/11/2021	[22379, 30163]	[22138, 30306]
21/11/2021	[20559, 29604]	[20362, 29800]
22/11/2021	[21391, 34756]	[21062, 34836]
23/11/2021	[23137, 35588]	[22938, 35726]
24/11/2021	[23492, 35854]	[23450, 35909]
25/11/2021	[23654, 36101]	[23361, 35982]
26/11/2021	[23897, 34887]	[23689, 34998]
27/11/2021	[23050, 32300]	[22792, 32500]
28/11/2021	[21420, 33120]	[21266, 33029]
29/11/2021	[22136, 36294]	[22062, 36455]
30/11/2021	[24044, 36446]	[24133, 36348]

Πίνακας 3.3 Διαστήματα ζήτησης ανά ημέρα



Εικόνα 3.2. Διάγραμμα διασποράς των κάτω ορίων των διαστημάτων.



Εικόνα 3.3. Διάγραμμα διασποράς των άνω ορίων των διαστημάτων.

Από τις Εικόνες 3.2 και 3.3 είναι φανερό ότι η εξαρτημένη και η ανεξάρτητη μεταβλητή έχουν γραμμική σχέση μεταξύ τους. Για την εφαρμογή των τεχνικών παλινδρόμησης παρακάτω, το σύνολο δεδομένων θα χωριστεί τυχαία σε ένα σύνολο εκπαίδευσης 24 γραμμών και σε ένα σύνολο ελέγχου με τις υπόλοιπες 6 γραμμές.

3.2 Lassoir

Η τεχνική Lasso-IR είναι μια τεχνική γραμμικής παλινδρόμησης που πρότεινε ο Paolo Giordani. Πρόκειται για μια τεχνική, στην οποία δημιουργούνται δύο μοντέλα παλινδρόμησης, ένα για τα κέντρα του διαστήματος και ένα για τις ακτίνες του διαστήματος. Το κύριο χαρακτηριστικό αυτής της τεχνικής είναι ότι οι συντελεστές του μοντέλου παλινδρόμησης των κέντρων υπολογίζονται με τέτοιον τρόπο, ώστε να είναι κοντά στους συντελεστές με το μοντέλο παλινδρόμησης των ακτίνων. Αυτό επιτυγχάνεται, θέτοντας ένα κατώφλι που αναπαριστά το μέγιστο επίπεδο διακύμανσης στις τιμές των συντελεστών μεταξύ των δύο μοντέλων παλινδρόμησης, που θυμίζει την κλασική τεχνική Lasso.

Στην δημοσίευση του Giordani δίνεται υλοποίηση της τεχνικής σε κώδικα Matlab, ο οποίος θα χρησιμοποιηθεί.

```
function[bMh,bAh,bRh,yMh,yRh,lfv,cpt]=lassoir(yM,yR,XM,XR,  
theta,rs,t);  
% Lasso-based interval-valued regression  
% Note: lassoir requires lsqlin (MATLAB Optimization  
Toolbox)  
% Input:  
% yM: vector of the midpoints of the dependent variable  
% yR: vector of the radii of the dependent variable  
% XM: matrix of the midpoints of the independent variables  
% XR: matrix of the radii of the independent variables  
% rs: number of random starts  
% t: shrinkage parameter for the lasso penalization term  
% Output:  
% bMh: vector of the estimated coefficients for the  
midpoints
```

```

% bAh: vector of the estimated additive coefficients
% bRh: vector of the estimated coefficients for the radii
% yMh: vector of the estimated midpoints of the dependent
variable
% yRh: vector of the estimated radii of the dependent
variable
% lfv: vector of the loss function values
% cpt: vector of the computation times
eps=10^-10;
fopt=10^6;
[n,p]=size(XM);
CONST=ones(2^p,p);
options=optimset('Display','off','LargeScale','off');
for col=1:p
    CONST(1:2^(p-col),col)=-ones(2^(p-col),1);
    inc=1;
    while 2^(p-col)+2*inc*2^(p-col)<=2^p;
        CONST(1+2*inc*2^(p-col):2^(p-col)+2*inc*2^(p-
col),col)=-ones(2^(p-col),1);
        inc=inc+1;
    end
end
for st=1:rs
    bA=rand(p,1);
    if sum(bA)>t
        bA=bA/sum(bA)*t;
    end
    func=10^6;
    fold=func+2*eps*func;
    iter=1;
    tic;
    while abs(fold-func)>eps*func
        fold = func;
        y=[yM; (yR-XR*bA)*theta^.5];
        X=[XM; XR*theta^.5];
        %bM=inv(X'*X)*X'*y;
        bM = inv(X'*X)*X'*y;

        G=[-XR; CONST];
        h=[XR*bM; t*ones(2^p,1)];
        %bRold=bR;
        [bA,nr,r,ef]=lsqlin(XR,yR-
XR*bM,G,h,[],[],[],[],options);
        if ef==-2
            bA=bAold; func=sum((yM-
XM*bM).^2)+theta*sum((yR-XR*(bM+bA)).^2);

```

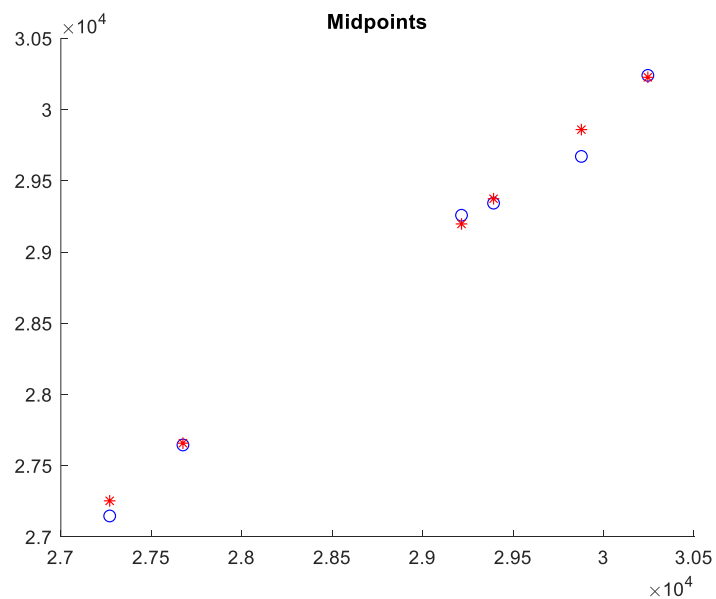
```

        fold=func;
    else
        iter=iter+1;
        func=sum((yM-XM*bM).^2)+theta*sum((yR-
XR*(bM+bA)).^2);
    end
end
cpt(st)=toc;
lfv(st)=func;
if func<fopt
    fopt=func;
    bMh=bM;
    bAh=bA;
    bRh=bMh+bAh;
end
end
yMh=XM*bMh;
yRh=XR*bRh;

end

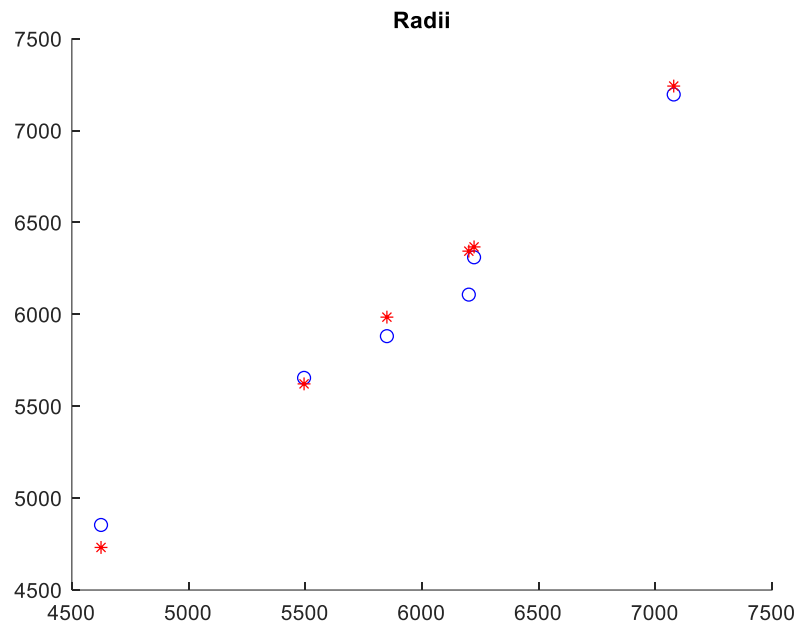
```

Στις παρακάτω εικόνες απεικονίζονται τα αποτελέσματα της πρόβλεψης σε σχέση με τις πραγματικές τιμές για διάφορες παραμέτρους. Στην εικόνα 3.4 μπορεί να παρατηρηθεί ότι το σφάλμα της πρόβλεψης στα κέντρα είναι πολύ μικρό.



Εικόνα 3.4. Εκτιμήτριες για τα κέντρα των διαστημάτων

Στην εικόνα 3.5 φαίνεται να υπάρχει μεγαλύτερη απόκλιση στις ακτίνες των διαστημάτων. Στις περιπτώσεις που η πρόβλεψη έχει μικρότερη ακτίνα από την πραγματική, χάνονται δεκτές τιμές, ενώ στην περίπτωση που η πρόβλεψη έχει μεγαλύτερη ακτίνα, μέσα σε όλες τις αποδεκτές τιμές του διαστήματος, υπάρχουν και μη αποδεκτές. Συνολικά όμως, οι προβλέψεις παραμένουν ικανοποιητικές.



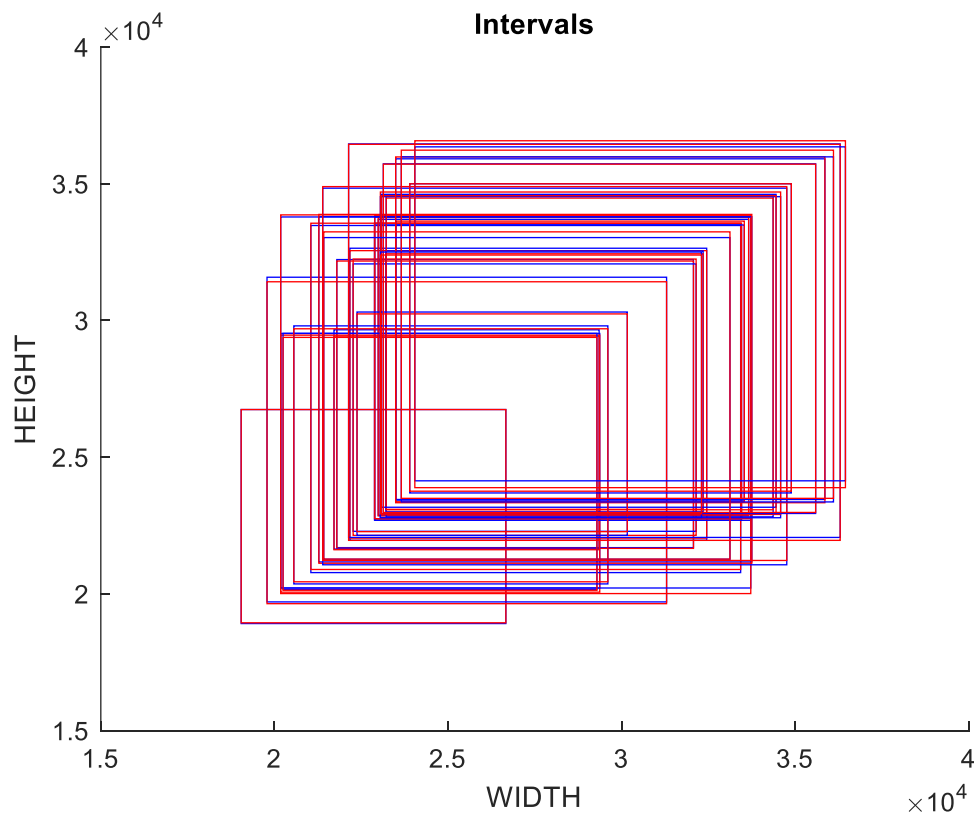
Εικόνα 3.5 Εκτιμήτριες για τις ακτίνες των διαστημάτων

Όπως αναφέρθηκε πιο πάνω, ενώ φαίνεται να υπάρχει μεγάλη απόκλιση στις ακτίνες (κατά συνέπεια το εύρος του διαστήματος), υπολογίζοντας την μέση τιμή σφάλματος, τα αποτελέσματα φαίνεται να είναι πολύ ικανοποιητικά.

```
error on midpoints is: 0.0023828
error on radius is: 0.017105
..
```

Εικόνα 3.6 Συνολικά σφάλματα για τη μέθοδο Lassoir

Μια πιο πλήρης εικόνα θα μπορούσε να επιτευχθεί, συγκρίνοντας τα διαστήματα που προβλέπει η τεχνική Lassoir με τα πραγματικά διαστήματα. Η σύγκριση απεικονίζεται στην παρακάτω εικόνα. Ένα διάστημα σε έναν άξονα μπορεί να αναπαρασταθεί με μια ευθεία, ενώ στις 2 διαστάσεις με ένα παραλληλόγραμμα. Τα παραλληλόγραμμα που προκύπτουν από την πρόβλεψη καλύπτουν σχεδόν πλήρως τα πραγματικά διαστήματα, το οποίο σημαίνει ότι η πρόβλεψη είναι πολύ καλή.



Εικόνα 3.7 Οι προβλέψεις (κόκκινο χρώμα) σε σύγκριση με τα πραγματικά διαστήματα (μπλε χρώμα)

3.3 Πλησιέστεροι Γείτονες

Σαν συνάρτηση απόστασης για τη τεχνική των πλησιέστερων γειτόνων χρησιμοποιήθηκε η απόσταση Hausdorff, στην οποία αν:

$$A = ([\alpha_1^1, \beta_1^1], [\alpha_1^2, \beta_1^2], \dots, [\alpha_1^\nu, \beta_1^\nu])$$

και

$$B = ([\alpha_2^1, \beta_2^1], [\alpha_2^2, \beta_2^2], \dots, [\alpha_2^\nu, \beta_2^\nu])$$

δύο διαστήματα n διαστάσεων, τότε η απόσταση ορίζεται ως εξής:

$$d(A, B) = \sum_{i=1}^n d([\alpha_1^i, \beta_1^i], [\alpha_2^i, \beta_2^i]) = \sum_{i=1}^n \max(|\alpha_1^i - \alpha_2^i|, |\beta_1^i - \beta_2^i|)$$

Η υλοποίηση της μεθόδου των k -πλησιέστερων γειτόνων δίνεται στη συνέχεια.

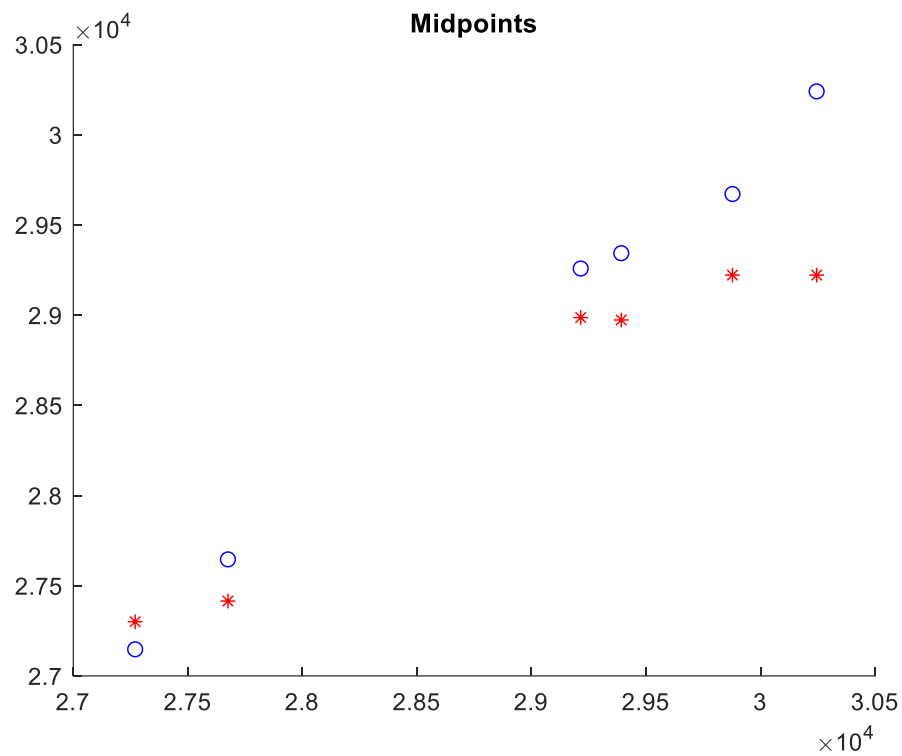
```
function [yMh, yRh] = knn_intervals(K, yM, yR, XM, XR,
input_xM, input_xR);
%Input:
% K: number of neighbors(hyperparameter)
% yM: vector of the midpoints of the dependent variable
% yR: vector of the radii of the dependent variable
% XM: matrix of the midpoints of the independent variables
% XR: matrix of the radii of the independent variables
% input_xR: radius of interval to predict
% input_XM: midpoint of interval to predict
%Output:
% yMh: value of the estimated midpoint of the dependent
variable
% yRh: value of the estimated radius of the dependent
variable

yM1 = zeros(K,1);
yR1 = zeros(K,1);
L_distance = abs((XM-XR)-(input_xM-input_xR));
U_distance = abs((XR+XM)-(input_xR+input_xM));
distances = max(L_distance, U_distance);

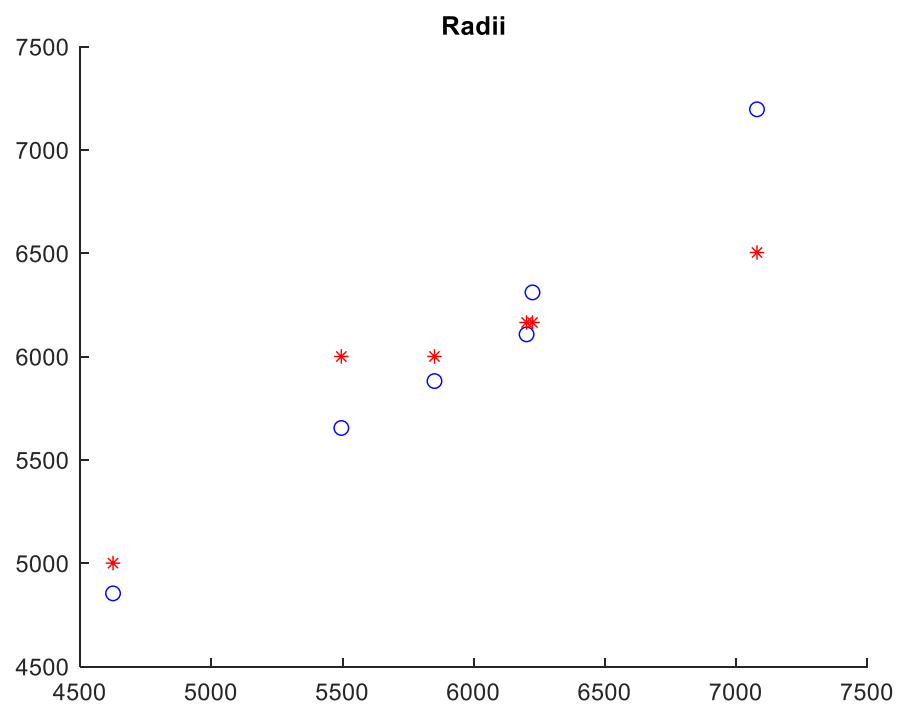
[~, idx] = sort(distances);

for i = 1:K
    yM1(i) = yM(idx(i));
    yR1(i) = yR(idx(i));
end

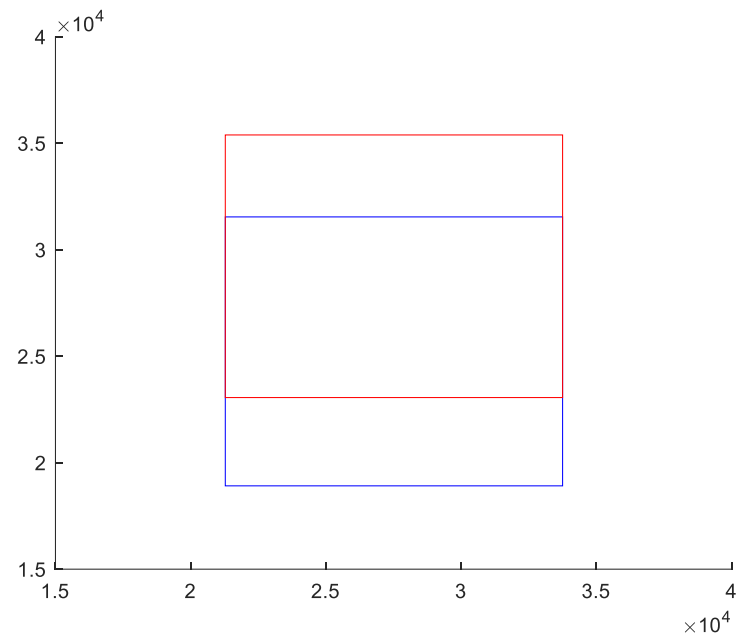
yMh = mean(yM1);
yRh = mean(yR1);
end
```



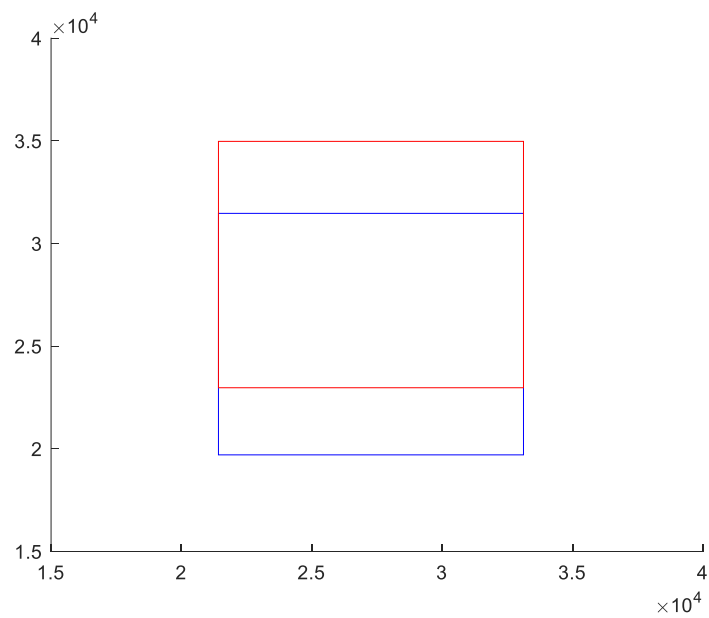
Εικόνα 3.8. Εκτιμήτριες για τα κέντρα, με βάση την τεχνική πλησιέστερων γειτόνων



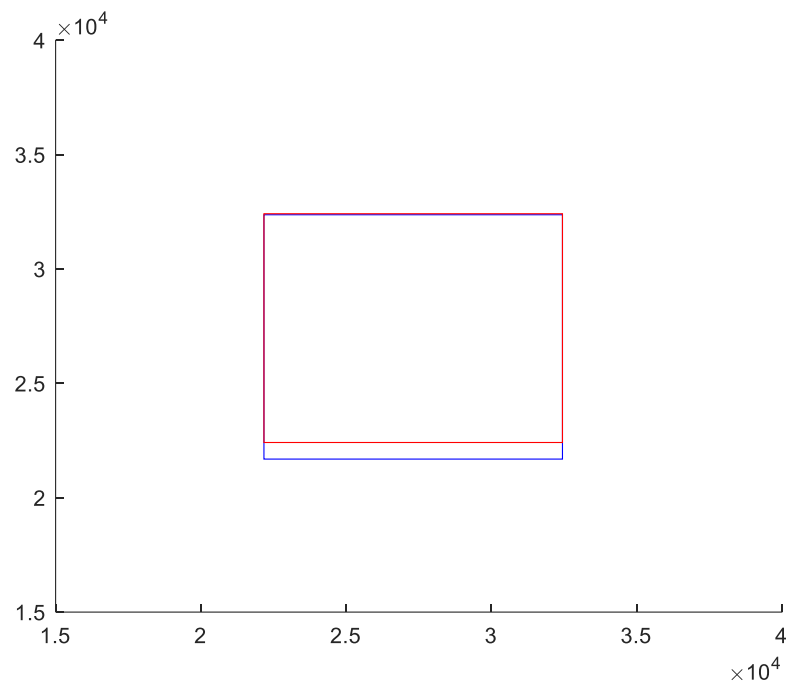
Εικόνα 3.9. Εκτιμήτριες για τις ακτίνες, με βάση την τεχνική πλησιέστερων γειτόνων



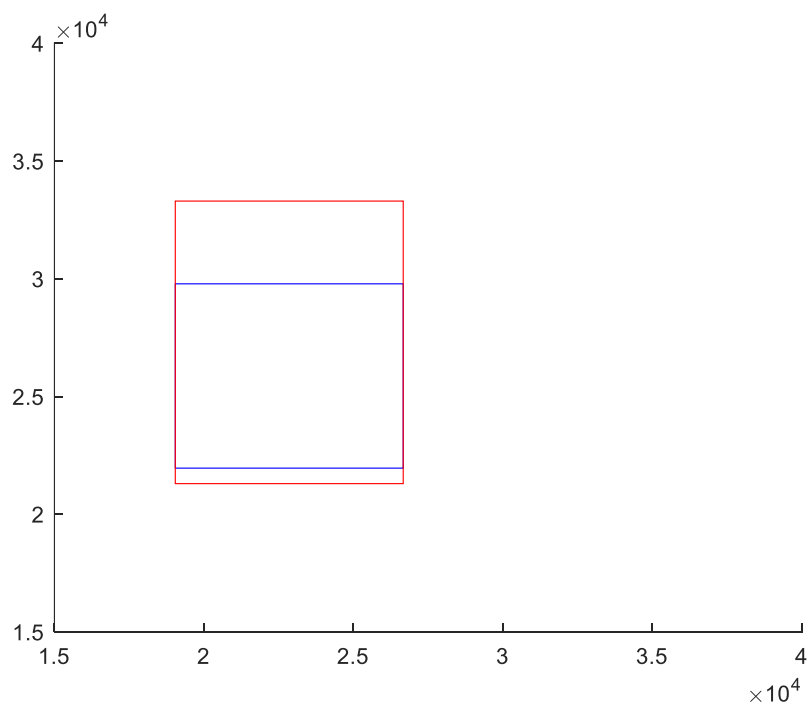
Εικόνα 3.10 Η πρόβλεψη (κόκκινο) του πρώτου στοιχείου σε σχέση με το πραγματικό διάστημα (μπλε).



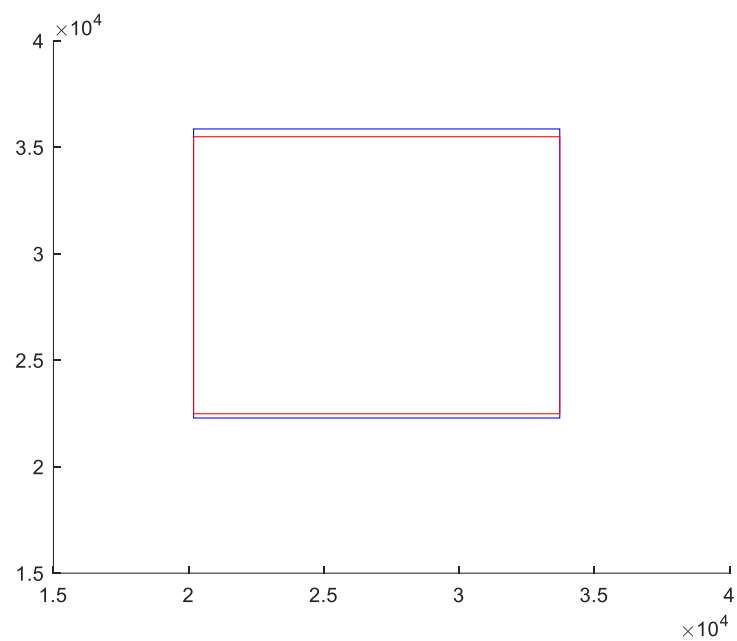
Εικόνα 3.11 Η πρόβλεψη (κόκκινο) του δεύτερου στοιχείου σε σχέση με το πραγματικό διάστημα (μπλε).



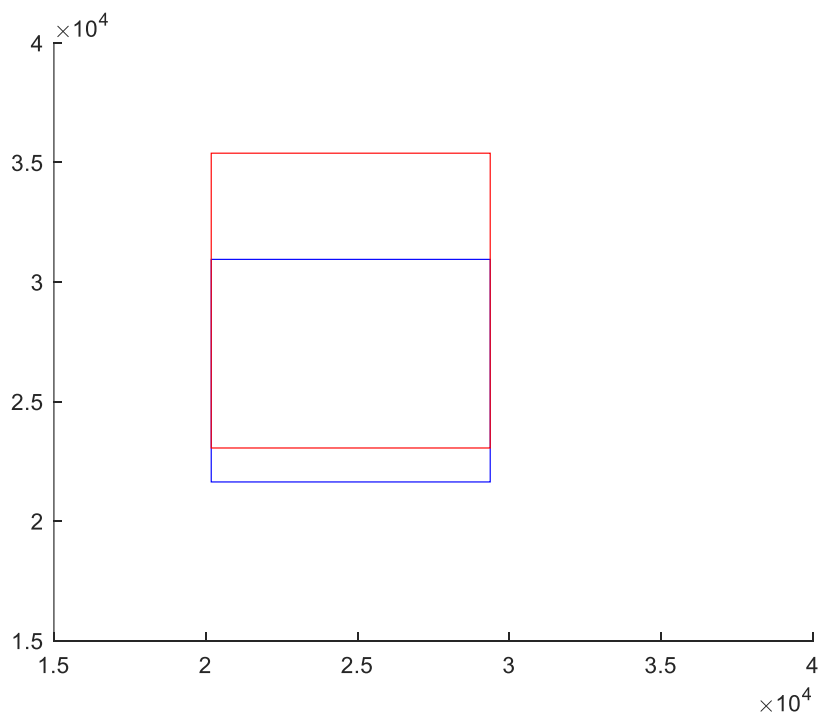
Εικόνα 3.12 Η πρόβλεψη (κόκκινο) του τρίτου στοιχείου σε σχέση με το πραγματικό διάστημα (μπλε).



Εικόνα 3.13. Η πρόβλεψη (κόκκινο) του τέταρτου στοιχείου σε σχέση με το πραγματικό διάστημα (μπλε).



Εικόνα 3.14. Η πρόβλεψη (κόκκινο) του πέμπτου στοιχείου σε σχέση με το πραγματικό διάστημα (μπλε).



Εικόνα 3.15 Η πρόβλεψη (κόκκινο) του έκτου στοιχείου σε σχέση με το πραγματικό διάστημα (μπλε).

Από τις εικόνες, φαίνεται οι προβλέψεις της τεχνικής των k- Πλησιέστερων γειτόνων να μην είναι τόσο αποτελεσματικές. Υπάρχουν μεγάλες αποκλίσεις τόσο στο εύρος, όσο και στα κέντρα των διαστημάτων. Η απόκλιση είναι αναμενόμενη, καθώς για την απόσταση μεταξύ δύο δεδομένων, λαμβάνεται υπόψιν μόνο η διαφορά των άκρων των διαστημάτων, με αποτέλεσμα να χάνονται σημαντικές πληροφορίες για τα διαστήματα. Άλλες τεχνικές μέτρησης για τον υπολογισμό απόστασης μεταξύ διαστημάτων υπάρχουν αλλά δεν ήταν δυνατό να χρησιμοποιηθούν στα διαθέσιμα χρονικά όρια για την πτυχιακή αυτή εργασία., Έτσι ενδέχεται η τεχνική αυτή, να μην είναι ιδιαίτερα αποδοτική για τον υπολογισμό παλινδρόμησης σε διαστηματικά δεδομένα με τη χρήση της τεχνικής των k-Πλησιέστερων Γειτόνων.

4. Συμπεράσματα και Μελλοντική Εργασία

Ο σκοπός της παρούσας πτυχιακής εργασίας ήταν η μελέτη και η εφαρμογή της παλινδρόμησης σε αβέβαια δεδομένα. Η αβεβαιότητα αντιμετωπίστηκε με διαστήματα, που αποτελούν το βασικότερο μοντέλο αναπαράστασης της αβεβαιότητας και της περιγραφής συνοπτικών πληροφοριών στον κλάδο της ανάλυσης των συμβολικών δεδομένων. Οι τεχνικές παλινδρόμησης σε διαστηματικά δεδομένα που μελετήσαμε θεωρητικά είναι η μέθοδος κέντρου, η μέθοδος κέντρου και εύρους, η διμεταβλητή μέθοδος κέντρου και εύρους και η περιορισμένη μέθοδος.

Οι μέθοδοι που υλοποιήθηκαν είναι η τεχνική Lassoir του Giordani, που είναι μια παραλλαγή της μεθόδου εύρους και κέντρου με έμπνευση από την κλασική στατιστική τεχνική παλινδρόμησης Lasso, καθώς επίσης και η τεχνική των k-Πλησιέστερων Γειτόνων με μέτρο απόστασης μεταξύ των διαστημάτων την απόσταση Hausdorff. Ως σύνολο δεδομένων για το πειραματικό σκέλος της υλοποίησης χρησιμοποιήθηκαν οι τιμές ζήτησης ηλεκτρικής ενέργειας από την ιστοσελίδα του Διαχειριστή δικτύου διανομής ηλεκτρικής ενέργειας της Ισπανίας, όπου ως διάστημα ορίστηκε η ελάχιστη και η μέγιστη τιμή ζήτησης κάθε ημέρας για όλες τις μέρες του μήνα Νοέμβρη του 2021.

Η αντιμετώπιση των αβέβαιων δεδομένων με συμβολικά δεδομένα, και συγκεκριμένα με διαστήματα φαίνεται να απέδωσε εξαιρετικά, καθώς οι προβλέψεις της τεχνικής Lassoir είχαν ελάχιστη έως και ανύπαρκτη διαφορά με τις πραγματικές τιμές. Από την άλλη, η απόκλιση των προβλέψεων της τεχνικής των k -πλησιέστερων γειτόνων ήταν αναμενόμενη, αφού η τεχνική είναι αναποτελεσματική σε δεδομένα μεγάλης κλίμακας και η μέτρηση απόστασης βασίζεται μόνο στην διαφορά μεταξύ των ορίων των διαστημάτων, χάνοντας σημαντικές πληροφορίες για τα διαστήματα.

Ο συνδυασμός των προτεινόμενων μεθόδων με άλλες τεχνικές παλινδρόμησης, είναι πολλά υποσχόμενη για αβέβαια δεδομένα που μπορούν να αναπαρασταθούν διαστηματικά και θα ήταν μια ενδιαφέρουσα μελλοντική έρευνα. Επιπλέον, η γενίκευση των μεθόδων αυτών και σε μη-γραμμικά μοντέλα παλινδρόμησης είναι εξίσου σημαντική μελλοντική έρευνα, καθώς σε πολλά φαινόμενα δεν υπάρχει γραμμική συσχέτιση.

Καταλητικά, ο συνδυασμός της Διαστηματικής Ανάλυσης με την παλινδρόμηση είναι μια σημαντική περιοχή έρευνας, αφού πέρα από το υψηλό ποσοστό ακρίβειας που δίνει για τα αποτελέσματα, προσφέρει και όλα τα πλεονεκτήματα της Ανάλυσης Διαστημάτων για την Ανάλυση Συμβολικών Δεδομένων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Aien, M., Hajebrahimi, A. and Fotuhi-Firuzabad, M., 2016. A comprehensive review on uncertainty modeling techniques in power system studies. *Renewable and Sustainable Energy Reviews*, 57, pp.1077-1089.

Bertrand, P. and Goupil, F. (2000): Descriptive Statistics for Symbolic Data. Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data (eds. H.-H. Bock and E. Diday). Springer-Verlag, Berlin, pp. 103-124.

Billard, L. (2007). Dependencies and Variation Components of Symbolic Interval-Valued Data. Selected Contributions in Data Analysis and Classification. Springer-Verlag, Berlin, pp. 3-13.

Billard, L. (2008). Sample Covariance Functions for Complex Quantitative Data. Processing, World Conferences International Association of Statistical Computing 2008, Yokohama, Japan.

Billard, L. and Diday, E. (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association* 98, pp. 470- 487.

Billard, L. and Diday, E. (2006). Descriptive Statistics for Interval-valued Observations in the Presence of Rules. *Computational Statistics*, 21, pp. 187-210.

Billard, L. and Diday, E. (2007). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley, Chichester.

Chacón, J. and Rodríguez, O., 2021. Regression Models for Symbolic Interval-Valued Variables. *Entropy*, 23(4), p.429.

de Carvalho F.A.T., Lima Neto, E.A. and Tenorio, C.P. (2004). A New Method to Fit a Linear Regression Model for Interval-valued Data. *Lecture Notes in Computer Science, KI2004 Advances in Artificial Intelligence*. Springer-Verlag, pp. 295-306.

Diday, E. (1987). Introduction `a l'Approache Symbolique en Analyse des Donn´ees. Premi`eres Journ´ees Symbolique - Num´erique. CEREMADE, Universit´e Paris, pp. 21-56.

Diday, E. (1995). Probabilist, Possibilist and Belief Objects for Knowledge Analysis. *Annals of Operations Research*, 55, pp. 227-276.

Diday, E. and Emilion, R. (1996). Lattices and Capacities in Analysis of Probabilist Objects. *Studies in Classification* (eds. E. Diday, Y. Lechevallier, and O. Opilz), pp.13-30.

Diday, E. and Emilion, R. (1998). Capacities and Credibilities in Analysis of Probabilistic Objects by Histograms and Lattices. In: *Data Science, Classification, and Related Methods* (eds. C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, and Y. Baba), pp. 353-357.

Diday, E., Emilion, R. and Hillali, Y. (1996). Symbolic Data Analysis of Probabilistic Objects by Capacities and Credibilities. *Societ`e Italiana di Statistica*, pp. 5-22.

Education, I., 2022. *What are Neural Networks?*. [online] Ibm.com. Available at: <<https://www.ibm.com/cloud/learn/neural-networks>> [Accessed 10 February 2022].

Emilion, R. (1997). Diff´erentiation des Capacit´es. *Comptes Rendus de l'Academie des Sciences - Series I - Mathematics*, 324, pp. 389-392.

En.wikipedia.org. 2022. *Logistic function - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/Logistic_function> [Accessed 10 February 2022].

Fisher RA. 1922. On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A Containing Papers of a Mathematical or Physical Character*, pp. 309–368.

Fisher RA. 1925. Theory of statistical estimation. *Proc. Cambridge Philos. Soc.*, pp. 700–725.

Galton, F. (1886). "Regression towards mediocrity in hereditary stature". *The Journal of the Anthropological Institute of Great Britain and Ireland*, pp. 246–263.

Gauss, C.F. (1809). *Theoria Motus Corporum Coelestium. Sectionibus Conicis Solem Ambientum*.

- Giordani, P., 2014. Lasso-constrained regression analysis for interval-valued data. *Advances in Data Analysis and Classification*, 9(1), pp.5-19.
- Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The elements of statistical learning*. pp.68-69
- Hüllermeier, E. and Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, pp.457-506.
- Lawson, C. and Hanson, R. (1974). *Solving Least Squares Problems*. Prentice-Hall.
- Legendre, A.M. (1805). Sur la Méthode des moindres carrés. Nouvelles méthodes pour la détermination des orbites des comètes.
- Lima Neto, E.A and de Carvalho F.A.T. (2010). Constrained Linear Regression Models for Symbolic Interval-valued Variables. *Computational Statistics & Data Analysis*, 54(2), pp. 333-347.
- Lima Neto, E.A., de Carvalho F.A.T. and Freire, E.S. (2005). Applying Constrained Linear Aggression Models to Predict Interval-Valued Data. *Lecture Notes in Computer Science, KI: Advances in Artificial Intelligence* (ed. U. Furbach). Springer-Verlag, Berlin, pp. 92-106.
- Lima Neto, E.A., de Carvalho F.A.T. and Tenorio, C.P. (2004). Univariate and Multivariate Linear Regression Methods to Predict Interval-valued Features. *Lecture Notes in Computer Science, AI 2004 Advances in Artificial Intelligence*. Springer-Verlag, Berlin, 526-537.
- Medium. 2022. *A Gentle Introduction To Math Behind Neural Networks*. [online] Available at: <<https://towardsdatascience.com/introduction-to-math-behind-neural-networks-e8b60dbbdeba>> [Accessed 10 February 2022].
- Montgomery, D., Peck, E. and Vining, G., 2012. *Introduction to linear regression analysis*. 4th ed. p.1.
- Moore, R., Kearfott, R. and Cloud, M., 2009. *Introduction to interval analysis*. Philadelphia, PA: Society for Industrial and Applied Mathematics, pp.1-5.

Pearson, K. and Lee, A. (1903) On the Laws of Inheritance in Man: Inheritance of Physical Characters. *Biometrika*, pp. 357-462.

Demanda.ree.es. 2022. *Seguimiento de la demanda de energía eléctrica*. [online] Available at: <<https://demanda.ree.es/visiona/peninsula/demanda/tablas/2021-11-01/1>> [Accessed 10 February 2022].

Yule G U, 1897a, ``On the theory of correlation" *Journal of the Royal Statistical Society* , pp. 812 -854

Zadeh, L.A. (1965): Fuzzy Sets. *Information and Control* 8(3), pp. 338-353.

[Οπισθόφυλλο. Κενή σελίδα]