Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εργαστήριο Επεξεργασίας Πληροφορίας και Υπολογισμών

Διπλωματική
Εργασία

Αυτοματοποιημένος Αναλυτής Δεδομένων

Σχεδίαση και Υλοποίηση ενός Συστήματος Αυτόματης Εκπαίδευσης Μοντέλων Μηχανικής Μάθησης

Εκπόνηση
Νησιώτη Ελένη 7737

Επίβλεψη
Επίκ. Καθ. Ανδρέας Συμεωνίδης

Συνεπίβλεψη
Δρ. Χατζηδημητρίου Κυριάκος

6 Ιουνίου 2017

Περίληψη

Η επιστήμη της μηχανικής μάθησης έχει κατορθώσει, βασιζόμενη σε αυστηρά μαθηματικά εργαλεία, να μετατρέψει τη σύγχρονη πληροφοριακή αφθονία σε κατανόηση κοινωνικών, οικονομικών και φυσικών μηχανισμών, γεγονός που οδήγησε στην εκτεταμένη δημιουργία προβλεπτικών μοντέλων. Η εξέλιξή της, ωστόσο, περιορίζεται σήμερα από την ύπαρξη απαιτητικών προβλημάτων και την εγγενή αδυναμία μεταφερσιμότητας των πειραμάτων και μοντέλων σε νέα προβλήματα. Ανέκυψε λοιπόν η ανάγκη ανακάλυψης μονοπατιών που θα οδηγήσουν σε βαθύτερη κατανόηση των μηχανισμών της μηχανικής μάθησης, ώστε πλέον να εκπαιδεύουμε μοντέλα που βελτιστοποιούν την ίδια τη διαδικασία της μάθησης, και όχι μεμονωμένα προβλήματα. Το πεδίο του AutoML αναδύθηκε πρόσφατα από αυτήν την προσπάθεια και σύγχρονοι ορισμοί του αποδίδουν την αυτοματοποίηση της εφαρμογής μηχανικής μάθησης. Συναντάται κυρίως με τη μορφή εργαλείων λογισμικού, τα οποία υποβοηθούν τον αναλυτή δεδομένων, αναλαμβάνοντας την αναζήτηση αυτόματων λύσεων που καθιστούν την εφαρμογή μηχανικής μάθησης πιο αποδοτική και αποτελεσματική. Χαρακτηριστικό αυτών των συστημάτων είναι η ύπαρξη μετα-γνώσης, δηλαδή γνώσης προερχόμενης από την εφαρμογή μηχανικής μάθησης σε παρελθοντικά προβλήματα, η οποία προσδίδει εμπειρία και προσαρμοστικότητα στο σύστημα. Την υλοποίηση ενός τέτοιου εργαλείου έχει ως στόχο η παρούσα διπλωματική εργασία, καθώς αναγνωρίζει την ανάγκη έρευνας και επέκτασης των εφαρμογών του AutoML. Εκμεταλλευόμενοι σύγχρονες τεχνολογίες, όπως την πολυπληθή αποθήκη πακέτων της γλώσσας R, εξερευνήσαμε τις δυνατότητες τεχνικών μηχανικής μάθησης και επιχειρήσαμε να επεκτείνουμε την τρέχουσα κατάσταση ενσωματώνοντας στο σύστημά μας μετα-μάθηση για τη βελτιστοποίηση υπερ-παραμέτρων και ensembles με προς τα εμπρός επιλογή μοντέλων. Κυρίαρχο στόχο της εργασίας μας αποτέλεσε η σχεδίαση και υλοποίηση ενός έμπειρου, κατανοητού και επεκτάσιμου αυτόματου αναλυτή δεδομένων. Θεωρούμε πως η βιβλιογραφική έρευνα, τα πειράματα και το εργαλείο λογισμικού που υλοποιήθηκαν κατά τη διάρκεια της δουλειάς μας αποτελούν σημαντική συνεισφορά στο πεδίο του AutoML.

Diploma Thesis

Automated Data Scientist

Abstract

The science of machine learning has achieved, based on solid mathematical tools, to convert the current informational abundance into the understanding of social, economical and nature mechanisms that lead to the general creation of predictive models. Its evolution, however, has stumbled upon the presence of computationally demanding problems and an inherent lack of transferability of machine learning experiments to new applications. The necessity, therefore, came up of discovering paths that lead to a deeper understanding of the machine learning mechanismus, with the ambition of training models that optimize the very process of learning, instead of individual applications. The field of AutoML emerged recently through this attempt and contemporary definitions acknowledge to it the automation of applying machine learning. Its most apparent manifestations include software systems that serve as productivity tools, instruments to make experts more efficient and effective, but not eliminate them. A common feature of these systems is the embedding of meta-knowledge, namely knowledge produced by the application of machine learning in past experiments, a trait that adds experience and adaptability to the system. This Diploma Thesis aims at the implementation of a software tool belonging to the above described family, an idea sparked from the realization of AutoML's need for research and enrichment of its applications. Exploiting current technologies, susch as the rich CRAN repository, we explored opportunities offered by machine learning techniques and attempted to push forward the state of the art by embedding meta-learning for optimal hyperparameter selection and forward model selection ensembles to our system. Main aspiration of our work consisted in designing and implementing an experienced, intuitive and expandable automated data analyst. We deem that the academic research, experiments and software produced during our work constitute an informative contribution to the area of AutoML.

Eleni Nisioti
Intelligent Systems & Software Engineerin Labgroup
Electrical & Computer Engineering Department
Aristotle University of Thessaloniki
June 2017

Ευχαριστίες

Η εκπόνηση της διπλωματικής εργασίας ήταν μια γεμάτη και διδακτική εμπειρία που συμπεριέλαβε άτομα, τα οποία θα ήθελα να ευχαριστήσω σε αυτό το σημείο.

Τον επίκουρο καθηγητή κ. Ανδρέα Συμεωνίδη για την εμπιστοσύνη που μου έδειξε με την ανάθεση της διπλωματικής εργασίας, την κατανόηση που επέδειξε μέχρι το πέρας της και την ακαδημαϊκή στήριξή του.

Τον μεταδιδακτορικό ερευνητή κ. Κυριάκο Χατζηδημητρίου για την καθοδήγηση, την εμπιστοσύνη του στην κρίση μου και την εισαγωγή μου σε ένα ανεξερεύνητο επιστημονικό αντικειμένο.

Την οικογένειά μου για τη στήριξη και την αλόγιστη εμπιστοσύνη της στις αποφάσεις μου.

Τους φίλους μου για τις εμπειρίες των τελευταίων 6 χρόνων.

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

"*When you have eliminated the impossible, whatever remains, no matter how improbable, must be the truth.*"
— The Sign of Four

1

ΕΙΣΑΓΩΓΗ

1.1 Γενικά

Όταν ο Arthur Lee Samuel εισήγαγε τον όρο *μηχανική μάθηση*, το 1959, μάλλον δεν ανέμενε την ταχεία εξέλιξή του σε τομέα με τεράστιο επιστημονικό ενδιαφέρον, εμπορική σημασία και καθολική αναγνωρισιμότητα. Μία δραστήρια κοινότητα μαθηματικών, αναλυτών δεδομένων, μηχανικών και προγραμματιστών έχει τροφοδοτήσει, τη βιβλιογραφία με πληθώρα αλγορίθμων και τεχνικών μηχανικής μάθησης, την αγορά με εφαρμογές και την κοινωνία με τις δυνατότητες, ή απειλές, της Τεχνητής Νοημοσύνης.

Τα τελευταία χρόνια έχει κυριαρχήσει η εικόνα της παγίωσης των τεχνικών μηχανικής μάθησης. Η πληθώρα των διαθέσιμων δεδομένων και η αναγνώριση της επιστημονικής και εμπορικής αξίας τους έχει αναδείξει απαιτητικά προβλήματα μηχανικής μάθησης, τάση απέναντι στην οποία η κοινότητα ανταποκρινόταν με τη σχεδίαση νέων τεχνικών και αλγορίθμων. Σήμερα ωστόσο ένα μεγάλο μέρος της βιβλιογραφίας αναλώνεται στην προσπάθεια εύρεσης εξειδικευμένων λύσεων σε ιδιαίτερα προβλήματα που παρά τα οφέλη, δεν είναι επεκτάσιμα. Η ουσία της δυσλειτουργικότητας έγκειται στο γεγονός ότι δεν έχει παραχθεί γνώση χρήσιμη για την επιστήμη της μηχανικής μάθησης, καθώς η προσέγγιση που έχει ακολουθηθεί δεν είναι μεταφέρσιμη σε νέα προβλήματα. Η προσέγγιση αυτή, δεδομένης της εμπειρίας της κοινότητας, θα μπορούσε να χαρακτηριστεί αφελής, καθώς η εξατομικευμένη αντιμετώπιση κάθε νέου προβλήματος απαιτεί την καταβολή μεγίστου κόπου και πόρων με το ελάχιστο κέρδος για την επιστήμη της μηχανικής μάθησης. Προκύπτει λοιπόν το εξής ερώτημα: ήρθε η ώρα να περάσουμε σε ένα *νέο στάδιο μάθησης*;

Η μετα-μάθηση έλαβε υπόσταση το 1992, με την εμφάνιση των πρώτων συστημάτων Craw et al. [1] and Brazdil, Gama, and Henery [2], που επιχειρούσαν να αυτοματοποιήσουν στάδια της εξόρυξης δεδομένων, όπως η επιλογή αλγορίθμου μάθησης. Κλειδί στην προσέγγιση της μετα-μάθησης αποτελεί η προσπάθεια συλλογής εμπειρίας από ένα σύστημα με μορφή γνώσης, παραγόμενης από παρελθοντικά πειράματα. Πρόκειται για έναν ευρύ τομέα, που σήμερα αποτελεί εφαλτήριο για την εξέλιξη της μηχανικής μάθησης. Αν και βήματα προς την αναθεώρηση της συμβατικής εφαρμογής μηχανικής μάθησης γίνονται από το 1995 (Ενότητα 2.3.1), η συνειδητοποιημένη κινητοποίηση της κοινότητας προς την αυτοματοποίηση ης μηχανικής μάθησης ξεκίνησε πολύ αργότερα, με τους πρώτους διαγωνισμούς να κάνουν την εμφάνισή τους το 2011 [1]. Εν έτει 2017 η κοινότητα προσπαθεί να ορίσει τη νέα τάση στη μηχανική μάθηση, το *AutoML*.

Ο κλάδος του *AutoML*, πατώντας στην εμπειρία δεκαετιών προσπαθεί να αντιμετωπίσει τα απαιτητικά προβλήματα, που απασχολούν την τρέχουσα αγορά, με μία νέα προσέγγιση: μαθαίνοντας στους υπολογιστές να μαθαίνουν, όχι πλέον την επίλυση μεμονωμένων προβλημάτων, αλλά *την ίδια τη διαδικασία της μάθησης*.

[1] `http://automl.chalearn.org/`

1.2 Στόχοι

Μία σύντομη ματιά στη βιβλιογραφία του AutoML αποκαλύπτει την επιτακτικότητα της ανάγκης σχεδιασμού και υλοποίησης εργαλείων μηχανικής μάθησης, τα οποία υποβοηθούν τον αναλυτή δεδομένων. Η συνεισφορά αυτών των εργαλείων λογισμικού μπορεί να αναλυθεί σε δύο κύριους άξονες. 1. αναλαμβάνουν την αυτοματοποίηση χρονοβόρων και τετριμμένων διαδικασιών 2. επιστρατεύουν μηχανισμούς εξαγωγής μετα-γνώσης για την προσαρμοσμένη αντιμετώπιση κάθε νέου προβλήματος με βάση την εμπερία παλαιότερων πειραμάτων

Το No free Lunch Theorem έχει προβληματίσει και καθοδηγήσει την κοινότητα των επιστημόνων που ερευνούν τη βελτιστοποίηση γενική φύσης συναρτήσεων κόστους [3] και ειδικότερα την κοινότητα μηχανικής μάθησης [4]. Σύμφωνα με αυτό η μέση απόδοση ενός αλγορίθμου βελτιστοποίησης σε όλες τις πιθανές συναρτήσεις κόστους δεν εξαρτάται από τον αλγόριθμο. Στην περίπτωση λοιπόν που κάποιος επιδιώκει τη σχεδίαση ενός συστήματος βελτιστοποίησης προβλημάτων γενικής φύσεως, όπως εμείς, δεν έχει τη δυνατότητα της εκ των προτέρων επιλογής του αποδοτικότερου αλγορίθμου βελτιστοποίησης, καθώς όποια μέθοδο και να επιλέξει θα έχει κατά μέσο όρο την ίδια απόδοση και, επομένως, η σχεδίασή του στερείται νοήματος. Είναι λοιπόν λογικό να επιδιώκεται η ενσωμάτωση μετα-μάθησης στη διαδικασία, καθώς καθιστά το σύστημα βελτιστοποίησης εκπαιδεύσιμο και προσαρμόσιμο σε νέα προβλήματα.

Η παρούσα εργασία στοχεύει στην αναγνώριση και αντιμετώπιση κενών, καθώς και την εκμετάλλευση δυνατοτήτων στο χώρο του AutoML μέσω της υλοποίησης ενός συστήματος αυτόματης ανάλυσης δεδομένων. Το σύστημα θα αναλαμβάνει τη βελτιστοποίηση προβλημάτων δυαδικής ταξινόμησης έχοντας ως πρότυπο τη μεθοδολογία ενός πραγματικού αναλυτή δεδομένων και εφαρμόζοντας τεχνικές του AutoML.

1.3 Μεθοδολογία

Πρώτο βήμα στη προσέγγιση του προβλήματος αποτέλεσε ο εντοπισμός των σημείων στη διαδικασία της μηχανικής μάθησης που επιδέχονται και χρήζουν αυτοματοποίησης. Χαρακτηριστικά αυτών των σημείων είναι η χρονική και υπολογιστική επιβάρυνση και η αναγνώριση κάποιου μηχανισμού βελτιστοποίησης του προβλήματος μέσω μαθηματικής διατύπωσής του. Χαρακτηριστικό παράδειγμα αυτής της κατηγορίας είναι η επιλογή βέλτιστων υπερ-παραμέτρων κατά τη ρύθμιση ενός μοντέλου μηχανικής μάθησης.

Το σύστημά μας θέτει ιδιαίτερη βαρύτητα στην τεχνική με την οποία γίνεται η βέλτιστη επιλογή υπερ-παραμέτρων ερευνώντας δύο άξονες αυτού του αντικειμένου: διαθέσιμους αλγορίθμους βελτιστοποίησης και τρόπους ενσωμάτωσης μετα-γνώσης στη διαδικασία. Το σύστημά μας υποστηρίζει και πειραματίζεται με άπλειστες (πλεγματική αναζήτηση) και στατιστικές (bayesian) τεχνικές βελτιστοποίησης, τις οποίες αξιολογεί μέσω στατιστικών τεστ υπόθεσης.

Η υλοποίηση ενός λογισμικού ανάλυσης δεδομένων αποτελεί σημαντική ευκαιρία εκμετάλλευσης της θεωρίας της μετα-γνώσης για την επίτευξη ενός έμπειρου, εκπαιδευόμενου και επεκτάσιμου προγράμματος. Η δυνατότητα εκμετάλλευσης της εμπειρίας που δημιουργείται με την επιτυχημένη αντιμετώπιση προβλημάτων συνεισφέρει στη συνειδητή λήψη αποφάσεων, την εύκολη προσαρμογή σε νέα προβλήματα και την αποφυγή της άσκοπης επανάληψης διαδικασιών. Συγκεκριμένα, το σύστημά μας ενσωματώνει τη χρήση μετα-χαρακτηριστικών για τη πρόβλεψη υπερ-παραμέτρων, μια τεχνική που μας απαλλάσσει από την ανάγκη βελτιστοποίησής τους.

Αποτελεί, πλέον, κοινή παραδοχή ότι η επιτυχημένη μηχανική μάθηση προϋποθέτει τη χρήση, ή έστω τον πειραματισμό με ποικιλία τεχνικών. Φαίνεται πως η κοινότητα της μηχανικής μάθησης έχει αρχίσει να θέτει υπό αμφισβήτιση την αρχή της απλότητας του μοντέλου μάθησης, γνωστής ως "ξυράφι του Όκαμ" για να περάσει στη πλευρά του Επίκουρου, σύμφωνα με τον οποίο "ο συνδυασμός σωστών λύσεων σε ένα πρόβλημα, δε μπορεί παρά να λύνει το πρόβλημα τουλάχιστον εξίσου καλά". Η μεταφορά βέβαια της αρχής αυτής στο χώρο της μηχανικής μάθησης απαιτεί ιδιαίτερη προσοχή κατά την αξιολόγηση του μοντέλου, καθώς ενέχει ο κίνδυνος υπερ-προσαρμογής. Αυτή η διαπίστωση αποτέλεσε βασικό παράγοντα στον καθορισμό της λειτουργικότητας του συστήματός μας, το οποίο υποστηρίζει πληθώρα αλγορίθμων και τεχνικών προ-επεξεργασίας και ανάλυσης δεδομένων, θέτοντας την απαίτηση για τη χρήση συλλογών μοντέλων (ensembles). Δεδομένης της απαιτητικότητας που δημιουργεί η παρουσία πολλών, ενδεχομένως ποιοτικά αμφισβητήσιμων μοντέλων, ενσωματώσαμε την τεχνική του σχηματισμού συλλογών μοντέλων με προς τα εμπρός επιλογή (forward model selection ensemble) [5], μία προσέγγιση που έχει ξαναχρησιμοποιηθεί σε σχετικές εργασίες.

Τέλος, υπάρχει μία προσέγγιση της μηχανικής μάθησης, η οποία δεν μπορεί να βελτιστοποιηθεί, να προκύψει από μετα-γνώση ή την εφαρμογή κάποιου αλγορίθμου μάθησης, αλλά συνιστά απαραίτητο εργαλείο στα χέρια του αναλυτή δεδομένων. Πρόκειται για τη χρήση ευριστικών κανόνων. Θεωρούμε πως η παράλειψη ενσωμάτωσής τους θα στερούσε από το σύστημά μας πρακτική γνώση, απαραίτητη για τη λήψη σχεδιαστικών αποφάσεων. Έχουμε επομένως αναζητήσει και συλλέξει ευριστική γνώση από τη βιβλιογραφία, την οποία ενσωματώσαμε στο λογισμικό, παραμετροποιώντας σχεδιαστικές αποφάσεις που παίρνει ο αλγόριθμος.

1.4 Διάρθρωση Κειμένου

Η εργασία αποτελείται από 7 κεφάλαια, συμπεριλαμβανομένου και του παρόντος εισαγωγικού.

Στο Κεφάλαιο 2 θέτουμε το θεωρητικό υπόβαθρο στο οποίο βασίστηκε το σύστημα μας. Συγκεκριμένα ορίζουμε τη διαδικασία της μηχανική μάθησης και αναλύουμε βασικές τεχνικές της. Στη συνέχεια εισάγουμε τον αναγνώστη στο χώρο του AutoML, παραθέτοντας ιστορικά στοιχεία, γνωρίζοντας τη τρέχουσα κατάσταση και αναλύοντας τις δύο κυρίαρχες εκφάνσεις αυτής της επιστήμης: τη βελτιστοποίηση των υπερ-παραμέτρων μοντέλων μηχανικής μάθησης και τη μετα-μάθηση.

Στο Κεφάλαιο 3 αναλύουμε το σύστημά μας Αρχικά παραθέτουμε τα κίνητρα που οδήγησαν στη σχεδίαση του συστήματος και διαμόρφωσαν τα κύρια χαρακτηριστικά του κατά την εκκίνηση της εργασίας. Στη συνέχεια περιγράφουμε τη βασική αρχιτεκτονική του λογισμικού που σχεδιάσαμε αναλύοντας τα δύο βασικά υποσυστήματά του: το υποσύστημα εκπαίδευσης και το υποσύστημα πειράματος. Ακολουθεί αναλυτική περιγραφή των τεχνικών που επιστρατεύτηκαν για την επίτευξη της επιθυμητής λειτουργικότητας. Εμπνευσμένες από τη πρόσφατη βιβλιογραφία όφειλαν να προσαρμοστούν στο σύστημα και να εξεταστεί η συνεισφορά τους. Πιστεύουμε πως η ανάλυση αυτή θα βοηθήσει στην κατανόηση της αρχιτεκτονικής που έχουμε αναλύσει.

Το Κεφάλαιο 4 επιχειρεί να αξιολογήσει το σύστημα ως ολότητα, καθώς και τις μεμονωμένες τεχνικές που χρησιμοποιεί. Αρχικά περιγράφουμε τη διαδικασία συλλογής των σετ δεδομένων που χρειαστήκαμε για την ανάλυση, καθώς και τις παραμετροποιήσεις εργαλείων που χρησιμοποιήσαμε, ώστε να είναι εφικτή η αναπαραγωγή των πειραμάτων. Στη συνέχεια περιγράφουμε τη διαδικασία αναγνώρισης εξωκείμενων σετ δεδομένων, μίας λειτουργικής απαίτησης που στοχεύει στην ενημέρωση του χρήστη για την ετοιμότητα του συστήματος ως προς το

ζητούμενο πείραμα. Τέλος, αξιολογούμε το υποσύστημα που ασχολείται με τη ρύθμιση των μοντέλων, το υποσύστημα του ensemble και τέλος, το συνολικό σύστημα. Σε κάθε περίπτωση αναφέρουμε σχεδιαστικές επιλογές και προ-απαιτούμενα του πειράματος.

Στο Κεφάλαιο 5 αναφέρουμε δημοσιεύσεις στις οποίες βασιστήκαμε και περιγράφουμε τη λειτουργία συστημάτων παρόμοιων με το δικό μας.

Στο Κεφάλαιο 6 αναθεωρούμε τη λειτουργία του συστήματός μας εξάγοντας γενικότερα συμπεράσματα από την πειραματική αξιολόγηση και τοποθετούμε το σύστημα ως προς τη συνεισφορά του στη σύγχρονη βιβλιογραφία.

Στο Κεφάλαιο 7 αφορμώμενοι από περιορισμούς, προβλήματα και ιδέες που προέκυψαν στη διάρκεια της εργασίας μας, παραθέτουμε μελλοντικές επεκτάσεις-βελτιώσεις του συστήματος.

2

ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Μηχανική Μάθηση

Στην ενότητα αυτή θα τοποθετήσουμε εννοιολογικά τον όρο μηχανική μάθηση, θα ορίσουμε μία καθολική ορολογία και θα περιγράψουμε συνοπτικά τη ροή ενός πειράματος μηχανικής μάθησης.

2.1.1 Η έννοια της μηχανικής μάθησης

Η μηχανική μάθηση αναδύθηκε από τον επιστημονικό τομέα της Τεχνητής Νοημοσύνης, η οποία μελετά την ικανότητα υπολογιστικών συστημάτων να επιδείξουν ευφυΐα. Στο άρθρο *Υπολογιστική Μηχανική και Νοημοσύνη* [6] ο Allan Turing επιχειρεί να αντιμετωπίσει την εγγενή ασάφεια των όρων "μηχανή" και "σκέφτομαι" μέσω ενός συλλογισμού, που αποκαλεί *Το Παιχνίδι της Μίμησης*. Το "παιχνίδι" έγκειται στην προσπάθεια ανίχνευσης Τεχνητής Νοημοσύνης ως απάντηση στο ερώτημα: "Μπορεί μία μηχανή να πείσει για την ανθρώπινη ιδιότητά της;" Με παρόμοια συλλογιστική ο Tom M. Mitchel [7] κατέληξε στον ακόλουθο φορμαλιστικό ορισμό της μηχανικής μάθησης:

> *Λέμε πως ένα πρόγραμμα υπολογιστή μαθαίνει από μια εμπειρία E, αναφερόμενοι σε ένα σύνολο καθηκόντων T και ένα μέτρο απόδοσης P, αν η απόδοσή του στα καθήκοντα T, όπως μετράται από το P, βελτιώνεται καθώς αποκτά εμπειρία E.*

Πρακτικά το πρόγραμμα αντιλαμβάνεται την εμπειρία ως δεδομένα, τα οποία περιγράφουν ένα πρόβλημα και καλείται να εξάγει συμπεράσματα ώστε να προβλέψει μελλοντικές συμπεριφορές. Ανάλογα με τη μορφή του καθήκοντος η μηχανική μάθηση διακρίνεται σε:

- **Επιβλεπόμενη μάθηση (Supervised learning)** Το πρόγραμμα λαμβάνει πληροφορία τόσο για τα χαρακτηριστικά του προβλήματος όσο και για τη συμπεριφορά που καλείται να προβλέψει. Για παράδειγμα, αν θέλουμε να προβλέψουμε την τιμή των ακινήτων μιας περιοχής θα συλλέξουμε χαρακτηριστικά όπως η τοποθεσία, τα τετραγωνικά μέτρα και η τιμή κάποιων κατοικιών και θα χρησιμοποιήσουμε το πρόγραμμα για τη πρόβλεψη των τιμών άλλων κατοικιών με βάση την τοποθεσία και το μέγεθός τους.

- **Μη επιβλεπόμενη μάθηση (Unsupervised learning)** Σε αυτήν την περίπτωση το πρόγραμμα αναλαμβάνει να ανακαλύψει δομικά πρότυπα στα δεδομένα χωρίς να διαθέτει πληροφορία για τη προβλεπόμενη συμπεριφορά. Αν λοιπόν η Amazon στοχεύει να αναγνωρίσει τους τύπους των πελατών της, ώστε να μεγιστοποιήσει το κέρδος της προβάλλοντας σε κάθε τύπο προσαρμοζόμενες διαφημίσεις, θα χρειαστεί ένα πρόγραμμα, το οποίο θα τους ομαδοποιεί σε ομοιογενείς ομάδες με βάση χαρακτηριστικά όπως οι αγορές, η καταγωγή κτλ.

- **Ενισχυτική Μάθηση (Reinforcement learning)** Ούτε σε αυτή τη μορφή το πρόγραμμα διαθέτει πληροφορία για τη προβλεπόμενη συμπεριφορά, η προσέγγιση ωστόσο είναι διαφορετική: το πρόγραμμα δρα σε ένα δυναμικό περιβάλλον, με το οποίο αλληλεπιδρά μέσω ανταμοιβών στις προβλέψεις του. Όπως ακριβώς ένα παιδί χρειάζεται να ακουμπήσει μερικές φορές κάτι καυτό για να μάθει ότι δεν πρέπει να το ξανακάνει, έτσι και ένας πράκτορας λογισμικού χρειάζεται να δοκιμάσει διάφορες κινήσεις στο σκάκι για να μάθει να κερδίζει.

Ένας ακόμη διαχωρισμός των προβλημάτων μηχανικής μάθησης προκύπτει από το είδος της προβλεπόμενης συμπεριφοράς:

- **Προβλήματα παλινδρόμησης (Regression)** Πρόκειται για προβλήματα πρόβλεψης μιας συνεχούς τιμής, όπως η τιμή πώλησης ακινήτων.

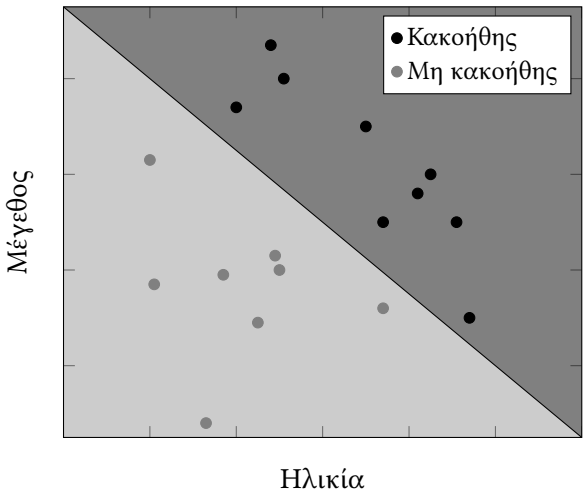
- **Προβλήματα ταξινόμησης (Classification)** Εδώ ενδιαφερόμαστε να αναγνωρίσουμε την κατηγορία, στην οποία ανήκει ένα δεδομένο. Για παράδειγμα ένα εργοστάσιο ενδιαφέρεται για τη πρόβλεψη ελαττωματικών εξαρτημάτων με βάση τα χαρακτηριστικά τους.

- **Προβλήματα ομαδοποίησης (Clustering)** Σε αυτή τη περίπτωση γίνεται αναγνώριση ομάδων με βάση την ομοιογένειά τους, όπως στο παράδειγμα που χρησιμοποιήσαμε για τη περιγραφή της Μη Επιβλεπόμενης Μάθησης.

Στη συνέχεια θα εστιάσουμε στην επιβλεπόμενη μάθηση σε προβλήματα ταξινόμησης, καθώς αποτελούν το πεδίο εφαρμογής της παρούσας διπλωματικής εργασίας.

Ορολογία

Με κίνητρο τη χρήση ενός κοινού λεξιλογίου θα ορίσουμε βασικές έννοιες που χρησιμοποιούνται συχνά στη βιβλιογραφία μέσω ενός παραδείγματος. Έστω το πρόβλημα πρόβλεψης της κακοήθειας ενός όγκου με βάση την ηλικία και το μέγεθός του. Τότε ορίζουμε ως:



Σχήμα 2.1: Ορολογία μηχανική μάθησης: Οι άξονες του διαγράμματος αντιστοιχούν στα χαρακτηριστικά του προβλήματος και τα σημεία στα παραδείγματα, για τα οποία η κλάση απεικονίζεται με το χρώμα. Η υπόθεση h αντιστοιχεί στη γραμμή, η οποία διαχωρίζει το πρόβλημα σε δύο υποχώρους.

- **Χαρακτηριστικά** x_n Τα στοιχεία που περιγράφουν το πρόβλημα, δηλαδή το μέγεθος και η ηλικία του όγκου.

- **Κλάση** y_n Πρόκειται για το στοιχείο που θέλουμε να προβλέψουμε, στην προκειμένη, τη κακοήθεια του όγκου.

- **Παραδείγματα** (x_n, y_n) Τα δεδομένα του προβλήματος δίνονται συνήθως σε μορφή πίνακα: κάθε γραμμή αποτελεί ένα παράδειγμα και οι στήλες περιέχουν τα χαρακτηριστικά και την κλάση.

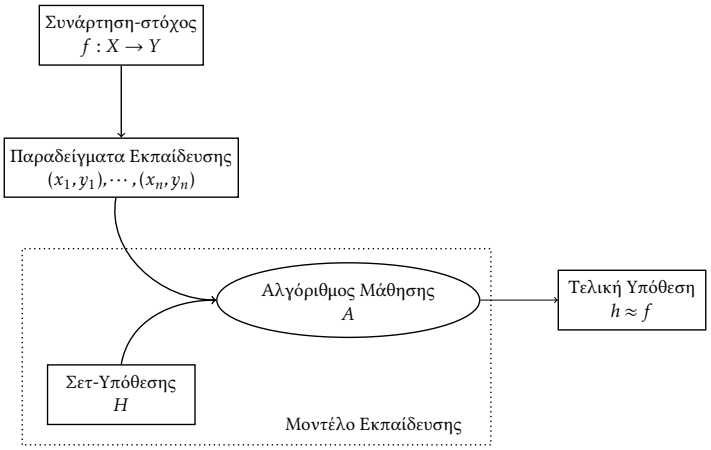
- **Συνάρτηση-στόχος** $f : X \to Y$ Είναι η άγνωστη συνάρτηση, που ορίζει πως προκύπτει η κλάση από τα χαρακτηριστικά του προβλήματος. Σκοπός της μηχανικής μάθησης είναι η προσέγγισή της, η οποία θα γίνει με τη βοήθεια των πεπερασμένων παραδειγμάτων που διαθέτουμε.

- **Υπόθεση** h Το αποτέλεσμα της εκπαίδευσης, δηλαδή η προσέγγιση της f, όπως φαίνεται στο Σχήμα 2.1. Στο παράδειγμά μας είναι μια νοητή γραμμή, η οποία χωρίζει το δισδιάστατο χώρο των χαρακτηριστικών σε δύο υποχώρους.

- **Μοντέλο Εκπαίδευσης** Προκειμένου να πραγματοποιήσουμε προβλέψεις σε άγνωστα δεδομένα, χρειαζόμαστε ένα μοντέλο, μία μαθηματική διαδικασία, η οποία έχει παραμετροποιηθεί πάνω στο συγκεκριμένο πρόβλημα και λαμβάνοντας τα χαρακτηριστικά ενός νέου δεδομένου μπορεί να δώσει τη κλάση του. Το μοντέλο αποτελείται από δύο συστατικά:

  - **Σετ υπόθεσης** $H = \{h\}$ Κάθε μοντέλο επιχειρεί να προσεγγίσει τη συνάρτηση-στόχο με διαφορετικό τρόπο. Το σετ υπόθεσης περιέχει όλες τις πιθανές υποθέσεις, που μπορούν να προκύψουν από ένα μοντέλο. Κάθε διαφορετική υπόθεση αντιστοιχεί σε διαφορετική ρύθμιση κάποιας παραμέτρου του μοντέλου και επιτελεί διαφορετική πρόβλεψη για τα δεδομένα. Είδη μοντέλων αποτελούν τα τεχνητά νευρωνικά δίκτυα (Artificial neural networks-ANN), οι μηχανές διανυσματικής στήριξης (Support vector machines-SVM) κτλ.

  - **Αλγόριθμος μάθησης** Ανάλογα με το μοντέλο που έχουμε επιλέξει, υπάρχει πληθώρα αλγορίθμων, οι οποίοι επιτελούν τη διαδικασία της μάθησης, προσπαθώντας να βελτιστοποιήσουν τις παραμέτρους της υπόθεσης. Για παράδειγμα οι νευρώνες (perceptrons) χρησιμοποιούν τον αλγόριθμο PLA, τα ANN τον αλγόριθμο οπισθοδιάδοσης (backpropagation) κτλ.

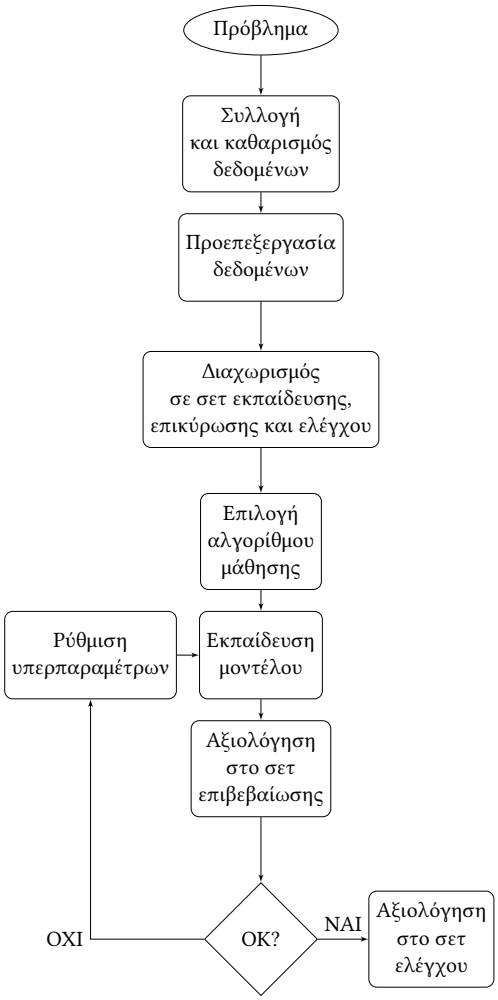


Σχήμα 2.2: Συστατικά Μηχανικής Μάθησης: Στόχος της παραγωγής ενός μοντέλου εκπαίδευσης είναι η προσέγγιση, μέσω της τελικής υπόθεσης, της συνάρτησης-στόχου , για την οποία το μοντέλο λαμβάνει πληροφορία μέσω των παραδειγμάτων. (Το σχήμα προέρχεται από τη σειρά διαλέξεων [1])

[1] `http://work.caltech.edu/telecourse.html`

2.1.2 Η διαδικασία της μηχανικής μάθησης

Η παρατήρηση της διαδικασίας εφαρμογής μηχανικής μάθησης σε ένα πραγματικό πρόβλημα εξηγεί γιατί οι ειδικοί σε αυτό τον τομέα χρειάζονται ένα πλούσιο, ετερογενές επιστημονικό υπόβαθρο, εμπειρία και εφευρετικότητα. Αποτελείται από διάφορα στάδια, που αλληλοεπηρεάζονται και με βάση την αξιολόγηση επαναλαμβάνονται κατά βούληση στη διάρκεια της μάθησης.



Σχήμα 2.3: Η διαδικασία της μηχανικής μάθησης, βασισμένη στην περιγραφή των Kotsiantis [8].

Συλλογή και καθαρισμός δεδομένων Η επίλυση ενός προβλήματος απαιτεί την ύπαρξη σχετικών δεδομένων, τα οποία συνήθως πρέπει να καθαριστούν από αγνοούμενες τιμές και θόρυβο.

Προεπεξεργασία Υπάρχουν διάφορες ενέργειες που μπορούν να εφαρμοστούν στα δεδομένα, με βάση τις ιδιαιτερότητες που παρουσιάζουν, ώστε να εξασφαλιστεί η καλή λειτουργία των αλγορίθμων μηχανικής μάθησης και εν γένει η καλή απόδοση του μοντέλου. Παραδείγματα αποτελούν η αφαίρεση αγνοούμενων τιμών και η κανονικοποίηση.

Διαχωρισμός σε σετ εκπαίδευσης, αξιολόγησης και ελέγχου Είναι αναγκαία η ύπαρξη δύο σετ δεδομένων, ανεξάρτητων από το σετ εκπαίδευσης, αλλά στατιστικά συσχετισμένων με αυτό (καθώς έχουν προκύψει από την ίδια άγνωστη συνάρτηση): το σετ αξιολόγησης, που θα χρησιμοποιηθεί για την επίτευξη της βέλτιστης παραμετροποίησης του μοντέλου και το σετ ελέγχου, το οποίο αποδεικνύει πόσο καλά δουλεύει το μοντέλο σε άγνωστα δεδομένα.

Εκπαίδευση μοντέλου Απαιτεί την επιλογή ενός αλγορίθμου μάθησης και την παραμετροποίησή του, ώστε να παραχθεί το τελικό μοντέλο.

Αξιολόγηση Αξιολογείται η ποιότητα του μοντέλου μέσω της εφαρμογής του στο σετ ελέγχου και του υπολογισμού μετρικών, που επιλέγονται με βάση της φύση του προβλήματος.

2.2 Τεχνικές Μηχανικής Μάθησης

Σε αυτήν την ενότητα αναλύουμε καθιερωμένες τεχνικές μηχανικής μάθησης διαχωρίζοντάς τες ως προς το στάδιο του πειράματος, στο οποίο εμφανίζονται.

2.2.1 Προεπεξεργασία

Σε αυτό το στάδιο ο αναλυτής οφείλει να αναγνωρίζει παθογένειες των δεδομένων, οι οποίες θα επηρεάσουν αρνητικά τη λειτουργία του αλγορίθμου μάθησης. Η βιβλιογραφία προσφέρει πληθώρα ετερογενών μεθοδολογιών, όπως η αναγνώριση ακατάλληλων τιμών που προέκυψαν κατά τη συλλογή των δεδομένων, η οπτικοποίηση του πληροφοριακού περιεχομένου των χαρακτηριστικών και ο μετασχηματισμός τους σε χρησιμότερες μορφές.

Ανάλυση κυρίαρχων συνιστωσών

Η τεχνική αυτή προέρχεται από τη γραμμική άλγεβρα και εφαρμόζεται με στόχο την εξαγωγή χρήσιμων χαρακτηριστικών σε προβλήματα μηχανικής μάθησης. Ανήκει στην ομάδα των μεθόδων φιλτραρίσματος, οι οποίες εφαρμόζονται στο σετ δεδομένων πριν την εκπαίδευση προκειμένου να "φιλτραριστούν" ανεπιθύμητα χαρακτηριστικά του. Το όνομά της προδίδει τη λειτουργία της: την εύρεση των κυρίαρχων συνιστωσών στα δεδομένα.

Τα δεδομένα σε ένα πρόβλημα ταξινόμησης αποτελούνται από τα χαρακτηριστικά και την κλάση πρόβλεψης. Γεωμετρικά, μπορούμε να αντιληφθούμε τα χαρακτηριστικά ως διανύσματα-βάσεις και τις τιμές κάθε παραδείγματος ως τις προβολές σε αυτή τη βάση. Σκοπός της ανάλυσης κυρίαρχων συνιστωσών είναι να βρει μια νέα βάση για τα δεδομένα, στην οποία αυτά θα περιγράφονται "καλύτερα". Αν λοιπόν τα αρχικά μας δεδομένα βρίσκονται στον πίνακα X, τότε αρκεί να βρούμε έναν πίνακα μετασχηματισμού P που θα μας μεταφέρει στη νέα βάση, δηλαδή:

$$Y = PX \tag{2.1}$$

Ο πίνακας P ορίζεται με τέτοιο τρόπο ώστε να αντιμετωπιστούν δύο παθογένειες των δεδομένων: ο θόρυβος και η περίσσεια πληροφορίας. Και τα δύο αυτά προβλήματα σχετίζονται άμεσα με την έννοια της ετεροσυσχέτισης: ο μεν θόρυβος είναι εξ ορισμού ασυσχέτιστος με όλα τα χαρακτηριστικά, η δε περίσσεια πληροφορίας ποσοτικοποιείται μέσω της ετεροσυσχέτισης μεταξύ των χαρακτηριστικών. Ο πίνακας ετεροσυσχέτισης των αρχικών δεδομένων ορίζεται ως:

$$S_X = \frac{1}{n-1} XX^T \tag{2.2}$$

Προκύπτει λοιπόν μία αναγκαιότητα για τα μετασχηματισμένα δεδομένα Y: ο πίνακας ετεροσυσχέτισής τους οφείλει να είναι διαγώνιος, δηλαδή κάθε χαρακτηριστικό να συσχετίζεται μόνο με τον εαυτό του. Ο πίνακας που επιθυμούμε να διαγωνοποιήσουμε είναι λοιπόν:

$$\begin{aligned} S_Y &= \frac{1}{n-1} YY^T \\ &= \frac{1}{n-1} (PX)(PX)^T \\ &= \frac{1}{n-1} PXX^T P^T \\ &= \frac{1}{n-1} P(XX^T)P^T \\ &= \frac{1}{n-1} PAP^T \end{aligned} \tag{2.3}$$

όπου $A = XX^T$.

Σύμφωνα με τη θεωρία της γραμμικής άλγεβρας, ένας πίνακας A διαγωνοποιείται με τη βοήθεια ενός πίνακα, κάθε στήλη του οποίου είναι ένα ιδιοδιάνυσμα του A, δηλαδή:

$$A = EDE^T \tag{2.4}$$

όπου ο D είναι ένας διαγώνιος πίνακας και E ένας πίνακας με στήλες τα ιδιοδιανύσματα του A.

Αν λοιπόν επιλέξουμε τον πίνακα P, έτσι ώστε κάθε γραμμή του να είναι ιδιοδιάνυσμα του A, τότε πετυχαίνουμε:

$$\begin{aligned} S_Y &= \frac{1}{n-1} PAP^T \\ &= \frac{1}{n-1} P(P^T DP)P^T \\ &= \frac{1}{n-1} (PP^T)D(PP^T) \\ &= \frac{1}{n-1} (PP^{-1})D(PP^{-1}) \\ &= \frac{1}{n-1} D \end{aligned} \tag{2.5}$$

δηλαδή ο πίνακας Y έχει διαγώνιο πίνακα ετεροσυσχέτισης και μπορούμε να πούμε πως οι κυρίαρχες συνιστώσες είναι οι γραμμές του P, δηλαδή τα ιδιοδιανύσματα του X και οι διαγώνιες τιμές του πίνακα S_Y είναι η διακύμανση κατά μήκος των κυρίαρχων συστατικών.

Συνήθως κατά την εξαγωγή χαρακτηριστικών προσπαθούμε να απλοποιήσουμε την περιγραφή των δεδομένων διατηρώντας όσο το δυνατόν περισσότερη πληροφορία. Έτσι, μετά την εφαρμογή της ανάλυσης κυρίαρχων συνιστωσών μπορούμε να επιλέξουμε να κρατήσουμε τα διανύσματα που μας δίνουν ένα ικανοποιητικό μέρος της διακύμανσης, συνήθως το 97% – 98%. Σε εφαρμογές που χαρακτηρίζονται από μεγάλη διαστασιμότητα στα δεδομένα, όπως μηχανικής όρασης, αυτή η μικρή απώλεια πληροφορίας μπορεί να μειώσει τα χαρακτηριστικά κατά εκατοντάδες.

Μετασχηματισμός Box-Cox

Στη διάρκεια ενός πειράματος μηχανικής μάθησης ανακύπτει συχνά η ανάγκη εξασφάλισης κανονικής κατανομής για ένα πληθυσμό. Παραδείγματος χάριν κατά την εκπαίδευση μοντέλων παλινδρόμησης η κανονικότητα των υπολειπόμενων τιμών (residuals) αποτελεί προϋπόθεση εγκυρότητας του μοντέλου. Την ανάγκη αυτή ικανοποιεί η οικογένεια των μετασχηματισμών ισχύος (power transformations), ένας εκ των οποίων είναι ο μετασχηματισμός Box-Cox. Εισήχθη από τους Box and Cox [9] και ορίζεται ως

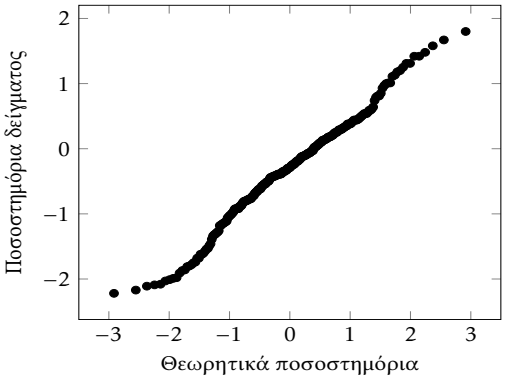
$$y' = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases} \tag{2.6}$$

Προκειμένου να επιλεχθεί το λ, το οποίο οδηγεί στη βέλτιστα κανονική κατανομή γίνεται χρήση της ιδιότητας των Q-Q (Quartile-Quartile) διαγραμμάτων να απεικονίζουν την κανονικότητα ενός πληθυσμού. Πρόκειται για διαγράμματα διασποράς των σημείων:

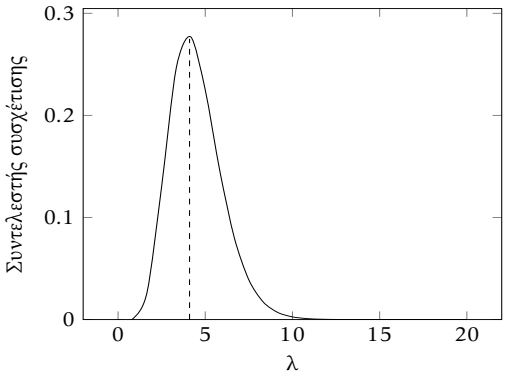
$$\left(\Phi^{-1}\left(\frac{i-0.5}{n}\right), x_i\right) \tag{2.7}$$

όπου Φ^{-1} η αντίστροφη συνάρτηση αθροιστικής κατανομής της κανονικής κατανομής και i το i-οστό ταξινομημένο σημείο του πληθυσμού. Η παρατήρηση γραμμικότητας σε ένα τέτοιο διάγραμμα αποτελεί απόδειξη κανονικότητας. Επομένως, ως λ του μετασχηματισμού Box-Cox

επιλέγεται αυτό που οδηγεί σε μέγιστο συντελεστή συσχέτισης μεταξύ των σημείων του Q-Q διαγράμματος.



Σχήμα 2.4: Ένα διάγραμμα διασποράς των πραγματικών τεταρτημορίων ενός πληθυσμού με τα τεταρτημόρια κανονικής κατανομής. Η διαπίστωση γραμμικότητας σε αυτό το διάγραμμα αποτελεί ένδειξη κανονικότητας της κατανομής.



Σχήμα 2.5: Η επιλογή του λ που βελτιστοποιεί το συντελεστή συσχέτισης του διαγράμματος Q-Q συνεπάγεται το μετασχηματισμό στη βέλτιστα κανονική κατανομή και χρησιμοποιείται στο μετασχηματισμό Box-Cox.

2.2.2 Εκπαίδευση

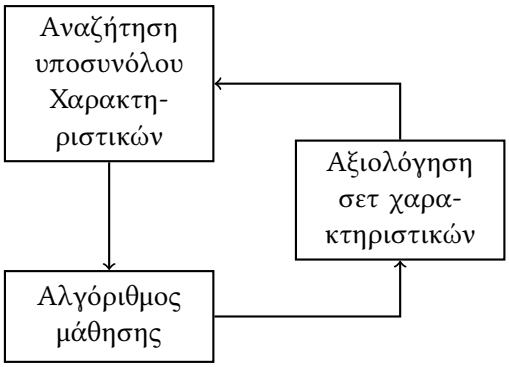
Πρωταρχική επιλογή κατά την εκπαίδευση ενός μοντέλου ταξινόμησης είναι αυτή του αλγορίθμου μάθησης. Ο αναλυτής δεδομένων έχει στη διάθεσή του ετερογενείς αλγορίθμους, όπως ο k-κοντινότερος γείτονας, οι μηχανές διανυσματικής στήριξης, ο απλοϊκός bayesian ταξινομητής (Naive Bayes), η Λογιστική Παλινδρόμηση (Logistic Regression), οι οποίοι αναλύονται στα Παραρτήματα Δ΄, Α΄, Β΄, Γ΄ αντίστοιχα. Μία τεχνική που ενσωματώνεται σε έναν αλγόριθμο μάθησης προκειμένου να αποφευχθεί το πρόβλημα της υπερ-προσαρμογής είναι αυτή της κανονικοποίησης (Παράρτημα Ε΄). Συνοπτικά θα αναφέρουμε ότι το πρόβλημα αυτό προκύπτει όταν το μοντέλο παραμετροποιείται τόσο καλά στη πρόβλεψη του σετ εκπαίδευσης ώστε να προβλέπει και θόρυβο εγγενή στα παραδείγματα, με αποτέλεσμα να έχει μειωμένη απόδοση σε άγνωστα δεδομένα.

Επιλογή χαρακτηριστικών

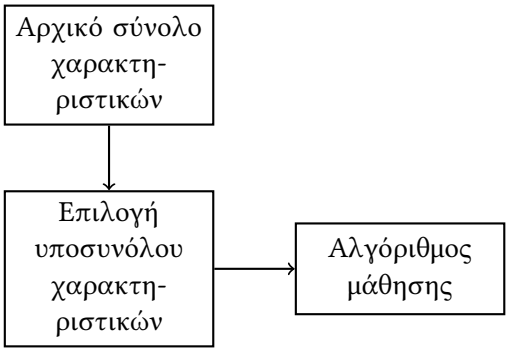
Σε αυτό το στάδιο θα αναλύσουμε τη δυνατότητα να εκπαιδεύσουμε το μοντέλο μας με διαφορετικά υποσύνολα των χαρακτηριστικών και να αξιολογήσουμε τις ακρίβειες των διαφορετικών μοντέλων, ώστε να συμπεράνουμε ποια χαρακτηριστικά συνεισφέρουν με βεβαιότητα στην πρόβλεψη. Οι μέθοδοι που το επιχειρούν αυτό λέγονται wrapper, επειδή εμπλέκουν τη διαδικασία της εκπαίδευσης και είναι πιο αποδοτικοί από τις μεθόδους φιλτραρίσματος που είδαμε κατά την προεπεξεργασία, καθώς λαμβάνουν αποφάσεις πολύ πιο συνειδητά.

Στο Σχήμα 2.6 βλέπουμε πώς οι μέθοδοι αυτοί διαλέγουν διαδοχικά ένα υποσύνολο των χαρακτηριστικών, εκπαιδεύουν ένα μοντέλο, το αξιολογούν και στη συνέχεια παραδίδουν το μοντέλο με την καλύτερη απόδοση και τα χαρακτηριστικά που επέλεξαν. Είναι σημαντικό πως κατά την επαναληπτική διαδικασία της επιλογής χαρακτηριστικών, η εκπαίδευση και η αξιολόγηση γίνεται σε 2 διαφορετικά υποσύνολα, το σετ εκπαίδευσης και το σετ επικύρωσης.

Η λογική με την οποία επιλέγονται τα υποσύνολα που θα δοκιμαστούν ακολουθεί συνήθως μία από δύο διαφορετικές συλλογιστικές:



Σχήμα 2.6: Μέθοδος wrapper για επιλογή χαρακτηριστικών: σε κάθε επανάληψη γίνεται επιλογή ενός υποσυνόλου χαρακτηριστικών, το μοντέλο εκπαιδεύεται και η απόδοσή του χρησιμοποιείται για την επιλογή του επόμενου υποσυνόλου.



Σχήμα 2.7: Μέθοδος φιλτραρίσματος για την επιλογή χαρακτηριστικών: γίνεται επιλογή ενός υποσυνόλου με βάση κάποιες ιδιότητες των χαρακτηριστικών, όπως η συσχέτισή τους στην Ανάλυση Κυρίαρχων Συνιστωσών.

- **προς τα εμπρός επιλογή** Ξεκινά με ένα άδειο σύνολο και προσθέτει διαδοχικά χαρακτηριστικά που μειώνουν το σφάλμα ταξινόμησης μέχρι καμία προσθήκη να μη το βελτιώνει.
- **προς τα πίσω επιλογή** Ξεκινά με όλα τα χαρακτηριστικά και αφαιρεί διαδοχικά χαρακτηριστικά που μειώνουν το σφάλμα ταξινόμησης μέχρι καμία αφαίρεση να μη το βελτιώνει.

Τα βασικά μειονεκτήματα αυτών των μεθόδων είναι πως είναι χρονοβόρα και μπορεί να οδηγήσουν σε υπερ-προσαρμογή.

Συνάθροιση μοντέλων

Στο σημείο αυτό θα γνωρίσουμε μια οικογένεια τεχνικών, που στοχεύουν στη βελτίωση της μηχανικής μάθησης, καθώς αποτελούν επιπρόσθετο κομμάτι της διαδικασίας και όχι πρωταρχικό της συστατικό. Η ιδέα που κρύβεται από πίσω τους, αν και αρχαία, ανατρέπει την παραδοσιακή προσέγγιση της μηχανικής μάθησης και προσδίδει περισσότερη σχεδιαστική ελευθερία στο στάδιο της εκπαίδευσης ενός μοντέλου.

Μία από τις βασικές αρχές της μηχανικής μάθησης αποτελεί το “ξυράφι του Όκαμ”: Δεδομένου ενός προβλήματος, μεταξύ ανταγωνιζομένων λύσεων επιλέγουμε αυτήν που απαιτεί τις λιγότερες υποθέσεις. Αυτή η αρχή ερμηνεύεται ως εξής: αν διαθέτω ετερογενείς και σωστές λύσεις σε ένα πρόβλημα, τότε θα προτιμήσω την απλούστερη. Στον τομέα της μηχανικής μάθησης αυτό αντιστοιχεί στην επιλογή της απλούστερης υπόθεσης, δηλαδή αυτής που προέκυψε από το μοντέλο με τις λιγότερες παραμέτρους και την απλούστερη προεπεξεργασία, μεταξύ υποθέσεων με παρόμοια ακρίβεια. Το συμπέρασμα φαντάζει λογικό: αν έχουμε βρει ένα απλό μοντέλο, που περιγράφει τη συνάρτηση-στόχο γιατί να διακινδυνέψουμε με ένα πιο απαιτητικό, χρονικά και υπολογιστικά, δυσνόητο και ευάλωτο σε υπερ-προσαρμογή μοντέλο;

Στον αντίποδα αυτής της επιχειρηματολογίας βρισκόταν ο Επίκουρος: “αν έχω βρει πολλές ερμηνείες για κάποιο φαινόμενο, γιατί να μην τις λάβω όλες υπόψιν μου, ώστε να έχω μια πιο ολοκληρωμένη αντίληψη;” Με την παραδοχή πως δεν υπάρχουν αυθεντίες, αλλά ειδικοί, ο συνδυασμός των απόψεων ειδικών σε διαφορετικούς τομείς ενός προβλήματος, μπορεί να οδηγήσει σε μια πιο εξισορροπημένη και βέλτιστη λύση. Αντιστοίχως, το βελτιστοποιημένο μοντέλο με το οποίο έχουμε επιλύσει ένα πρόβλημα ταξινόμησης δε συνιστά εγγυημένα καλή λύση, καθώς

υπόκειται σε περιορισμούς, που δεν εμφανίζονται σε άλλα μοντέλα.

Υπάρχουν διάφορες τεχνικές με τις οποίες μπορούμε να συνδυάσουμε τη γνώση διαφορετικών μοντέλων με σκοπό η ακρίβεια της συνισταμένης γνώσης να είναι καλύτερη από το βέλτιστο μοντέλο που επιτεύχθηκε με τη χρήση ενός αλγορίθμου.

- **Bootstrap- aggregating** Η τεχνική αυτή, που συνήθως αποκαλείται *bagging*, συνιστά τον απλούστερο τρόπο συνάθροισης: Από τα παραδείγματα εκπαίδευσης, λαμβάνουμε Κ υποσύνολα με n στοιχεία το καθένα, δειγματοληπτώντας με αντικατάσταση. Για κάθε διαφορετικό υποσύνολο εκπαιδεύουμε ένα μοντέλο με έναν αλγόριθμο ομαδοποίησης, για παράδειγμα με ένα δέντρο ταξινόμησης. Όταν θέλουμε να προβλέψουμε την κλάση ενός νέου στοιχείου, χρησιμοποιούμε τα Κ μοντέλα και τελικά προβλέπουμε την κλάση που επέλεξε η πλειοψηφία. Ο λόγος για τον οποίο επιλέξαμε τα δέντρα ως παράδειγμα δεν είναι τυχαίος: η τεχνική αυτή χρησιμοποιείται κυρίως για μοντέλα που επηρεάζονται από την τυχαιότητα των δεδομένων εκπαίδευσης, γεγονός που εξασφαλίζεται με την τυχαία δειγματοληψία.

- **Boosting** Η προηγούμενη τεχνική θα μπορούσε να χαρακτηριστεί ως naive, καθώς υποθέτει πως τα διαφορετικά μοντέλα παρουσιάζουν μη αλληλεπικαλυπτόμενες αδυναμίες και άρα απλά συνδυάζοντάς τα θα καλύψουμε ικανοποιητικά όλους τους τύπους εισόδου. Η υπόθεση αυτή δεν είναι ωστόσο ρεαλιστική, καθώς τα μοντέλα τείνουν να δυσκολεύονται σε παρόμοιες περιπτώσεις. Η τεχνική *boosting* ακολουθεί επίσης τη λογική εκπαίδευσης Κ μοντέλων, τα οποία ωστόσο δεν είναι ανεξάρτητα: κάθε μοντέλο δίνει περισσότερη βαρύτητα στην ταξινόμηση παραδειγμάτων, τα οποία τα προηγούμενα μοντέλα απέτυχαν να ταξινομήσουν σωστά. Επίσης, η ψήφος των μοντέλων δεν είναι ισοδύναμη, αλλά ενισχύεται για τα ακριβέστερα μοντέλα.

- **Stacked generalization** Η ψηφοφορία των διαφορετικών μοντέλων γίνεται δυσκολότερη, όταν έχουν εκπαιδευθεί με τη χρήση διαφορετικών αλγορίθμων. Διαφοροποιήσεις ως προς τις προϋποθέσεις, τη λειτουργία και το είδος της εξόδου των αλγορίθμων οδηγούν σε μη-συγκρίσιμα και διαφορετικής ποιότητας αποτελέσματα. Η τεχνική αυτή, που αποκαλείται εν συντομία *stacking*, δίνει λύση σε αυτό το πρόβλημα εισάγοντας την έννοια του μετα-μοντέλου εκπαίδευσης. Σε πρώτο στάδιο τα μοντέλα εκπαιδεύονται και παράγεται η πρόβλεψη για κάθε παράδειγμα. Το δεύτερο στάδιο, που αποτελεί το μετα-μοντέλο, παίρνει ως είσοδο την πρόβλεψη κάθε μοντέλου και την πραγματική κλάση για κάθε παράδειγμα και εκπαιδεύει ένα νέο μοντέλο μηχανικής μάθησης, που θα αποφασίσει πώς θα συνδυάσει τα επιμέρους ώστε να επιτύχει την καλύτερη ακρίβεια.

2.2.3 Αξιολόγηση

Είδαμε πως τόσο κατά τη ρύθμιση του μοντέλου στη διάρκεια της εκπαίδευσης, όσο και για την τελική αξιολόγηση του μοντέλου για τη διαπίστωση της ικανότητάς του να γενικεύει χρειαζόμαστε δύο ανεξάρτητα σετ: ένα στο οποίο θα γίνεται η εκπαίδευση και ένα στο οποίο θα γίνεται η αξιολόγηση του μοντέλου. Φυσικά τα δεδομένα μας δίνονται ενιαία και ο τρόπος με τον οποίο θα διαχωριστούν αποτελεί σχεδιαστική επιλογή. Ο ειδικός οφείλει να συμβιβαστεί μεταξύ δύο σκοπών: τη χρήση όσο το δυνατόν μεγαλύτερου σετ εκπαίδευσης, για να παραχθεί ένα πιο "σοφό" μοντέλο, αλλά και σετ ελέγχου, ώστε η γενίκευση να είναι εγγυημένη, λαμβάνοντας υπόψιν το πεπερασμένο του σετ δεδομένων και του διαθέσιμου χρόνου.

Μέθοδοι

Hold out Πρόκειται για την απλούστερη τεχνική: αναθέτουμε ένα μέρος των δεδομένων για εκπαίδευση και τα διατίθενται για αξιολόγηση. Οι συνήθεις αναλογίες είναι 80% – 20% και 75% – 25% για εκπαίδευση και αξιολόγηση αντίστοιχα. Αν και γρήγορη, η τεχνική αυτή δεν προτιμάται, καθώς δεν εγγυάται την αξιοπιστία του αποτελέσματος και συρρικνώνει πολύ το σετ εκπαίδευσης.

Leave one out Με την τεχνική αυτή μεγιστοποιούμε το μέγεθος των δύο σετ εις βάρος του χρόνου της μάθησης: κάθε φορά εκπαιδεύουμε το μοντέλο με όλα τα δεδομένα εκτός από ένα και το αξιολογούμε σε αυτό, επαναλαμβάνοντας τη διαδικασία τόσες φορές όσα είναι τα δεδομένα μας.

k-fold cross-validation Αυτή είναι η συνηθέστερη τεχνική, καθώς εξασφαλίζει μικρούς χρόνους μάθησης και αξιόπιστο αποτέλεσμα τόσο από πλευρά εκπαίδευσης όσο και από πλευρά αξιολόγησης. Τα δεδομένα χωρίζονται σε k υποσύνολα, το μοντέλο εκπαιδεύτει με τα $k-1$ και αξιολογείται με το εναπομείναν. Η διαδικασία επαναλαμβάνεται k φορές, ώστε όλα τα υποσύνολα να χρησιμοποιηθούν μία φορά για αξιολόγηση. Τελικά η απόδοση του μοντέλου υπολογίζεται ως ο μέσος όρων των επιδόσεων στα k υποσύνολα. Συνήθως το k επιλέγεται ως 10, οπότε μιλάμε για 10-fold cross-validation.

.632 bootstrap Η τεχνική αυτή προσπαθεί να πετύχει το στόχο του cross-validation, με διαφορετική όμως λογική: αντί να τεμαχίζουμε το σετ εκπαίδευσης, δημιουργούμε k αντίγραφά του με τυχαία δειγματοληψία με αντικατάσταση, το καθένα από τα οποία αποτελείται από N στοιχεία. Αν λοιπόν το αρχικό σετ δεδομένων ήταν:

$$S = \begin{bmatrix} y_1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ y_N & x_{N1} & \dots & x_{Np} \end{bmatrix} \tag{2.8}$$

τότε κάθε δείγμα ορίζεται ως:

$$S_b = \begin{bmatrix} y_1^{*b} & x_{11}^{*b} & \dots & x_{1p}^{*b} \\ \vdots & \vdots & \ddots & \vdots \\ y_N^{*b} & x_{N1}^{*b} & \dots & x_{Np}^{*b} \end{bmatrix} \tag{2.9}$$

Στη συνέχεια μπορούμε να εκπαιδεύσουμε ένα διαφορετικό μοντέλο $\bar{f}^{*b}(x)$ με κάθε δείγμα και να υπολογίσουμε το σφάλμα ως εξής:

$$\bar{E}_{boot} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \bar{f}^{*b}(x)) \tag{2.10}$$

Ο παραπάνω δείκτης είναι πολωμένος, καθώς υπάρχει η πιθανότητα δεδομένα που έχουν χρησιμοποιηθεί για την εκπαίδευση ενός μοντέλου να χρησιμοποιηθούν και για την αξιολόγησή του. Η πιθανότητα αυτή, όπως προκύπτει λόγω της τυχαίας δειγματοληψίας με αντικατάσταση είναι πολύ μεγάλη:

$$P[(y_i, x_i) \in S_b] = 1 - (1 - \frac{1}{N})^N \approx 1 - e^{-1} \approx 0.632 \tag{2.11}$$

Την παθογένεια αυτή επιχειρεί να λύσει η τεχνική του leave one out bootstrap cross-validation, όπου τα στοιχεία δε χρησιμοποιούνται για την αξιολόγηση ενός μοντέλου, στην εκπαίδευση του οποίου έχουν συμμετάσχει:

$$\bar{E}_{boot(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \bar{f}^{*b}(x)) \tag{2.12}$$

Η παραπάνω μετρική συνεχίζει ωστόσο να είναι πολωμένη λόγω της συχνής επαναχρησιμοποίησης δεδομένων: κάθε δείγμα περιέχει κατά μέσο όρο $0.632 \cdot N$ διαφορετικά στοιχεία, χαρακτηριστικό που θυμίζει 2-fold cross-validation.

Έτσι, προτάθηκε η συμβιβαστική λύση του .632 bootstrap():

$$\bar{E}^{(0.632)} = 0.368 \cdot \overline{err} + 0.632 \cdot \bar{E}_{boot(1)} \tag{2.13}$$

όπου $\overline{err}$ είναι το σφάλμα που υπολογίζεται για τα σημεία που έχουν συμμετάσχει στην εκπαίδευση.

Αυτή η μετρική προσπαθεί να σταθμίσει τη συνεισφορά δύο αντίθετα πολωμένων όρων. Ο πρώτος αποτελεί τον απλό εκτιμητή και λειτουργεί σωστά για σημεία που δεν απέχουν καθόλου από το σετ εκπαίδευσης. Ο δεύτερος έχει υποθέσει τη μεγαλύτερη απόσταση των νέων δεδομένων από το σετ εκπαίδευσης (η πιθανότητα ένα στοιχείο να μη συμμετέχει σε ένα δείγμα είναι 0.368).

Ο σχεδιασμός της αξιολόγησης αποτελεί βασικό στάδιο για την επιτυχή εκπαίδευση ενός μοντέλου. Η κακή σχεδίαση οδηγεί είτε στην εκπαίδευση ενός υποβέλτιστου μοντέλου είτε στη λανθασμένη εκτίμηση της απόδοσής του. Όπως επισήμαναν οι Ambroise and McLachlan [10] αμφισβητήσιμη είναι η ποιότητα των ερευνών που χρησιμοποιούν τα παραδείγματα ελέγχου κατά τη ρύθμιση του μοντέλου, καθώς παράγουν θετικά πολωμένες εκτιμήσεις. Οι ίδιοι συνεχίζουν υπενθυμίζοντας πως η τεχνική leave one out οδηγεί σε μοντέλα υψηλής διακύμανσης, δηλαδή αμφισβητήσιμης απόδοσης σε νέα προβλήματα, γεγονός που αποδίδεται στην ομοιότητα των σετ εκπαίδευσης (κάθε ζεύγος διαφέρει ως προς ένα παράδειγμα). Ως καλύτερη συμβιβαστική λύση προτείνεται η τεχνική k-fold cross-validation για k=10. Καθώς μικρότερες τιμές του k οδηγούν σε υποβέλτιστα μοντέλα λόγω περιορισμένων παραδειγμάτων στο σετ εκπαίδευσης και μεγαλύτερες ενσωματώνουν τα προβλήματα του leave one out, η τιμή αυτή προσφέρει μια ευριστικά καλή λύση.

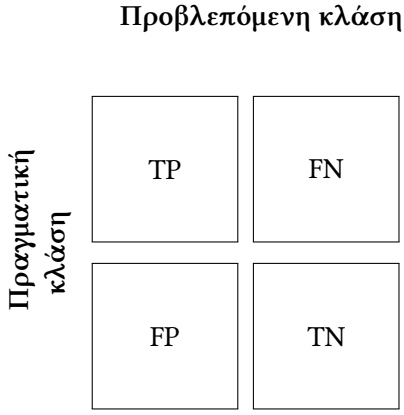
Μετρικές

Σκοπός μίας μετρικής είναι η ποσοτικοποίηση της ποιότητας ενός μοντέλου. Καθώς λοιπόν η ποιότητα ορίζεται μέσω της επίτευξης ενός προσδοκώμενου στόχου, η επιλογή της μετρικής που θα χρησιμοποιηθεί για δεδομένο πρόβλημα θα εξαρτηθεί από τη φύση του.

Σε ένα πρόβλημα ταξινόμησης οι συνήθεις μετρικές προκύπτουν από τεχνικές σύνοψης της λειτουργίας του ταξινομητή, όπως ο Πίνακας Σύγχυσης και η καμπύλη ROC (Receiver Operating characteristic).

Πίνακας Σύγχυσης Ο Πίνακας Σύγχυσης εισήχθη από τους Provost and Kohavi [11], ως μια ειδική περίπτωση πίνακα ενδεχομένων (contigency table), δηλαδή ενός πίνακα που περιγράφει

την κατανομή πιθανοτήτων τυχαίων μεταβλητών. Αποτελεί τρόπο παρουσίασης της λειτουργίας ενός δυαδικού ταξινομητή συνοψίζοντας τις σωστές και λανθασμένες προβλέψεις του ως προς τις διαφορετικές κλάσεις του προβλήματος. Οι κυρίαρχες μετρικές σε τέτοια προβλήματα μπορούν να οριστούν μέσω αυτού.



Πίνακας 2.1: Ο Πίνακας Σύγχυσης συνοψίζει τη λειτουργία του δυαδικού ταξινομητή, ο οποίος προβλέψει μεταξύ θετικών και αρνητικών παραδειγμάτων. TP: σωστά θετικά, TN: σωστά αρνητικά, FN: λάθος αρνητικά, FP: λάθος θετικά

Οι συνήθεις μετρικές που προκύπτουν από τον Πίνακα 2.1 είναι οι εξής:

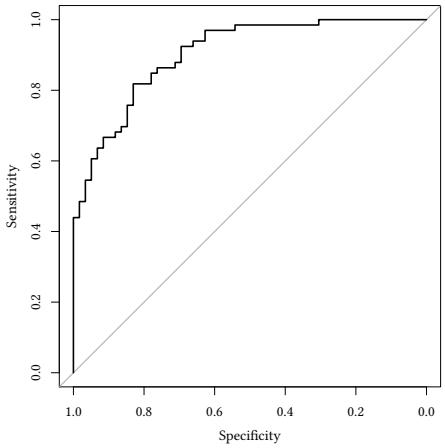
$$\text{1. Ακρίβεια (Accuracy)} = \frac{TP + TN}{TP + TN + FP + FN}$$
$$\text{2. Ανάκληση (Recall)} = \frac{TP}{TP + FN}$$
$$\text{3. Ακρίβεια (Precision)} = \frac{TP}{TP + FN}$$
$$\text{4. F-μετρική (F-measure)} = \frac{2 \cdot (\text{Precision} + \text{Recall})}{\text{Precision} + \text{Recall}}$$

Καμπύλη ROC Οι καμπύλες αυτές απεικονίζουν την απόδοση ενός δυαδικού ταξινομητή για μεταβλητό κατώφλι διάκρισης (αν θεωρήσουμε ότι ο ταξινομητής είναι πιθανοτικός, τότε το κατώφλι διάκρισης ορίζει την τιμή της πιθανότητας πέρα από την οποία προβλέπεται διαφορετική κλάση). Το διάγραμμα του Σχήματος 2.8 σχηματίζεται τοποθετώντας ένα σημείο για κάθε τιμή κατωφλίου με τετμημένη την Ειδικότητα (Specificity) και τεταγμένη την Ευαισθησία (Sensitivity), οι οποίες ορίζονται ως

$$\text{1. Ειδικότητα (Specificity)} = \frac{TP}{TP + TN}$$
$$\text{2. Ευαισθησία (Sensitivity)} = \frac{FP}{FP + TN}$$

Στατιστικά τεστ υπόθεσης

Τα πειράματα ταξινόμησης απαιτούν τη στατιστική ανάλυση πληθυσμών για την εξαγωγή συμπερασμάτων. Αν θεωρήσουμε ένα σύνολο από σετ δεδομένων δυαδικής ταξινόμησης, μερικά



Σχήμα 2.8: Καμπύλη ROC: η γραμμή αντιστοιχεί σε τυχαίο ταξινομητή, αποτελεί δηλαδή το σημείο αναφοράς του διαγράμματος. Όσο πιο πάνω και αριστερά βρίσκεται η καμπύλη του τόσο καλύτερη η απόδοσή του, με το σημείο $(0,1)$ *να αντιστοιχεί σε τέλειο ταξινομητή.*

ερωτήματα που μπορούν να προκύψουν είναι: υπάρχει κάποια στατιστική συσχέτιση μεταξύ της κλάσης και κάποιου χαρακτηριστικού για ένα συγκεκριμένο σετ δεδομένων; Ποιος αλγόριθμος παρουσιάζει γενικά καλύτερη συμπεριφορά;

Τα στατιστικά τεστ εφαρμόζονται σε πίνακες ενδεχομένων και έχουν ως στόχο της απόρριψη ή μη της μηδενικής υπόθεσης, η οποία αντιστοιχεί σε ανεξαρτησία των δεδομένων και τυχαιότητα των διαφορών που παρουσιάζονται μεταξύ διαφορετικών πληθυσμών. Οι πίνακες αυτοι είναι 1-way, 2-way ή 3-way, ενώ για την εμπλοκή περισσότερων πληθυσμών οι ερευνητές καταφεύγουν σε γενικευμένα γραμμικά μοντέλα [12].

Μερικές έννοιες που σχετίζονται με τα στατιστικά τεστ είναι:

- **Εναλλακτική υπόθεση (alternative hypothesis)** Προτείνεται από τον ερευνητή και χαρακτηρίζει τη στατιστική σχέση μεταξύ των δύο πληθυσμών υπό σύγκριση. Συγκρίνεται ως η εναλλακτική μιας ιδανικής μηδενικής υπόθεσης, η οποία αποκλείει οποιαδήποτε σχέση μεταξύ των δύο δειγμάτων. Στόχος του πειράματος είναι η απόρριψη της μηδενικής υπόθεσης, ενώ σε περίπτωση αποτυχίας το συμπέρασμα είναι η αδυναμία απόρριψης της μηδενικής υπόθεσης και όχι η επιβεβαίωση της εναλλακτικής.

- **Στατιστική σημασία (statistical significance)** Το πείραμα έχει στατιστική σημασία, όταν η σχέση μεταξύ των δειγμάτων είναι απίθανο να προκύψει από τη μηδενική υπόθεση με βάση ένα κατώφλι πιθανότητας α.

- **Στατιστική του τεστ** Πρόκειται για μία στατιστική μετρική ενός δείγματος, σκοπός της οποίας είναι να ποσοτικοποιήσει χαρακτηριστικά του δείγματος, τα οποία θα βοηθήσουν στο διαχωρισμό της μηδενικής από την εναλλακτική υπόθεση.

- **Διάστημα εμπιστοσύνης(confidence interval)** Προκύπτει από το κατώφλι πιθανότητας ως $1-\alpha$ και ερμηνεύεται ως εξής: Αν έχουμε 95% διάστημα εμπιστοσύνης τότε είμαστε κατά ίση πιθανότητα σίγουροι ότι η μέση τιμή του πληθυσμού (και όχι των δειγμάτων) θα κινείται σε συγκεκριμένα πλαίσια (που προκύπτουν από την κατανομή του test statistic).

- **p-τιμή (p-value)** Είναι η τιμή που οδηγεί στο συμπέρασμα του στατιστικού πειράματος. Αποτελεί απόδειξη κατά της μηδενική υπόθεσης και όσο χαμηλότερη είναι η τιμή του τόσο ισχυρότερη η απόρριψή της. Για δεδομένο κατώφλι πιθανότητας αρκεί να είναι μι-

κρότερο από αυτό.

- **Σφάλματα τύπου I/II (type I/II errors)** Η απόρριψη μιας έγκυρης μηδενικής υπόθεσης χαρακτηρίζεται ως σφάλμα τύπου I, ενώ η αδυναμία απόρριψης μιας άκυρης σφάλμα τύπου II.

- **Ισχύς (Power)** Πρόκειται για τη πιθανότητα το τεστ να απορρίψει μια λανθασμένη μηδενική υπόθεση.

Για έναν πλήρη κατάλογο των συμβατικών τεστ που χρησιμοποιούνται στη βιβλιογραφία μπορούμε να ανατρέξουμε στο Παράρτημα ΣΤ΄.

2.3 Αυτοματοποιημένη Μηχανική Μάθηση

Έχοντας αναλύσει μερικά από τους αλγορίθμους που προσφέρει η επιστήμη της μηχανικής μάθησης, μάλλον μας έχει δοθεί η εντύπωση πως απευθύνεται σε μία ολιγομελη κοινωνία ειδικών, που με χρόνια εμπειρίας, εξειδικευμένη έρευνα και λίγη δόση τύχης, καταφέρνει να δημιουργήσει μοντέλα μηχανικής μάθησης που λύνουν πραγματικά προβλήματα. Είναι αλήθεια πως σε κάθε στάδιο οι σχεδιαστικές επιλογές που επηρεάζουν την απόδοση του μοντέλου προσθέτουν στο πρόβλημα αρκετούς βαθμούς ελευθερίας και καθιστούν τη βέλτιστη επιλογή χρονοβόρα και αμφισβητήσιμη. Τα οφέλη ωστόσο που προσφέρει ένα αποδοτικό μοντέλο είναι τόσο άμεσα και το πεδίο εφαρμογών της μηχανικής μάθησης τόσο ευρύ, που η ιδέα της αυτοματοποίησης της διαδικασίας βελτιστοποίησης έχει κινητοποιήσει μια μεγάλη μερίδα ειδικών. Είναι χαρακτηριστικό το γεγονός ότι στη συνολική διαδικασίας της μηχανικής μάθησης, που ουσιαστικά αφορά την εύρεση του μοντέλου πρόβλεψης, το 75% αφορά την προετοιμασία των δεδομένων και το 15% την ανάλυση των αποτελεσμάτων. [2]

Ο όρος AutoML είναι σχετικά πρόσφατος στη βιβλιογραφία και αφορά κάθε τεχνική αυτοματοποίησης οποιουδήποτε σταδίου της διαδικασίας της μηχανικής μάθησης. Ο Matthew Mayo [3] αποτυπώνει την ουσία του AutoML ως εξέλιξη της σχέσης ανθρώπου-μηχανής:

> *Ο προγραμματισμός στοχεύει στην αυτοματοποίηση, η μηχανική μάθηση στην αυτοματοποίηση της αυτοματοποίησης και η αυτοματοποιημένη μηχανική μάθηση στην αυτοματοποίηση του να αυτοματοποιείς την αυτοματοποίηση. Δεδομένου λοιπόν ότι ο προγραμματισμός αναλαμβάνει τη διεκπεραίωση τετριμμένων καθηκόντων και η μηχανική μάθηση επιτρέπει στους υπολογιστές να εκπαιδευτούν στην καλύτερη επίλυση των καθηκόντων, το AutoML καταφτάνει για να επιτρέψει στους υπολογιστές να αυτοματοποιήσουν το αποτέλεσμα της εκπαίδευσης.*

2.3.1 Ιστορική Αναδρομή

Ο τομέας της αυτοματοποίησης της μηχανικής μάθησης βρίσκεται σε πειραματικό στάδιο, όχι όμως και σε εμβρυικό. Τα σύγχρονα, εντυπωσιακά εργαλεία που έχουν στη διάθεσή τους σήμερα οι ειδικοί, προέκυψαν από την εικοσαετή εκκόλαψη της ιδέας της αυτοματοποίησης των βασικών σταδίων της μηχανικής μάθησης. Το 1995 η εταιρία Unica εισήγαγε στην αγορά το Pattern Recognition Workbench, ένα πακέτο λογισμικού που ενσωμάτωσε την αυτομαποίηση της ρύθμισης μοντέλων με νευρωνικά δίκτυα. Το λογισμικό Model 1 αποτέλεσε απόγονο του παραπάνω

[2] https://indico.lal.in2p3.fr/event/2914/session/1/contribution/4/material/slides/0.pdf
[3] http://www.kdnuggets.com/2017/01/current-state-automated-machine-learning.html

προϊόντος, καθώς το επέκτεινε και σε άλλες οικογένειες αλγορίθμων. Τα τέλη της δεκαετίας του 90 έχουν να επιδείξουν ακόμη δύο προσπάθειες: το Marketswitch, και το KXEN, εργαλεία που απευθύνονταν κυρίως στην αγορά του marketing, παρέχοντας διεπαφές για αυτοματοποίηση των προβλεπτικών μοντέλων. Πιο πρόσφατα παραδείγματα αποτελούν οι κολοσσοί στην αγορά των πωλητών λογισμικού για στατιστικές αναλύσεις: η SAS και η IBM SPSS, που με τα προϊόντα τους SAS Rapid Modeler και IBM SPSS Modeler αντίστοιχα προσπάθησαν από το 2010 να αυτοματοποιήσουν την προ επεξεργασία των δεδομένων, παραχωρώντας ταυτόχρονα στο χρήστη λειτουργικότερες διεπαφές [4].

Αν και βραχύβια, η ιστορία του AutoML μπορεί να μας διδάξει κάτι: η αρχική θεώρηση της αυτοματοποίησης της μηχανικής μάθησης ως λύτρωση από τη χρονοβόρα και πνευματικά απαιτητική επίτευξη ενός αποτελεσματικού μοντέλου ήταν λανθασμένη. Τα εργαλεία που αντιμετώπισαν την εκπαίδευση ως ένα μαύρο κουτί, ώστε ο χρήστης να πετυχαίνει εντυπωσιακά αποτελέσματα αγνοώντας τις βασικές αρχές και λειτουργίες των αλγορίθμων, απέτυχαν κατά την εφαρμογή τους σε πραγματικά προβλήματα. Πλέον αντιλαμβανόμαστε αυτόν τον τομέα ως ένα εργαλείο στα χέρια του ειδικού, που επιταχύνει, διευκολύνει και επεκτείνει τη μηχανική μάθηση, όπως ένα εργαλείο ρομποτικής ιατρικής στα χέρια ενός χειρούργου.

Στη συνέχεια θα αναλύσουμε κυρίαρχες τεχνικές σε δύο βασικούς τομείς της πρόσφατης βιβλιογραφίας του AutoML: της βελτιστοποίησης των υπερ-παραμέτρων αλγορίθμων μηχανικής μάθησης και της μετα-μάθησης.

2.3.2 Βελτιστοποίηση Υπερ-παραμέτρων

Ένα βασικό στάδιο κατά την εκπαίδευση αλγορίθμων μηχανικής μάθησης είναι αυτό της επιλογής των υπερ-παραμέτρων του μοντέλου.

Μαθηματικά το πρόβλημα μπορεί να διατυπωθεί ως εξής: σκοπός ενός πειράματος μηχανικής μάθησης είναι η εκπαίδευση ενός μοντέλου M, το οποίο ελαχιστοποιεί μία προκαθορισμένη συνάρτηση κόστους L σε ένα σετ δεδομένων X. Το μοντέλο κατασκευάζεται από έναν αλγόριθμο μάθησης , ο οποίος παραμετροποιείται από ένα σύνολο παραμέτρων λ.

Καταλήγουμε λοιπόν στον μαθηματικό ορισμό της εύρεσης του συνόλου των υπερ-παραμέτρων λ^*, που ορίζουν το βέλτιστο μοντέλο M^*

$$\lambda^* = \arg\min \left\{ L(X^{(te)}; M(X^{(tr)}; \lambda)) \right\} = \arg\min_{\lambda} \left\{ L(\lambda; M, X^{(tr)}, L) \right\} \tag{2.14}$$

όπου $X^{(tr)}$ το σετ δεδομένων και $X^{(te)}$ το σετ ελέγχου.

Αν συμπεριλάβουμε στη διατύπωση του προβλήματος και την επιλογή του βέλτιστου αλγορίθμου μηχανικής μάθησης, τότε η εξίσωση 2.14 μπορεί να αναδιατυπωθεί ώστε να περιλαμβάνει όλα τα M μοντέλα, καθένα εκ τω οποίων έχει διαφορετικές υπερ-παραμέτρους λ.

$$\lambda^* = \arg\min_{M^j \in M, \lambda \in \Lambda} \left\{ L(X^{(te)}; M^j(X^{(tr)}; \lambda)) \right\} \tag{2.15}$$

όπου $M = M^1, \cdots, M^k$ είναι ο χώρος των πιθανών αλγορίθμων και $\Lambda = \Lambda^1 \cup \cdots \cup \Lambda^k \cup \lambda_r$ ο χώρος των υπερ-παραμέτρων όλων των αλγορίθμων, όπου λ_r μία βοηθητική υπερ-παράμετρος

[4] `https://www.datarobot.com/blog/automated-machine-learning-short-history/`

για την εναλλαγή μεταξύ αλγορίθμων.

Το πρόβλημα αυτό αναφέρεται ως Πρόβλημα Συνδυασμένης Επιλογής αλγορίθμου και Βελτιστοποίησης Υπερ-παραμέτρων (Combined Algorithm Selection and Hyperparameter optimization - CASH) [13]

Μερικά χαρακτηριστικά της παραπάνω συνάρτησης είναι τα εξής:

- **είναι μια συνάρτηση μαύρου κουτιού, δηλαδή περιγράφεται μόνο μέσω εισόδων-εξόδων**
- **δεν έχουμε γνώση για τις παραγώγους της, το οποίο είναι άμεσο επακόλουθο της προηγούμενης πρότασης**
- **είναι μη-κυρτή**
- **δεν εξαρτάται εξίσου από όλες τις παραμέτρους**
- **ο υπολογισμός της για δεδομένο λ είναι υπολογιστικά και χρονικά απαιτητικός**

Η θεωρία της βελτιστοποίησης συναρτήσεων έχει προσφέρει ποικίλλες επιλογές στην επίλυση του υπό μελέτη προβλήματος. Εξελικτικοί αλγόριθμοι [14], κατάβαση κλήσης (gradient decent) [15], αλγόριθμοι βασισμένοι σε ευριστικές [16, 17]. Τα χαρακτηριστικά ωστόσο που αναφέρουμε προσδίδουν στη βελτιστοποίηση της Εξίσωσης 2.14 ιδιαιτερότητες, που συγκεκριμενοποιούν τον κατάλληλο αλγόριθμο βελτιστοποίησης. Η Bayesian βελτιστοποίηση έχει αναδειχθεί σε κυρίαρχο αλγοριθμο, καθώς, όπως θα δούμε στην Ενότητα 2.3.4, χρησιμοποιείται στην πλειοψηφία των AutoML εργαλείων.

Bayesian Βελτιστοποίηση

Βελτιστοποίηση blackbox συναρτήσεων Η αναζήτηση των βέλτιστων παραμέτρων γίνεται με άξονα τη μεγιστοποίηση της γενικευμένης απόδοσης: ενός δείκτη που δηλώνει πόσο καλά λειτουργεί το μοντέλο μας σε άγνωστα δεδομένα. Όπως είδαμε στην Ενότητα 2.1.1 σκοπός ενός μοντέλου, και επομένως παράγοντας αξιολόγησής του, είναι η προσέγγιση της πραγματικής συνάρτησης, που περιγράφει πώς προκύπτει η υπό μελέτη κλάση από τα χαρακτηριστικά. Η υπόθεση κυρτότητας, που θα εξασφάλιζε την εύρεση ολικού μεγίστου με τοπικό αλγόριθμο αναζήτησης είναι άτοπη, καθώς η επίλυση ενός πραγματικού προβλήματος συνήθως συνεπάγεται πολυπλοκότητα της συνάρτησης που το περιγράφει. Το μόνο που γνωρίζουμε για αυτήν είναι τα δεδομένα που έχουμε, δηλαδή κάποιες εισόδους και εξόδους της, εξού και ο χαρακτηρισμός της ως μαύρο κουτί. Τα διαθέσιμα δεδομένα είναι μάλιστα περιορισμένα, καθώς η απόκτησή τους μπορεί να απαιτεί χρόνο, κόπο και χρήματα (δοκιμές φαρμάκων, οικονομικές επενδύσεις). Η βελτιστοποίηση μιας συνάρτησης f(x) με παραμέτρους από ένα σύνολο A, συμβολίζεται ως:

$$\max_{x \in A} f(x) \tag{2.16}$$

Ο όρος Bayesian βελτιστοποίηση εισήχθη από τους Mockus and Mockus [18] σε μια σειρά μελετών τους για ολική βελτιστοποίηση συναρτήσεων. Βασικά χαρακτηριστικά αυτής της διαδικασίας είναι πως είναι ακολουθιακή, ενσωματώνει κάποια εκ των προτέρων πεποίθηση που έχουμε για την υπό βελτιστοποίηση συνάρτηση, χρησιμοποιεί το θεώρημα Bayes και καθοδηγεί την αναζήτηση του μεγίστου με βάση ένα συνδυασμό μεταξύ εξερεύνησης και εκμετάλλευσης.

Θεώρημα Bayes Σύμφωνα με αυτό το θεώρημα, δεδομένου ενός μοντέλου πρόβλεψης M και ενός συνόλου παρατηρήσεων E, η εκ των υστέρων πιθανότητα, δηλαδή η πιθανότητα δεδομένων των παρατηρήσεων E να προκύψει το μοντέλο M είναι ανάλογη της πιθανότητας των παρατηρήσεων δεδομένου του μοντέλου επί την εκ των προτέρων πιθανότητα του μοντέλου, με μαθηματικούς όρους:

$$P(M \mid E) \propto P(E \mid M) \cdot P(M) \tag{2.17}$$

Η παραπάνω πιθανότητα μας δίνει ένα μέσο για να απαντήσουμε στο πραγματικό ερώτημα: "Δεδομένων των διαθέσιμων παρατηρήσεων ποιο μοντέλο τα περιγράφει καλύτερα και επομένως είναι πιθανότερο να προσεγγίζει την πραγματική συνάρτηση;"

Η Γκαουσιανή διαδικασία (Gaussian process) ως εκ των προτέρων πιθανότητα Η εκ των προτέρων πιθανότητα αντικατοπτρίζει την πεποίθηση που έχουμε για την άγνωστη συνάρτηση, όπως για παράδειγμα την ομαλότητα.

Η γκαουσιανή διαδικασία ορίζεται ως μια επέκταση μιας γκαουσιανής κατανομής πολλών μεταβλητών σε μια στοχαστική διαδικασία απείρων διαστάσεων, όπου κάθε πεπερασμένος συνδυασμός διαστάσεων δίνει μια γκαουσιανή κατανομή. Όπως ακριβώς η γκαουσιανή κατανομή δίνει την κατανομή κάποιων μεταβλητών και χαρακτηρίζεται πλήρως από τη μέση τιμή και τη διακύμανσή της, έτσι και η γκαουσιανή διαδικασία αποτελεί μια κατανομή συναρτήσεων που χαρακτηρίζεται από κάποια μέση συνάρτηση και μια συνάρτηση διακύμανσης. Για να κατανοήσουμε καλύτερα τη διαδικασία αυτή, μπορούμε να σκεφτούμε πως όπως μία συνάρτηση επιστρέφει έναν αριθμό $f(x)$ για μία τιμή x, αυτή επιστρέφει τη μέση τιμή και διακύμανση μιας κανονικής κατανομής, που δίνει όλες τις πιθανές τιμές της $f(x)$ για το συγκεκριμένο x.

Συνάρτηση απόκτησης Σε κάθε επανάληψη της Bayesian βελτιστοποίησης επιλέγονται τα επόμενα σημεία αναζήτησης της βέλτιστης τιμής με βάση τα σημεία που έχουν ήδη αξιολογηθεί και την τρέχουσα αντίληψη της πιθανότητας. Για το σκοπό αυτό χρησιμοποιούνται οι συναρτήσεις απόκτησης, οι οποίες μεγιστοποιούνται για σημεία που ενδεχομένως να μεγιστοποιούν και την άγνωστη συνάρτηση. Υπάρχουν διάφορες τεχνικές, ωστόσο η βασική αρχή επιλογής αποτελεί ένα συμβιβασμό μεταξύ "εξερεύνησης" (exploration) και "εκμετάλλευσης" (exploitation): ο εξερευνητικός χαρακτήρας της αναζήτησης εκδηλώνεται με την εισχώρηση σε ανεξερεύνητες περιοχές προκειμένου να αποφευχθεί ο κίνδυνος εμμονής σε κάποιο τοπικό μέγιστο, ενώ ο εκμεταλλευτικός με την προτίμηση περιοχών με εξακριβωμένα καλή απόδοση.

- **Πιθανότητα βελτίωσης (probability of improvement)** Μία από τις πρώτες τεχνικές, που εισήχθη από τους Kushner and Mockus [19], ήταν αυτή της επιλογής του σημείου x^+ με τη μεγαλύτερη πιθανότητα βελτίωσης:

$$PI(x) = P(f(x) \geq f(x^+)) = \Phi\left(\frac{\mu(x) - f(x^+)}{\sigma(x)}\right) \tag{2.18}$$

  όπου Φ είναι η συνάρτηση κανονικής αθροιστικής κατανομής.

  Το βασικό της μειονέκτημα τότε ήταν ότι δεν λάμβανε καθόλου υπόψιν της το στοιχείο της εξερεύνησης, το οποίο εισήχθη μέσω της παραμέτρου ξ ως εξής:

$$PI(x) = P(f(x) \geq f(x^+)) = \Phi\left(\frac{\mu(x) - f(x^+) - \xi}{\sigma(x)}\right) \tag{2.19}$$

  Η εισαγωγή της παραμέτρου ξ λύνει μεν το πρόβλημα δίνοντας ένα αντιληπτό κατώφλι στη βελτίωση που επιτυγχάνεται, αποτελεί δε μία σημαντική σχεδιαστική επιλογή.

Καθώς για ακατάλληλα μικρή τιμή υπάρχει ο κίνδυνος εγκλωβισμού σε τοπικό μέγιστο, ενώ για μεγάλη τιμή η διαδικασία επιβραδύνεται, η αναζήτηση μπορεί να οδηγηθεί σε σφάλμα.

Μία πιο ικανοποιητική προσέγγιση θα ήταν κατά την επιλογή του επόμενου σημείου να μη ληφθεί υπόψιν μόνο η πιθανότητα βελτίωσης, αλλά και το μέγεθος της βελτίωσης. Πιο συγκεκριμένα, θα θέλαμε να ελαχιστοποιήσουμε την προσδοκώμενη απόκλιση από το πραγματικό μέγιστο:

$$x_{t+1} = argmin\{E(|f_{t+1}(x) - f(x^*)|D_{1:t})\} \tag{2.20}$$

Ο Mockus όρισε τη συνάρτηση βελτίωσης ως εξής:

$$I(x) = \max\Big(f_{t+1}(x) - f(x^*)\Big) \tag{2.21}$$

δηλαδή η βελτίωση είναι θετική όταν η πρόβλεψη είναι μεγαλύτερη από την μέχρι τώρα καλύτερη τιμή, ειδάλλως μηδέν. Το νέο σημείο βρίσκεται μεγιστοποιώντας την προσδοκώμενη βελτίωση:

$$x = argmax\{E(\max(f_{t+1}(x) - f(x^*) \mid D_t)\} \tag{2.22}$$

Η πιθανότητα διαπίστωσης βελτίωσης I σε μία κανονική κατανομή, που χαρακτηρίζεται από μέση τιμή $\mu(x)$ και διακύμανση $\sigma(x)^2$ υπολογίζεται ως εξής:

$$\frac{1}{\sqrt{2\pi}\sigma(x)} e^{-\frac{(\mu(x)-f(x^*-I))^2}{\sigma(x)^2}} \tag{2.23}$$

και η προσδοκώμενη βελτίωση είναι το ολοκλήρωμα της παραπάνω συνάρτησης ως προς I.

- **Χρήση άνω ορίων εμπιστοσύνης (Upper confidence bounds)** Στην τεχνική αυτή είναι ξεκάθαρη η προσπάθεια συμβιβασμού μεταξύ εξερεύνησης και εκμετάλλευσης. Όπως έχουμε αναφέρει, η μέση τιμή της γκαουσιανής διαδικασίας αποτελεί την τρέχουσα εντύπωση που έχουμε για την άγνωστη συνάρτηση. Μπορούμε να επιλέξουμε πόσο εξερευνητικοί θα είμαστε ορίζοντας πόσες τυπικές αποκλίσεις πέρα από τη μέση τιμή της διαδικασίας θα κινηθούμε.

Η προσέγγιση της συνάρτησης L που εμφανίζεται στην Εξίσωση 2.14 μπορεί να γίνεται κατεξοχήν με τη χρήση μοντέλου, οπότε ονομάζεται Ακολουθιακή Βελτιστοποίηση βασισμένη σε Μοντέλο (Sequential Model-Based Optimization - SMBO). Παραδείγματα χωρίς χρήση μοντέλου είναι οι αλγόριθμοι Random Online Adaptive Racing (ROAR) [20], Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [13] και η Ακολουθιακή Βελτιστοποίηση χωρίς Μοντέλο (Sequential Model-free Optimization - SMBO) [21].

Ο αλγόριθμος που υλοποιεί τη Bayesian βελτιστοποίηση παραθέτεται στο Παράρτημα Z′

Ακολουθιακή βελτιστοποίηση βασισμένη σε μοντέλο

Οι αλγόριθμοι βελτιστοποίησης που χρησιμοποιούν μοντέλα εκπαίδευσης εκκολάφτηκαν από τη διαπίστωση πως η συνάρτηση, $f : \Theta \to Y$ που καθορίζει πως επιδρούν οι υπερ-παράμετροι στην απόδοση του μοντέλου είναι περίπλοκη και οφείλει να προσεγγιστεί από κάποιο μοντέλο μηχανικής μάθησης. Έτσι, τα δεδομένα εκπαίδευσης έχουν τη μορφή $(\theta_1, y_1), ..., (\theta_n, y_n)$, όπου

$\theta_i = (\theta_{i,1}..., \theta_{i,d})$ οι d παράμετροι και y_i η απόδοση που επετεύχθη με αυτές. Ένας τυπικός αλγόριθμος αυτής της κατηγορίας περιέχει έναν εσωτερικό βρόγχο, όπου επιλέγουμε το σημείο x^*, που βελτιστοποιεί την πρόβλεψη, ως το επόμενο σημείο αξιολόγησης της f. Οι αλγόριθμοι διαφοροποιούνται ως προς το κριτήριο που χρησιμοποιούν για να βελτιστοποιήσουν την πρόβλεψη και ποιο μοντέλο εκπαίδευσης χρησιμοποιούν.

Η υπεροχή των SMBO έγκειται στη δυνατότητα παρεμβολής (interpolation) μεταξύ παρατηρούμενων σετ υπερ-παραμέτρων και παρέκτασης (extrapolation) σε άγνωστες περιοχές του χώρου υπερ-παραμέτρων. Επίσης, ποσοτικοποιούν τη σημασία κάθε υπερ-παραμέτρου και των μεταξύ τους αλληλεξαρτήσεων.

Για μία αλγοριθμική περιγραφή των SMBO μπορούμε να ανατρέξουμε στο παράρτημα Z′. Στη συνέχεια θα αναφέρουμε δύο παραδείγματα SMBO αλγορίθμων.

Ακολουθιακή ρύθμιση αλγορίθμου βασισμένη σε μοντέλο (SMAC) Παρουσιάστηκε το 2011 από τους Hutter, Hoos, and Leyton-Brown [20] και αποτελεί χαρακτηριστικό παράδειγμα αυτής της οικογένειας. Χρησιμοποιεί βασισμένα σε παραδείγματα (instance-based - IBL) μοντέλα, όπως γκαουσιανά μοντέλα ακτινικής βάσης, αλλά και δέντρα, όπως τον αλγόριθμο Random Forest. Το κριτήριο επιλογής που χρησιμοποιεί είναι αυτό της προσδοκώμενης βελτίωσης, που αναλύσαμε στο προηγούμενο κεφάλαιο και εδώ ορίζεται ως:

$$EI_{y^*}(\theta) = \int_{-\infty}^{y^*} (y^* - y) p(y \mid \theta) dy \tag{2.24}$$

Δέντρο εκτιμητών Parzen (TPE) Παρουσιάστηκε από τους Bergstra et al. [22]. Η εμφάνιση του όρου Parzen στην ονομασία αυτή της τεχνικής αποδίδεται στη χρήση του Parzen εκτιμητή, μίας τεχνικής παρεμβολής για προσέγγιση της πυκνότητας πιθανότητας γύρω από ένα σημείο με χρήση γκαουσιανών συναρτήσεων βάσης. Ενώ ο αλγόριθμος SMAC υπολογίζει απευθείας την ποσότητα $p(y \mid \theta)$ κατά την εύρεση της προσδοκώμενης βελτίωσης, εδώ προσεγγίζεται με τη βοήθεια των $p(\theta \mid y)$ και $p(y)$. Συγκεκριμένα, μοντελοποιούμε την πιθανότητα $p(\theta \mid y)$ ως μία εκ δύο εκτιμήσεων πυκνότητας, ανάλογα με το αν η απόδοση y έχει ξεπεράσει κάποιο κατώφλι y^*:

$$p(\theta \mid y) = \begin{cases} l(\theta) & \text{εάν } y < y^* \\ g(\theta) & \text{εάν } y \geq y^* \end{cases}$$

Το σημείο y^* επιλέγεται με τη χρήση μιας παραμέτρου γ, ώστε να αντιστοιχεί στο γ-μόριο των απωλειών του TPE αλγορίθμου μέχρι την παρούσα στιγμή. Η συνάρτηση $l(\theta)$ είναι μια κατανομή που προήλθε από όλες τις προηγούμενες υπερ-παραμέτρους θ που οδήγησαν σε σφάλμα μικρότερο από y^* και η $g(\theta)$ από τις υπόλοιπες. Έτσι, μπορούμε να ερμηνεύσουμε την πρώτη ως μία εκτίμηση της κατανομής των υπερ-παραμέτρων που έχουν καλή απόδοση και τη δεύτερη αυτών που οδηγούν σε χαμηλή.

Οι κατανομές $l(\theta)$ και $g(\theta)$ παρουσιάζουν μία ιεραρχική δομή, καθώς αντιπροσωπεύουν τις υπερ-παραμέτρους και τη συσχέτιση μεταξύ τους. Όσο αφορά τις υπερ-παραμέτρους με συνεχείς τιμές, μπορούμε να φανταστούμε πως έχουμε υπολογίσει τον Parzen εκτιμητή για κάθε μια από αυτές. Για να υπολογίσουμε την πιθανότητα ενός παραδείγματος υπερ-παραμέτρων θ, ξεκινούμε από την κορυφή του δέντρου και κατευθυνόμαστε προς τα φύλλα ακολουθώντας τις υπερ-παραμέτρους που έχουμε. Η πιθανότητα σε κάθε κόμβο αντιστοιχεί στον Parzen εκτιμητή και τους συνδυάζουμε ακολουθώντας την αντίθετη διαδρομή προς τη ρίζα.

Τελικά, το κριτήριο που μεγιστοποιείται είναι:

$$EI_{y^*}(\theta) \propto (\gamma + \frac{g(\theta)}{l(\theta)} \cdot (1 - \gamma))^{-1}$$

δοκιμάζοντας διάφορους υποψήφιους συνδυασμούς των υπερ-παραμέτρων και επιλέγοντας αυτόν με τη μικρότερη τιμή $g(\theta)/l(\theta)$.

2.3.3 Μετα-μάθηση

Κατά την προσπάθεια αυτοματοποίησης, αλλά και γενικότερα βελτίωσης της διαδικασίας της μηχανικής μάθησης, συναντάμε τους ακόλουθους περιορισμούς των συμβατικών μοντέλων μάθησης (base learners):

- Τα πρότυπα, τα οποία αναγνωρίζονται στα δεδομένα, ενσωματώνονται στο μοντέλο, με αποτέλεσμα η επανεφαρμογή του να μη δημιουργεί νέα γνώση [23].
- Δεν υπάρχει προφανής τρόπος εξαγωγής και επαναχρησιμοποίησης της γνώσης που παράχθηκε σε νέα προβλήματα.

Κλειδί για την επίλυση αυτών των προβλημάτων αποτέλεσε η εισαγωγή της έννοιας της μετα-γνώσης. Πρόκειται για γνώση σχετική με την ίδια τη διαδικασία της μάθησης, την οποία προσπαθεί να βελτιώσει ο τομέας της μετα-μάθησης.

Η μετα-μάθηση στοχεύει στην ικανότητα ενός συστήματος να μαθαίνει από παρελθοντικά προβλήματα και να προσαρμόζεται με βάση την εμπειρία του. Δημιουργεί συστήματα ικανά να λάβουν εμπεριστατωμένες αποφάσεις σχετικά με την αυτοματοποίηση προβλημάτων μηχανικής μάθησης και να προσαρμοστούν σε νέα εμπόδια, όπως ένας αναλυτής δεδομένων επιστρατεύει την εμπειρία του κατά την αντιμετώπιση ενός νέου προβλήματος.

Η μετα-γνώση λαμβάνει τη μορφή μετα-χαρακτηριστικών, τα οποία εξάγονται από το εκάστοτε σετ δεδομένων και προσπαθούν να αποτυπώσουν τη φύση του προβλήματος μάθησης. Σύμφωνα με τη βιβλιογραφία [23] τα μετα-χαρακτηριστικά ανήκουν στις ακόλουθες κατηγορίες:

- **απλά, στατιστικά και της θεωρίας πληροφορίας (information-theoretic)** Πρόκειται για μετα-χαρακτηριστικά που περιγράφουν εξολοκλήρου το σετ δεδομένων, όπως το πλήθος των παραδειγμάτων, η συσχέτιση μεταξύ των χαρακτηριστικών, η εντροπία της κλάσης κτλ.
- **βασισμένα σε μοντέλο (model-based)** Σε αυτά γίνεται εκμετάλλευση των χαρακτηριστικών κάποιας υπόθεσης h, για παράδειγμα εκπαιδεύεται ένα δέντρο απόφασης και συλλέγονται οι υπερ-παράμετροι του.
- **ορόσημα (landmarks)** Η απόδοση ετερογενών αλγορίθμων μάθησης αποτελεί πληροφορία για τη φύση ενός σετ δεδομένων.

Πεδίο εφαρμογής της μετα-μάθησης μπορεί να αποτελέσει οποιοδήποτε στάδιο της διαδικασίας μηχανικής μάθησης, όπως η προεπεξεργασία, η επιλογή αλγορίθμου και η ρύθμιση ενός μοντέλου. Σχετικές προσπάθειες στον τομέα της αυτοματοποιημένης μηχανικής μάθησης αποτελούν οι Feurer, Springenberg, and Hutter [24], οι οποίοι χρησιμοποιούν μετα-μάθηση για να αρχικοποιήσουν την αναζήτηση υπερ-παραμέτρων και οι Soares, Brazdil, and Kuba [25], οι οποίοι εισάγουν μία μέθοδο επιλογής του πλάτους ενός γκαουσιανού πυρήνα για ένα μοντέλο SVM παλινδρόμησης.

2.3.4 Σύγχρονα εργαλεία

Αν αναλογιστεί κανείς το εύρος των εφαρμογών μηχανικής μάθησης, θα κατανοήσει την ύπαρξη πληθώρας εργαλείων που επιχειρούν να την αυτοματοποιήσουν. Βιβλιοθήκες σε διάφορες γλώσσες παρέχουν διεπαφές προς τεχνικές αυτοματοποίησης, διαδικτυακά περιβάλλοντα αναλαμβάνουν τη διαιτησία ολόκληρης της διαδικασίας της μηχανικής μάθησης προσφέροντας δυνατότητες αυτόματης βελτιστοποίησης της [5] και λογισμικά εξειδικευμένα στην ανάλυση δεδομένων ενσωματώνουν διεπαφές προς υλοποιημένους αλγορίθμους βελτιστοποίησης. Η εμπορική σημασία της αυτόματης επίτευξης μοντέλων πρόβλεψης έχει οδηγήσει στην κυκλοφορία πολλών εμπορικών εργαλείων, αλλά και οι κοινότητες ελεύθερου λογισμικού έχουν κινητοποιηθεί μπροστά σε αυτήν την πολύπλευρη ανάγκη. Στη συνέχεια θα δούμε μερικά χαρακτηριστικά εργαλεία.

HPOlib [6] Πρόκειται για μία βιβλιοθήκη βελτιστοποίησης υπερ-παραμέτρων, που παρέχει μία κοινή διεπαφή προς τρία σύγχρονα, αναγνωρισμένα πακέτα:

- **SMAC** Ο αλγόριθμος αυτός, γραμμένος σε python, έχει περιγραφεί στην Ενότητα 2.3.2.
- **Spearmint** Γραμμένο σε python, χρησιμοποιείται για την εφαρμογή bayesian βελτιστοποίησης.
- **Hyperopt** Γραμμένο σε python, αναλαμβάνει τη βελτιστοποίηση σε απαιτητικούς χώρους αναζήτησης με τη χρήση τυχαία αναζήτησης και του αλγορίθμου TPE.

auto-sklearn [7] Μία εργαλειοθήκη αυτοματοποιημένης μηχανικής μάθησης,η οποία με βάση την python βιβλιοθήκη scikit-learn [8] και χρήση Bayesian βελτιστοποίησης, μετα-μάθησης και ensembles αναλαμβάνει την παραγωγή μοντέλων μηχανικής μάθησης.

Auto-WEKA [9] Το Waikato Environment for Knowledge Analysis (Weka) είναι ένα λογισμικό σχετικό με τους τομείς της ανάλυσης δεδομένων και μοντέλων πρόβλεψης. Υλοποιεί πληθώρα αλγορίθμων μηχανικής μάθησης, γραμμένων σε Java, και παρέχει γραφικές διεπαφές και εργαλεία οπτικοποίησης για διευκόλυνση των χρηστών. Το λογισμικό αυτό έχει ενσωματώσει την αυτοματοποίηση της μηχανικής μάθησης στο Auto-WEKA, που εισήχθη το 2013 [26] και αναλαμβάνει την επίλυση του προβλήματος CASH (εξίσωση 2.15) . Συνεχίζοντας την παράδοση του εργαλείου αυτού στην απλότητα χρήσης, το Auto-WEKA αντιμετωπίζεται ως ένας απλός αλγόριθμος μάθησης, που αναλαμβάνει την επιλογή των χαρακτηριστικών, του μοντέλου και τη βελτιστοποίηση των υπερ-παραμέτρων, ανάμεσα σε όλες τις τεχνικές και αλγορίθμους που προσφέρει το WEKA. Για να επιλύσει αυτό το πολυδιάστατο πρόβλημα έχει βασιστεί σε SMBO αλγορίθμους (SMAC, TPE).

R & caret [10] Το πακέτο caret είναι γραμμένο σε R, μία γλώσσα προγραμματισμού και ένα περιβάλλον λογισμικού εξειδικευμένο στη στατιστική. Η R είναι το κατεξοχήν εργαλείο για

[5] `https://azure.microsoft.com/en-us/`
[6] `https://github.com/automl/HPOlib`
[7] `https://github.com/automl/auto-sklearn`
[8] `http://scikit-learn.org/stable/`
[9] `http://www.cs.ubc.ca/labs/beta/Projects/autoweka/`
[10] `http://caret.r-forge.r-project.org/`

συγγραφή κώδικα σε εφαρμογές στατιστικής, ανάλυσης δεδομένων και μηχανικής μάθησης και διαθέτει πακέτα που επιτελούν πληθώρα αλγορίθμων και τεχνικών, καθώς και εργαλείων οπτικοποίησης. Οι κοινότητες ελεύθερου λογισμικού έχουν εξοπλίσει την R με πακέτα που επιχειρούν τη βελτιστοποίηση της μηχανικής μάθησης, όπως αυτόματης προεπεξεργασίας δεδομένων και ρύθμισης μοντέλου, οι διεπαφές των οποίων είναι αναμενόμενα ανομοιόμορφες. Το πακέτο classification and regression training, caret, αποτελεί προσπάθεια τυποποίησης της διαδικασίας της εκπαίδευσης παρέχοντας ομοιόμορφες διεπαφές και εργαλεία για διαχωρισμό των δεδομένων, προ επεξεργασία, επιλογή χαρακτηριστικών, ρύθμιση των υπερ-παραμέτρων του μοντέλου, εκτίμηση της σημασίας των χαρακτηριστικών και άλλων λειτουργιών χρήσιμων στην προσπάθεια αυτοματοποίησης της εκπαίδευσης ενός μοντέλου.

Google AutoML Η ερευνητική ομάδα AutoML της Google παρουσιάστηκε στο συνέδριο Google I/O '17 [11]. Επικεντρώνεται σε μοντέλα βαθιάς μάθησης (deep learning) και βασίζεται στην ακόλουθη λογική: ένα νευρωνικό δίκτυο-ελεγκτής προτείνει μία αρχιτεκτονική, η οποία εκπαιδεύεται και αξιολογείται η απόδοσή της σε κάποιο συγκεκριμένο πρόβλημα. Στη συνέχεια το δίκτυο-ελεγκτής αποφασίζει πως θα βελτιώσει την υπάρχουσα αρχιτεκτονική με βάση την πληροφορία της αξιολόγησης. Το εργαλείο αυτό έχει εφαρμοστεί σε προβλήματα αναγνώρισης εικόνας και ομιλίας και στοχεύει στη διευκόλυνση της εφαρμογής deep learning, ώστε να είναι προσβάσιμη σε μη-ειδικούς.

[11] `https://www.youtube.com/watch?v=Y2VF8tmLFHw`

3

ΠΕΡΙΓΡΑΦΗ ΣΥΣΤΗΜΑΤΟΣ

Σε αυτό το κεφάλαιο θα περιγράψουμε το σύστημα Automated Data Scientist (ADS), έναν έμπειρο αυτοματοποιημένο αναλυτή δεδομένων για προβλήματα δυαδικής ταξινόμησης. Το λογισμικό είναι γραμμένο σε R και αποτελεί ένα command-line εργαλείο. Στη συνέχεια θα εξετάσουμε τη λειτουργικότητα, την αρχιτεκτονική και τις τεχνικές που επιστρατεύει προκειμένου να επιτελέσει το σκοπό του.

3.1 Σκοπός

Η εργασία μας ανταποκρίνεται στην ανάγκη της κοινότητας της μηχανικής μάθησης για εργαλεία AutoML, τα οποία συμβιβάζουν την αυτοματοποίηση με την κατανοησιμότητα, ώστε να υποβοηθούν, χωρίς να υποκαθιστούν τον αναλυτή. Συγκεκριμένα:

- **Στοχεύει στην επέκταση της τρέχουσας κατάστασης στη μετα-μάθηση εφαρμόζοντας μία νέα τεχνική πρόβλεψης υπερ-παραμέτρων για τη ρύθμιση ενός μοντέλου**.
- **Ενσωματώνει την εμπειρία της κοινότητας της μηχανικής μάθησης μέσω ευριστικών κανόνων που έχουν προκύψει από άρθρα και τη γενικότερη βιβλιογραφία**. Η λειτουργικότητα αυτή μιμείται τη προσέγγιση ενός αναλυτή δεδομένων, ο οποίος συχνά βασίζεται σε ευριστικές κατά τη λήψη σχεδιαστικών αποφάσεων.
- **Εισάγει τη γλώσσα προγραμματισμού R στο σύνολο γλωσσών που χρησιμοποιούνται από εργαλεία AutoML**, το οποίο απ' όσο γνωρίζουμε κυριαρχείται από τις γλώσσες python (SMAC, Spearmint, HPOlib, auto-sklearn, HyperOpt, TPO) και Java (AutoWEKA). Έτσι, ανοίγονται ευκαιρίες για εξερεύνηση και χρήση πακέτων της R κοινότητας, μιας δραστήριας και ετερογενούς ομάδας αναλυτών δεδομένων, μαθηματικών και προγραμματιστών.
- **Αναγνωρίζει τη σημασία της διεπαφής μεταξύ χρήστη και συστήματος**. Αν και το σύστημα θα είναι ανεξάρτητο χάρις στην εφαρμογή μετα-μάθησης και ευριστικών, είναι σημαντική η υποστήριξη της δυνατότητας επέμβασης του αναλυτή. Όσο αφορά την έξοδο του συστήματος θα εξασφαλίζεται η δυνατότητα επαναχρησιμοποίησης του παραγόμενου μοντέλου και η κατανοητή παρουσίαση χρήσιμης γνώσης που παράχθηκε στη διάρκεια του πειράματος.

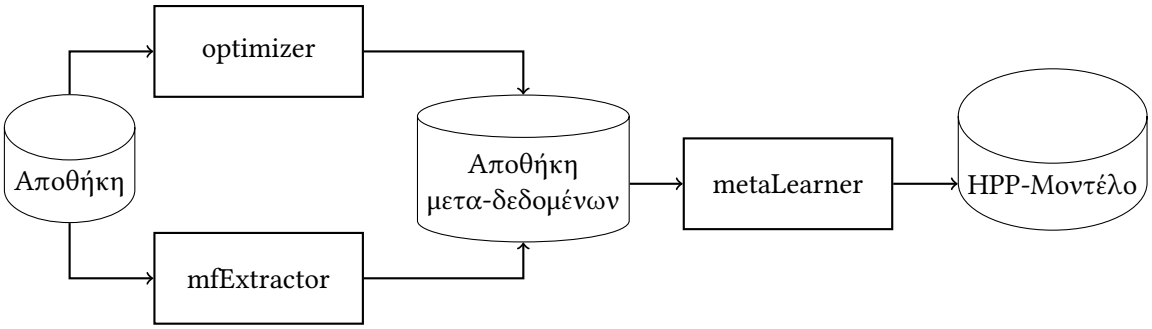
3.2 Αρχιτεκτονική ADS

Το σύστημα Automated Data Scientist αποτελείται από δύο ευκρινώς διαχωριζόμενα υποσυστήματα:

- **Υποσύστημα εκπαίδευσης** Η εκπαίδευση του συστήματος είναι απαραίτητη προκειμένου να έχει την ικανότητα πρόβλεψης των υπερ-παραμέτρων των αλγορίθμων μάθησης που χρησιμοποιεί ο ensemble. Μέσω αυτού του υποσυστήματος είναι δυνατή η παραγωγή των HPP μοντέλων, η οποία απαιτεί την εξαγωγή μετα-χαρακτηριστικών, την εφαρμογή αλγορίθμων βελτιστοποίησης και τέλος, την εκπαίδευση των HPP μοντέλων.

- **Υποσύστημα πειράματος** Περιέχει τη βασική λειτουργικότητα του συστήματος προς τον αναλυτή δεδομένων. Τα συστατικά του αναλαμβάνουν την παραγωγή ενός βέλτιστου ensemble μοντέλων για το σετ δεδομένων εισόδου παρέχοντας τεχνικές προ-επεξεργασίας, οπτικοποίησης, βελτιστοποίησης υπερ-παραμέτρων αλγορίθμων μηχανικής μάθησης, εκπαίδευσης μοντέλων, σχηματισμού ensemble και αξιολόγησης μοντέλου.

3.2.1 Υποσύστημα εκπαίδευσης



Σχήμα 3.1: Το υποσύστημα εκπαίδευσης: Για κάθε σετ δεδομένων δυαδικής ταξινόμησης της Αποθήκης γίνεται εξαγωγή των μετα-χαρακτηριστικών και εύρεση των βέλτιστων υπερ-παραμέτρων με αποτέλεσμα τη δημιουργία ενός σετ-μεταδεδομένων για κάθε υπερ-παράμετρο ενός αλγορίθμου μάθησης. Το πακέτο metaLearner αναλαμβάνει την εκπαίδευση των HPP μοντέλων.

Συστατικά αυτού του υποσυστήματος αποτελούν:

- **Αποθήκη** Πρόκειται για το σύνολο σετ δεδομένων δυαδικής ταξινόμησης, τα οποία έχουν συλλεχθεί για την εκπαίδευση των HPP μοντέλων.

- **optimizer** Το πακέτο αυτό αναλαμβάνει τη βελτιστοποίηση των υπερ-παραμέτρων για δεδομένο αλγόριθμο μάθησης και σετ δεδομένων. Για να το πετύχει αυτό διαθέτει διεπαφή προς την εξωτερική βιβλιοθήκη HPOlib.

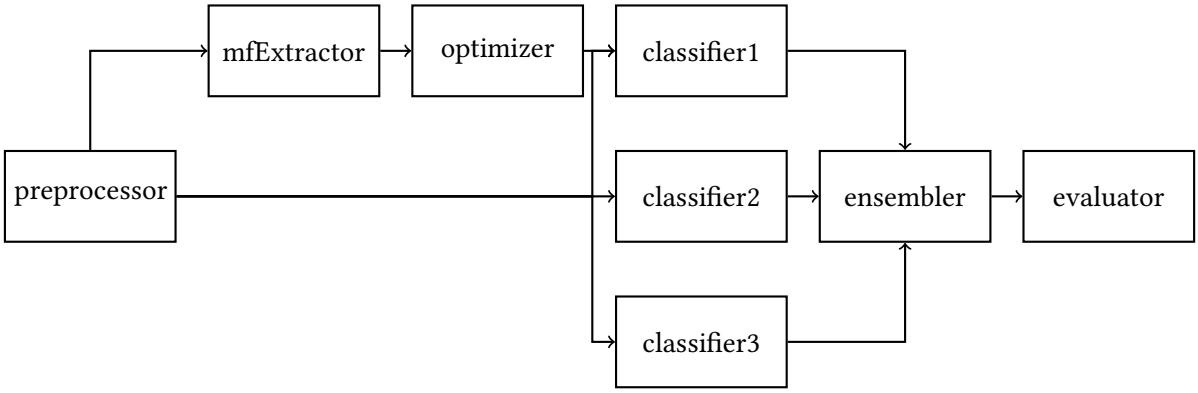
- **mfExtractor** Πακέτο υπεύθυνο για την εξαγωγή μετα-χαρακτηριστικών για ένα σετ δεδομένων.

- **Αποθήκη μετα-δεδομένων** Πρόκειται για μια συλλογή σετ δεδομένων στην οποία κάθε παράδειγμα έχει ως χαρακτηριστικά τα μετα-χαρακτηριστικά και κλάση τη βελτιστοποιημένη υπερ-παράμετρο για ένα σετ δεδομένων και συγκεκριμένο αλγόριθμο.

- **metaLearner** Το πακέτο αυτό αναλαμβάνει την εκπαίδευση ενός HPP μοντέλου για κάθε υπερ-παράμετρο ενός αλγορίθμου μάθησης.

- **HPP μοντέλο** Τελική έξοδος του υποσυστήματος αποτελεί το HPP μοντέλο, το οποίο θα χρησιμοποιηθεί από το υποσύστημα πειράματος. Εκτός από το μοντέλο παρέχεται και πληροφορία χρήσιμη για το καθορισμό των διαστημάτων πρόβλεψης κατά τη πρόβλεψη των υπερ-παραμέτρων.

3.2.2 Υποσύστημα πειράματος



Σχήμα 3.2: Το υποσύστημα πειράματος: Δεδομένου ενός σετ δεδομένων δυαδικής ταξινόμησης στην είσοδο το υποσύστημα αυτό εφαρμόζει την κατάλληλη προ-επεξεργασία και στη συνέχεια εξάγει τα μετα-χαρακτηριστικά, ώστε το πακέτο optimizer να προβλέψει τις βέλτιστες υπερ-παραμέτρους με τη βοήθεια των HPP μοντέλων. Στη συνέχεια εκπαιδεύεται ένα πλήθος μοντέλων για κάθε αλγόριθμο μάθησης και το πακέτο ensembler αναλαμβάνει το σχηματισμό του τελικού ensemble. Τελευταίο στάδιο αποτελεί η αξιολόγηση του πειράματος.

Τα πακέτα που υλοποιούν τη πλήρη διαδικασία της μηχανικής μάθησης για ένα σετ δεδομένων είναι:

- **preprocessor** Περιέχει τεχνικές προ-επεξεργασίας όπως καθαρισμού δεδομένων (αντιμετώπιση άγνωστων και άπειρων τιμών), κανονικοποίησης (z-score και min-max), μετασχηματισμού χαρακτηριστικών (PCA, μετασχηματισμός Box-Cox).

- **mfExtractor** Πρόκειται για το ίδιο πακέτο με αυτό που περιγράφηκε στο υποσύστημα εκπαίδευσης.

- **optimizer** Στην προκειμένη το πακέτο αυτό αναλαμβάνει την πρόβλεψη των υπερ-παραμέτρων κάθε αλγορίθμου μάθησης χρησιμοποιώντας τα ήδη εκπαιδευμένα HPP μοντέλα.

- **classifier**$_i$ Κάθε classifier αντιστοιχεί σε ένα μοντέλο με μοναδικό συνδυασμό υπερ-παραμέτρων και αλγορίθμου μάθησης.

- **ensembler** Το πακέτο αυτό σχηματίζει τον τελικό ensemble από τα διαθέσιμα μοντέλα με την τεχνική της προς τα εμπρός επιλογής μοντέλων.

- **evaluator** Υπεύθυνο για την αξιολόγηση του τελικού ensemble με χρήση τεχνικών που είδαμε στην ενότητα 2.2.3 και τη σύγκριση της μεθόδου μας με άλλες μεθόδους αναφοράς, οι οποίες αναλύονται στην ενότητα 4.5 με χρήση στατιστικών τεστ.

3.3 Τεχνικές ADS

Προκειμένου να ικανοποιήσει το στόχο του το σύστημα χρησιμοποιεί διάφορες τεχνικές εφαρμογής μηχανικής μάθησης, εμπνευσμένες από τη βιβλιογραφία και προσαρμοσμένες στις ανάγκες του. Ένα εργαλείο AutoML θα μπορούσε να αποτελείται εξολοκλήρου από έναν άκριτο συγκερασμό μεθόδων, μια τέτοια προσέγγιση ωστόσο θα ήταν υπολογιστικά και χρονικά ασύμφορη και, το σημαντικότερο, δε θα παρήγαγε γνώση χρήσιμη για την επιστήμη της μηχανικής μάθησης. Έχοντας αναγνωρίσει (Ενότητα 1.1) αυτήν την αντιμετώπιση ως κύριο αίτιο των παθογενειών της μηχανικής μάθησης, σχεδιάσαμε το σύστημα έτσι ώστε να ενσωματώνει τεχνικές του AutoML, οι οποίες του εξασφαλίζουν αποδοτικότητα και το καθιστούν εκπαιδευόμενο, έμπειρο και επεκτάσιμο. Στη συνέχεια θα αναλύσουμε τις τεχνικές αυτές ως προς τη συνεισφορά και την υλοποίησή τους.

3.3.1 Σύστημα βελτιστοποίησης υπερ-παραμέτρων με μετα-μάθηση και χρήση διαστημάτων πρόβλεψης

Η τεχνική αυτή αφορά το στάδιο της εκπαίδευσης ενός μοντέλου, συγκεκριμένα τη ρύθμισή του. Η επιλογή μας να υποκαταστήσουμε την αναζήτηση των υπερ-παραμέτρων με πρόβλεψή τους προσφέρει τα οφέλη της:

- **υπολογιστικής και χρονικής βελτίωσης**. Το πρόβλημα της βελτιστοποίησης, όπως έχουμε δει στην ενότητα 2.3.2, απαιτεί μία επαναληπτική διαδικασία αξιολόγησης μίας κοστοβόρας συνάρτησης που ενέχει το κίνδυνο προσκόλλησης σε τοπικά μέγιστα, αποτελώντας το κατεξοχήν χρονοβόρο στάδιο ενός πειράματος. Αντιθέτως η πρόβλεψη των τιμών με χρήση μοντέλων μηχανικής μάθησης είναι χρονικά σύντομη και υπολογιστικά απλή.

- **της εκπαιδευσιμότητας της μεθόδου μας**, όφελος πολύ σημαντικότερο. Ο εμπλουτισμός του συστήματός μας με νέα προβλήματα το καθιστά πιο έμπειρο, γεγονός που αποτελεί μελλοντική επένδυση στη προσπάθειά του. Σε αντίθετη περίπτωση η αναζήτηση ξεκινά από μηδενική βάση σε κάθε πρόβλημα, καθώς δεν ενσωματώνει μετα-μάθηση.

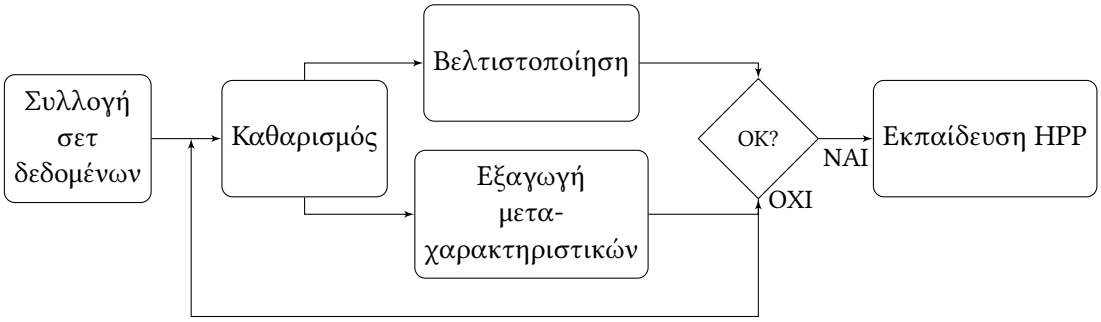
Προκειμένου να εκπαιδεύσουμε το μοντέλο θα χρειαστεί να συλλέξουμε επαρκή σετ δεδομένων, για τα οποία θα βελτιστοποιήσουμε τις υπερ-παραμέτρους και θα υπολογίσουμε τα μεταχαρακτηριστικά. Η αναζήτηση των βέλτιστων υπερ-παραμέτρων έγινε με χρήση του αλγορίθμου TPE της βιβλιοθήκης HPOlib και η εξαγωγή των μετα-χαρακτηριστικών βασίστηκε σε εκτεταμένη βιβλιογραφική έρευνα. Στο σχήμα 3.4 βλέπουμε τη διαδικασία εκπαίδευσης του μοντέλου πρόβλεψης υπερ-παραμέτρων, το οποίο στο εξής θα αποκαλούμε HPP (HyperParameter Prediction) μοντέλο.

Όσο αφορά τις υπερ-παραμέτρους διακρίνουμε 3 είδη, τα οποία απαιτούν διαφορετική αντιμετώπιση:

- **συνεχείς τιμές**, για παράδειγμα η παράμετρος που ορίζει το πλάτος της συνάρτησης ακτινικής βάσης ενός SVM με χρήση γκαουσιανού πυρήνα ή η παράμετρος επιβολής βαρών κανονικοποίησης σε ένα ANN. Για τη πρόβλεψή τους απαιτείται η εκπαίδευση ενός μοντέλου παλινδρόμησης.

- **ακέραιες τιμές**, όπως το πλήθος των γειτόνων στον αλγόριθμο k-κοντινότερου γείτονα ή το βάθος ενός ANN. Εδώ εκπαιδεύεται επίσης ένα μοντέλο παλινδρόμησης και στη συνέχεια επιλέγεται η πλησιέστερη ακέραια τιμή.

- **κατηγορικές τιμές**, όπως η χρήση πυρήνα σε ένα bayesian μοντέλο. Στην προκειμένη απαιτείται η εκπαίδευση ενός μοντέλου ταξινόμησης.



Σχήμα 3.3: Διάγραμμα ροής της διαδικασίας εκπαίδευσης του HPP μοντέλου: αρχικά συλλέγονται τα σετ δεδομένων και στη συνέχεια για το καθένα γίνεται εξαγωγή μετα-χαρακτηριστικών και βελτιστοποίηση υπερ-παραμέτρων. Η συνθήκη τερματισμού ελέγχει αν έχει ολοκληρωθεί η διαδικασία για όλα τα σετ δεδομένων. Τέλος, εκπαιδεύεται το μοντέλο, για το οποίο παράγεται επίσης πληροφορία για τα διαστήματα πρόβλεψης.

Εκμετάλλευση διαστημάτων πρόβλεψης Η απαίτηση ακριβούς πρόβλεψης της βέλτιστης τιμής μιας υπερ-παραμέτρου κρίνεται, τουλάχιστον με τα τρέχοντα μεταχαρακτηριστικά, υπερβολικά απαιτητική, διαπίστωση που οδηγεί στην απαίτηση τεχνικών αποδοτικότερης εκμετάλλευση των μετα-μοντέλων. Οι Feurer, Springenberg, and Hutter [27] αρκέστηκαν στη χρήση των προβλέψεών τους για την εκκίνηση αλγορίθμων βελτιστοποίησης, ενώ οι Soares, Brazdil, and Kuba [25] πρότειναν την κατάταξη προκαθορισμένων συνδυασμών υπερ-παραμέτρων έναντι της πρόβλεψή τους. Η δική μας προσέγγιση συνίσταται στην εκμετάλλευση των διαστημάτων πρόβλεψης που παράγονται από ένα μοντέλο παλινδρόμησης, ώστε να ορίσουμε ένα σύνολο βέλτιστων υπερ-παραμέτρων για κάθε σετ δεδομένων και τελικά να δημιουργήσουμε έναν ensemble με αυτά. Το σύνολο αυτό ορίζεται ως οι υπερ-παράμετροι που βρίσκονται στο p-οστό διάστημα εμπιστοσύνης της πρόβλεψης, όπου το p επιλέχτηκε έτσι ώστε να βελτιστοποιεί την απόδοση του ensemble. Αν η βέλτιστη τιμή βρίσκεται μέσα σε αυτό το διάστημα τότε με χρήση του ensemble θεωρητικά θα εξασφαλιστεί αποτέλεσμα ισάξιο με ένα μοντέλο που θα προέβλεπε επακριβώς τη βέλτιστη τιμή. Μια σύντομη ανάλυση των τρόπων εξαγωγής διαστημάτων πρόβλεψης από μοντέλα παλινδρόμησης βρίσκεται στο Παράρτημα Η′.

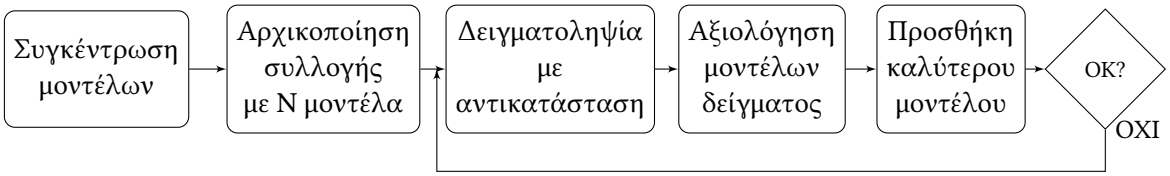
3.3.2 Σύστημα δημιουργίας ensemble με προς τα εμπρός επιλογή μοντέλων

Η τεχνική της χρήσης διαστημάτων πρόβλεψης από το HPP μοντέλο σε συνδυασμό με τη χρήση διαφορετικών αλγορίθμων μηχανικής μάθησης από το σύστημα δημιουργεί ένα πολυπληθές σύνολο μοντέλων προς εκμετάλλευση. Δεδομένης της ετερογένειας και της εφαρμογής βελτιστοποίησης για κάθε αλγόριθμο θεωρούμε πως το σύνολο μοντέλων που δημιουργούμε πληρεί τις προϋποθέσεις που είχε ορίσει ο Dietterich [28]:

> *Μία αναγκαία και ικανή συνθήκη για να είναι μία συλλογή μοντέλων πιο ακριβής από τα μοντέλα που την απαρτίζουν είναι αυτά να είναι ακριβή και ετερογενή.*

Το σύστημά μας βασίζεται στη δουλειά των Caruana et al. [5], οι οποίοι παρουσιάζουν τη μέθοδο του σχηματισμού μίας συλλογής μοντέλων με τη τεχνική της προς τα εμπρός επιλογής μοντέλων. Η τεχνική αυτή, η οποία θυμίζει την προς τα εμπρός επιλογή χαρακτηριστικών που περιγράψαμε στην ενότητα 2.2.2 επιλέγει επαναληπτικά να προσθέσει στη συλλογή το μοντέλο που μεγιστοποιεί την απόδοσή της και το ενσωματώνει παίρνοντας το μέσο όρο των προβλέψεων της συλλογής μετά την προσθήκη του. Στο σχήμα 3.4 μπορούμε να παρατηρήσουμε τον

τρόπο με τον οποίο λειτουργεί η προτεινόμενη μέθοδος.



Σχήμα 3.4: Διάγραμμα ροής της διαδικασίας σχηματισμού μίας συλλογής μοντέλων με την τεχνική της προς τα εμπρός επιλογής μοντέλων: Η συλλογή αρχικοποιείται με τα Ν καλύτερα μοντέλα και στη συνέχεια εφαρμόζεται bootstrapping, σε κάθε επανάληψη του οποίου επιλέγεται ένα υποσέτ των μοντέλων για αξιολόγηση και προστίθεται το βέλτιστο στη συλλογή. Η συνθήκη τερματισμού αντιστοιχεί στο σχηματισμό ενός ensemble προκαθορισμένου πλήθους μοντέλων ή στην ικανοποίηση κάποιας άλλης ποιοτικής συνθήκης (π.χ. ακρίβεια ensemble).

Κατά το σχηματισμό της συλλογής ακολουθούνται κάποιες τεχνικές που στοχεύουν στην αποδοτικότερη σχεδίαση και την αποφυγή υπερ-προσαρμογής:

- **Επιλογή μοντέλων με αντικατάσταση** Στην περίπτωση που κάθε μοντέλο επιτρέπεται να χρησιμοποιηθεί μόνο μία φορά παρατηρήθηκε το πρόβλημα της απότομης πτώσης της απόδοσης της συλλογής, λόγω της αναγκαστικής συμπερίληψης των εναπομεινάντων "κακών" μοντέλων.

- **Αρχικοποίηση συλλογής με τα καλύτερα μοντέλα** Έτσι, αποφεύγεται η υπερ-προσαρμογή στην περίπτωση που διαθέτουμε λίγα μοντέλα.

- **Εφαρμογή συνάθροισης (bootstrapping) κατά τη δειγματοληψία μοντέλων** Σε κάθε επανάληψη της συνάθροισης επιλέγεται ένα δείγμα από τα διαθέσιμα μοντέλα με πιθανότητα συμπερίληψης ενός μοντέλου $p = 0.5$, το οποίο αξιολογείται για την επιλογή του βέλτιστου. Έτσι, αποφεύγεται η υπερ-προσαρμογή στην περίπτωση που διαθέτουμε πολλά μοντέλα, καθώς μειώνεται η πιθανότητα να επιλέξουμε το συνδυασμό μοντέλων που οδηγούν σε αυτή.

3.3.3 Σύστημα ενσωμάτωσης ευριστικών

Συχνά στην πορεία ενός πειράματος μηχανικής μάθησης οι αναλυτές δεδομένων καταφεύγουν στη χρήση ευριστικών. Ως ευριστική ορίζουμε την προσέγγιση της λύσης σε ένα πρόβλημα μέσω μίας πρακτικής μεθόδου, η οποία δεν εγγυάται τη θεωρητικά βέλτιστη λύση, αλλά είναι επαρκώς καλή για το δεδομένο πρόβλημα. Σημαντικές σχεδιαστικές επιλογές βασίζονται, συνειδητά ή ασυνείδητα, σε τέτοιες μεθόδους, που έχουν δοκιμαστεί στο χρόνο και φαίνεται να έχουν ενσωματωθεί στη θεωρία της μηχανικής μάθησης.

Το υπό σχεδίαση σύστημα δε στερείται αυτής της γνώσης, η οποία έχει ενσωματωθεί στον κώδικά του με τη μορφή παραμέτρων στις αποφάσεις που λαμβάνονται στην πορεία ενός πειράματος. Στη συνέχεια παραθέτουμε μερικά παραδείγματα ευριστικών, τα οποία συλλέξαμε από τη βιβλιογραφία:

Το ξυράφι του Όκαμ Η αρχή αυτή αποδίδεται στον William of Ockham και, ως όρος, εισήχθη από τον Libert Froidmont [1]. Συμβουλεύει προς την επιλογή της απλούστερης υπόθεσης μεταξύ ισάξιων, ανταγονιζόμενων υποθέσεων. Στον τομέα της μηχανικής μάθησης χρησιμοποιείται κατά το σχηματισμό του μοντέλου, δίνοντας προτεραιότητα σε απλούστερους αλγορίθμους

[1] `https://en.wikipedia.org/wiki/Occam's_razor`

και απλούστερες παραμετροποιήσεις αλγορίθμων και αποτελεί ευριστική λύση στο πρόβλημα της υπερ-προσαρμογής.

Κανόνας επάρκειας παραδειγμάτων Ο Yaser S. Abu-Mostafa στις διαλέξεις του [2] παρουσιάζει την εξής ευριστική: προκειμένου να είναι εφικτή η εκπαίδευση ενός αλγορίθμου μηχανικής μάθησης πρέπει να ικανοποιείται η σχέση:

$$N \geq 10 * d_{vc} \tag{3.1}$$

όπου N είναι το πλήθος των παραδειγμάτων και d_{vc} ο βαθμός του VC-dimension του αλγορίθμου, ο οποίος ορίζεται ως το μέγιστο πλήθος των σημείων που μπορεί να διαχωρίσει το σετ υπόθεσης H του αλγορίθμου και ισούται με:

- $d + 1$, όπου d η διάσταση της εισόδου, για τους perceptrons.
- $d + 1$, όπου d το πλήθος των βαρών, για ένα ΤΝΔ.
- d, το πλήθος των διανυσμάτων στήριξης, για ένα SVM.

Κανόνας για επιλογή μεγέθους τεστ ελέγχου Ο ίδιος καθηγητής παρουσιάζει και την ακόλουθη ευριστική. Κατά το διαχωρισμό του σετ δεδομένων σε υποσέτ για εκπαίδευση και έλεγχο πρέπει να επέλθει συμβιβασμός, καθώς είναι σημαντική τόσο η καλή εκτίμηση της απόδοσης όσο και η αντιπροσωπευτικότητά της για το τελικό μοντέλο, ανάγκες που σπρώχνουν προς την αύξηση και των δύο υποσέτ. Ευριστικό κανόνα αποτελεί η επιλογή

$$k = \frac{N}{5} \tag{3.2}$$

όπου k το πλήθος των παραδειγμάτων στο σετ ελέγχου και N το συνολικό πλήθος παραδειγμάτων. Η κοινότητα βέβαια συμφωνεί πως η καλύτερη τεχνική είναι αυτή του 10-fold cross-validation (Ενότητα 2.2.3).

Διατηρούμενη διακύμανση PCA Από μία σειρά διαλέξεων [3] προέρχεται και η επόμενη ευριστική: κατά την εφαρμογή PCA επιλέγουμε να κρατήσουμε το πλήθος των κυρίαρχων συνιστωσών που εξασφαλίζουν τη διατήρηση του 98% της διακύμανσης των αρχικών χαρακτηριστικών.

Κανόνας Tukey για αναγνώριση ακραίων τιμών Η αναγνώριση των ακραίων τιμών σε ένα δείγμα γίνεται συνήθως οπτικά, καθώς όπως δήλωσε ο Grubbs [29] ως ακραία τιμή ορίζεται αυτή που απέχει πολύ από τις υπόλοιπες. Ο Tukey [30] ποσοτικοποίησε το γενικό ορισμό, ορίζοντας άνω και κάτω όρια, πέρα από τα οποία οι τιμές θεωρούνται ακραίες

$$min = Q_1 - (IQR * 1.5) max = Q_3 + (IQR * 1.5) \tag{3.3}$$

όπου Q_1 και Q_3 το πρώτο και τρίτο τεταρτημόριο και IQR το διατεταρτημοριακό εύρος (interquartile range) της κατανομής ενός δείγματος.

[2] `http://work.caltech.edu/telecourse.html`
[3] `https://www.coursera.org/learn/machine-learning`

4

ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Περιγραφή πειραμάτων

Στόχος της παρούσας ενότητας είναι ο έλεγχος του συστήματος Automated Data Scientist, καθώς και της συνεισφοράς των τεχνικών που εφαρμόσαμε και περιγράψαμε στην Ενότητα 3.3. Προς αυτό το σκοπό σχεδιάσαμε τα ακόλουθα πειράματα, τα οποία θα αναλύσουμε στη συνέχεια:

- **σχεδιασμός μετρικής για ανίχνευση εξωκείμενων σετ δεδομένων**
- **αξιολόγηση των HPP μοντέλων**
- **αξιολόγηση του ensemble με προς-τα-εμπρός επιλογή μοντέλων**
- **συνολική αξιολόγηση του συστήματος**

Περιγραφή σετ δεδομένων Για τη διεξαγωγή των πειραμάτων συλλέξαμε ένα πλήθος 123 σετ δεδομένων από διάφορες πηγές. (Στο παράρτημα Θ′ βρίσκεται ένας λεπτομερής κατάλογος περιγραφής τους.) Άξονας αναζήτησης κατά τη συλλογή ήταν η εύρεση σετ δεδομένων δυαδικής ταξινόμησης με ετερογενή χαρακτηριστικά, ώστε ο έλεγχος του συστήματος να είναι αντιπροσωπευτικός για το πραγματικό πλήθος σετ δεδομένων. Προκειμένου να υπάρχει μία κοινή διεπαφή για τα πειράματα ήταν απαραίτητος ο "καθαρισμός" των σετ δεδομένων μέσω των ακόλουθων βημάτων:

- **μετατροπή αρχείων σε comma-delimited .csv** Τα πηγαία αρχεία βρίσκονταν σε μορφές .csv, .txt, .xlsx, .arff και .mysql.
- **καθορισμός κλάσης** Στη πλειοψηφία των περιπτώσεων η κλάση αναγνωριζόταν χειροκίνητα από την περιγραφή του σετ δεδομένων. Συλλέχθηκαν και σετ δεδομένων που ήταν πολλαπλής ταξινόμησης και παλινδρόμησης. Στην πρώτη περίπτωση έγινε αντιστοίχηση σε δύο ουσιώδεις κλάσεις, ενώ στη δεύτερη βρέθηκε η μέση τιμή της μεταβλητής κλάσης και χρησιμοποιήθηκε ως κατώφλι για το διαχωρισμό των παραδειγμάτων σε δύο κλάσεις.
- **αναγνώριση άγνωστων τιμών** Στα αρχεία που περιείχαν άγνωστες τιμές χρησιμοποιούνταν διάφοροι συμβολισμοί ("?", "*", "") οι οποίοι αντικαταστάθηκαν από κενά, ώστε να αναγνωρίζονται από την R ως NAs (Not Available).

4.2 Σχεδιασμός μετρικής για ανίχνευση εξωκείμενων σετ δεδομένων

4.2.1 Περιγραφή προβλήματος

Το σύστημα Automated Data Scientist απευθύνεται σε οποιοδήποτε σετ δεδομένων δυαδικής ταξινόμησης. ένα σύνολο πολύ ετερογενές ως προς πεδίο εφαρμογής, φύση χαρακτηριστικών, μέγεθος κλπ. Καθώς το σύστημά μας περιέχει μετα-μάθηση, έχουμε συλλέξει και εκπαιδευτεί σε μια αποθήκη σετ δεδομένων, η οποία αντιπροσωπεύει το κομμάτι του "κόσμου" για το οποίο έχουμε μετα-γνώση. Όταν ο χρήστης τροφοδοτεί το σύστημά μας με ένα νέο σετ δεδομένων υπάρχει το ενδεχόμενο αυτό να βρίσκεται σε ένα ανεξερεύνητο κομμάτι του "κόσμου" με αποτέλεσμα το πείραμα να είναι υποβέλτιστο. Αποτελεί λοιπόν λειτουργική απαίτηση του συστήματος ο υπολογισμός μίας μετρικής, η οποία θα εκφράζει την έλλειψη ετοιμότητας του συστήματός μας να βελτιστοποιήσει ένα άγνωστο σετ δεδομένων.

4.2.2 Μεθοδολογία

Το πρόβλημα θα αντιμετωπιστεί ως ανίχνευση εξωκείμενων παραδειγμάτων (outliers) σε μια συλλογή σετ δεδομένων, καθώς ταιριάζει νοηματικά και πρόκειται για μία καλά μελετημένη τεχνική της μηχανικής μάθησης. Αν και υπάρχουν διάφορες προσεγγίσεις, το γεγονός ότι έχουμε ήδη εξάγει τα μετα-χαρακτηριστικά μας βοηθά να μετατρέψουμε εύκολα το πρόβλημα από ανίχνευση outliers σε μια συλλογή σετ δεδομένων σε ανίχνευση outliers σε ένα σετ μετα-δεδομένων.

Είναι γεγονός πως δεν υπάρχει αυστηρός μαθηματικός ορισμός του τι καθιστά ένα παράδειγμα outlier, οπότε έχουν υλοποιηθεί πολλές προσεγγίσεις Είναι σύνηθες φαινόμενο να γίνεται η παραδοχή συγκεκριμένων κατανομών προκειμένου να γίνει χρήση έτοιμης θεωρίας, κάτι που θέλουμε να αποφύγουμε. Θα ακολουθήσουμε την distance-based προσέγγιση [31] καθώς έχει αποδειχτεί κατάλληλη για πολυδιάστατα δεδομένα και δεν απαιτεί παραδοχές. Η τεχνική του k-κοντινότερου γείτονα αποτελεί χαρακτηριστικό παράδειγμα αυτής της κατηγορίας.

Τα στάδια που ακολουθήθηκαν κατά την εκπαίδευση είναι τα εξής:

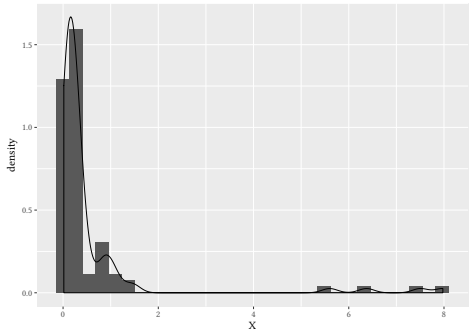
- **εξαγωγή και κανονικοποίηση των μετα-χαρακτηριστικών**
- **υπολογισμός του k με βάση τον ευριστικό κανόνα των Duda, Hart, and Stork [32]**
- **υπολογισμός της μέσης απόστασης από τους k κοντινότερους γείτονες για κάθε σετ δεδομένων εκπαίδευσης**
- **εύρεση του κατωφλίου για θετικές εξωκείμενες τιμές με βάση το τεστ Tukey (Εξίσωση 3.3)**

Κατά το πείραμα εξάγονται τα μετα-χαρακτηριστικά του άγνωστου σετ δεδομένων, κανονικοποιούνται με βάση τις παραμέτρους που έχουν υπολογιστεί κατά την εκπαίδευση και υπολογίζεται η μέση απόσταση από τους k-κοντινότερους γείτονες για να ελεγχθεί αν ξεπερνά το κατώφλι.

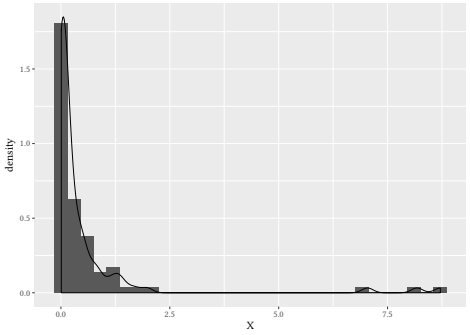
4.2.3 Πειράματα

Στη συνέχεια παρουσιάζουμε τα αποτελέσματα των πειραμάτων που είχαν ως στόχο τον καθορισμό του κατωφλίου, πέρα από το οποίο ένα σετ δεδομένων θεωρείται εξωκείμενο. Καθώς κάθε HPP μοντέλο χρησιμοποιεί διαφορετικό σετ μετα-χαρακτηριστικών (που προέκυψαν από τα πειράματα που περιγράφονται στην Ενότητα 4.3) αρχικά υπολογίστηκαν οι αποστάσεις για κάθε υπερ-παράμετρο και τελικά χρησιμοποιήθηκε ο μέσος όρος τους για τον υπολογισμό της συνολικής μετρικής του συστήματος. Πέρα από τον υπολογισμό του κατωφλίου, ενδιαφερόμαστε και για την στατιστική απόδειξη συσχέτισης μεταξύ του χαρακτηρισμού ενός σετ δεδομένων ως εξωκείμενο και της απόδοσης του αντίστοιχου HPP μοντέλου.

Στα Σχήματα 4.1, 4.2, 4.3, 4.4, 4.5, 4.6 που ακολουθούν παρουσιάζονται τα ιστογράμματα των αποστάσεων στα σετ δεδομένων εκπαίδευσης, η συνάρτηση πυκνότητας πιθανότητας που προσεγγίζεται με χρήση της συνάρτησης density του πακέτου stats της γλώσσας R, το κατώφλι που προκύπτει με βάση το Tukey τεστ και ο Pearson συντελεστής συσχέτισης μεταξύ των αποστάσεων και του σφάλματος [1] του HPP μοντέλου.



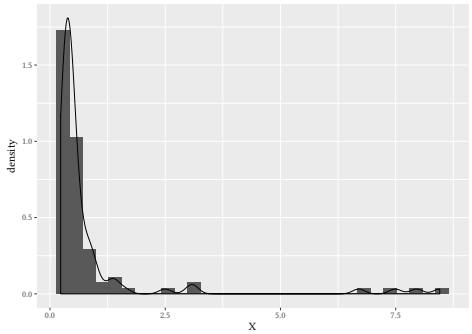
Σχήμα 4.1: Ιστόγραμμα και συνάρτηση πυκνότητας πιθανότητας αποστάσεων σετ δεδομένων για το C-HPP: η συνάρτηση πυκνότητας πιθανότητας έχει γκαουσιανό πυρήνα, παράμετρο bw nrd0 και adjust 2.5. Το κατώφλι προέκυψε 12.9328 και δεν αναγνωρίστηκαν εξωκείμενα σετ δεδομένων στο σετ ελέγχου. Ο συντελεστής συσχέτισης ισούται με 0.438369.



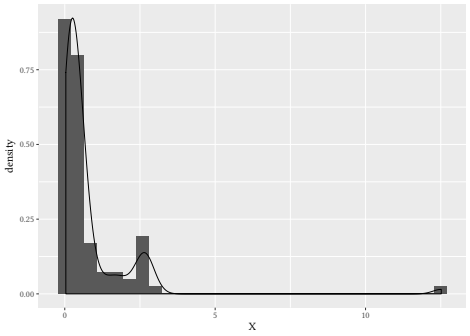
Σχήμα 4.2: Ιστόγραμμα και συνάρτηση πυκνότητας πιθανότητας αποστάσεων σετ δεδομένων για το sigma-HPP: η συνάρτηση πυκνότητας πιθανότητας έχει γκαουσιανό πυρήνα, παράμετρο bw nrd0 και adjust 0.7. Το κατώφλι προέκυψε 31.345375 και αναγνωρίστηκε ένα εξωκείμενο σετ δεδομένων στο σετ ελέγχου με σφάλμα 2.5. Ο συντελεστής συσχέτισης ισούται με 0.0852608.

Συμπέρασμα Τα πειράματά μας αποδεικνύουν κάποια συσχέτιση μόνο στην περίπτωση του C-HPP μοντέλου (0.438369), ενώ στις υπόλοιπες περιπτώσεις η συσχέτιση είναι αμελητέα. Επίσης, παρατηρούμε πως με βάση τις συναρτήσεις πυκνότητας πιθανότητας που έχουμε υπολογίσει τα εξωκείμενα σετ δεδομένων είναι ελάχιστα (δεν υπάρχει κανένα εξωκείμενο σετ δεδομένων στο σετ ελέγχου, εκτός από το sigma-HPP μοντέλο), γεγονός που καθιστά αδύνατη την εξαγωγή στατιστικού συμπεράσματος. Ωστόσο, η μετρική, που παρουσιάζεται στο Σχήμα 4.7 για τα σετ δεδομένων εκπαίδευσης, συνεχίζει να συνιστά χρήσιμη πληροφορία για την τοποθέτηση ενός νέου σετ δεδομένων στην αποθήκη των υπάρχοντων και η στατιστική της συσχέτιση με την απόδοση των HPP μοντέλων ενδεχομένως να αποδειχτεί στο μέλλον.

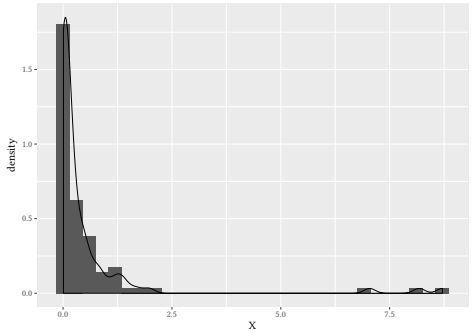
[1] Το σφάλμα υπολογίζεται ως η ρίζα του μέσου τετραγωνικού σφάλματος



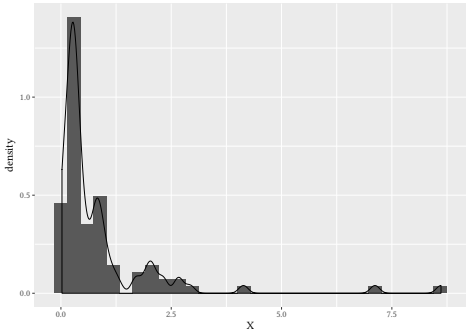
Σχήμα 4.3: Ιστόγραμμα και συνάρτηση πυκνότητας πιθανότητας αποστάσεων σετ δεδομένων για το size-HPP: η συνάρτηση πυκνότητας πιθανότητας έχει γκαουσιανό πυρήνα, παράμετρο bw nrd0 και adjust 1.5. Το κατώφλι προέκυψε 13.35625 και δεν αναγνωρίστηκαν εξωκείμενα σετ δεδομένων στο σετ ελέγχου. Ο συντελεστής συσχέτισης ισούται με -0.017244



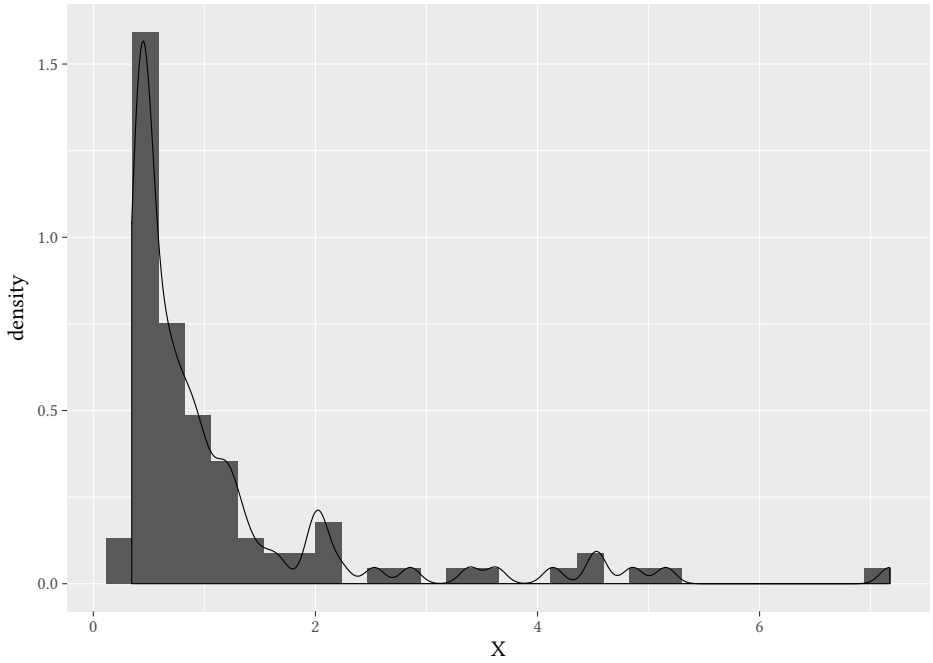
Σχήμα 4.4: Ιστόγραμμα και συνάρτηση πυκνότητας πιθανότητας αποστάσεων σετ δεδομένων για το decay-HPP: η συνάρτηση πυκνότητας πιθανότητας έχει γκαουσιανό πυρήνα, παράμετρο bw nrd0 και adjust 2. Το κατώφλι προέκυψε 20.41505 και δεν αναγνωρίστηκαν εξωκείμενα σετ δεδομένων στο σετ ελέγχου. Ο συντελεστής συσχέτισης ισούται με - 0.151937.



Σχήμα 4.5: Ιστόγραμμα και συνάρτηση πυκνότητας πιθανότητας αποστάσεων σετ δεδομένων για το k-HPP: η συνάρτηση πυκνότητας πιθανότητας έχει γκαουσιανό πυρήνα, παράμετρο bw nrd0 και adjust 1.1. Το κατώφλι προέκυψε 13.8389 και δεν αναγνωρίστηκαν εξωκείμενα σετ δεδομένων στο σετ ελέγχου. Ο συντελεστής συσχέτισης ισούται με 0.218436.



Σχήμα 4.6: Ιστόγραμμα και συνάρτηση πυκνότητας πιθανότητας αποστάσεων σετ δεδομένων για το cp-HPP: η συνάρτηση πυκνότητας πιθανότητας έχει γκαουσιανό πυρήνα, παράμετρο bw nrd0 και adjust 2. Το κατώφλι προέκυψε 13.5359 και δεν αναγνωρίστηκαν εξωκείμενα σετ δεδομένων στο σετ ελέγχου. Ο συντελεστής συσχέτισης ισούται με -0.2123.



Σχήμα 4.7: Ιστόγραμμα και συνάρτηση πυκνότητας πιθανότητας μέσης τιμής αποστάσεων σετ δεδομένων για όλα τα HPP μοντέλα: η συνάρτηση πυκνότητας πιθανότητας έχει γκαουσιανό πυρήνα, παράμετρο bw nrd0 και adjust 0.45. Το κατώφλι προέκυψε 11.114865.

4.3 Αξιολόγηση της τεχνικής βελτιστοποίησης υπερ-παραμέτρων με μετα-μάθηση και χρήση διαστημάτων πρόβλεψης

Όπως είδαμε στην ενότητα 3.3.1 προϊόντα αυτής της τεχνικής είναι τα HPP μοντέλα, καθένα εκ των οποίων έχει εκπαιδευτεί στη πρόβλεψη μίας υπερ-παραμέτρου ενός αλγορίθμου μηχανικής μάθησης. Σε αυτό το σημείο θα αξιολογήσουμε τα μοντέλα αυτά ως προς το σκοπό τους, δηλαδή πόσο καλά προβλέπουν τις βελτιστοποιημένες υπερ-παραμέτρους. Επίσης, θα σχολιάσουμε τη συνεισφορά της χρήσης διαστημάτων πρόβλεψης.

Για την παραγωγή των σετ μετα-δεδομένων, τα οποία χρησιμοποιούνται για την εκπαίδευση των HPP μοντέλων, είναι απαραίτητα δύο στάδια:

- Εξαγωγή των μετα-χαρακτηριστικών κάθε σετ δεδομένων. Τα 81 μετα-χαρακτηριστικά που χρησιμοποιήσαμε περιγράφονται στον Πίνακα 4.1 και υπολογίστηκαν με χρήση του πακέτου mfExtractor του συστήματός μας. Βασίστηκαν σε εκτεταμένη βιβλιογραφική έρευνα και προσπαθούν να συμπεριλάβουν όλα τα είδη μετα-χαρακτηριστικών που εμφανίζονται σε παρόμοιες εργασίες [23, 33, 27]. [2]

Πίνακας 4.1: Λίστα μετα-χαρακτηριστικών, τα οποία χρησιμοποιήθηκαν για την εκπαίδευση των HPP μοντέλων

Απλά	**Στατιστικά Αριθμητικά**
Πλήθος χαρακτηριστικών	Άθροισμα
Λογάριθμος πλήθους χαρακτηριστικών	Μέση τιμή
Πλήθος παραδειγμάτων	Τυπική απόκλιση
Λογάριθμος πλήθους παραδειγμάτων	Ελάχιστη τιμή
Πλήθος χαρακτηριστικών με άγνωστες τιμές	Μέγιστη τιμή
Ποσοστό πλήθους χαρακτηριστικών με άγνωστες τιμές	Κυρτότητα
Πλήθος παραδειγμάτων με άγνωστες τιμές	Λοξότητα
Ποσοστό πλήθους παραδειγμάτων με άγνωστες τιμές	Ποσοστό PCs για 95% διακύμανση
Πλήθος άγνωστων τιμών	Κυρτότητα πρώτης PC
Λογάριθμος πλήθους άγνωστων τιμών	Λοξότητα πρώτης PC
Πλήθος αριθμητικών χαρακτηριστικών	
Πλήθος κατηγορικών χαρακτηριστικών	**Στατιστικά Κατηγορικά**
Πιθανότητες κλάσης	Πλήθος επιπέδων
Ελάχιστη πιθανότητα κλάσης	
Μέγιστη πιθανότητα κλάσης	**Μετα2-**
Μέση τιμή πιθανοτήτων κλάσης	Άθροισμα
Τυπική απόκλιση πιθανοτήτων κλάσης	Μέση τιμή
	Τυπική απόκλιση
Θεωρίας Πληροφορίας	Ελάχιστη τιμή
Εντροπία Κλάσης	Μέγιστη τιμή
	Κυρτότητα
	Λοξότητα

Καθώς το πλήθος των μετα-χαρακτηριστικών είναι δυσανάλογο των διαθέσιμων σετ δεδομένων για εκπαίδευση των HPP μοντέλων, θα επιστρατευθούν τεχνικές επιλογής των βέλτιστων. Σε πρώτη φάση αφαιρέσαμε τα γραμμικά συσχετισμένα χαρακτηριστικά, όπως

[2] Πριν τον υπολογισμό τους έγινε μετατροπή των κατηγορικών χαρακτηριστικών σε μεταβλητές-δείκτες, ώστε να υπάρχει ομοιόμορφη αντιμετώπιση.

αυτά υπολογίστηκαν στο σετ δεδομένων εκπαίδευσης. Η τελική λίστα παρουσιάζεται στον Πίνακα 4.2:

Πίνακας 4.2: Λίστα μετα-χαρακτηριστικών μετά από εφαρμογή φιλτραρίσματος

Άθροισμα αθροισμάτων	Τυπική απόκλιση επιπέδων
Άθροισμα μέγιστων τιμών	Κυρτότητα επιπέδων
Μέση τιμή τυπικών αποκλίσεων	Λοξότητα επιπέδων
Μέση τιμή ελαχίστων τιμών	Πλήθος χαρακτηριστικών
Μέση τιμή κυρτοτήτων	Λογάριθμος πλήθους χαρακτηριστικών
Μέση τιμή λοξοτήτων	Πλήθος παραδειγμάτων
Τυπική απόκλιση ελαχίστων τιμών	Λογάριθμος πλήθους παραδειγμάτων
Ελάχιστη τιμή μέσων τιμών	Ποσοστό αγνώστων τιμών
Ελάχιστη τιμή τυπικών αποκλίσεων	Πλήθος αριθμητικών χαρακτηριστικών
Ελάχιστη τιμή ελαχίστων τιμών	Πλήθος κατηγορικών χαρακτηριστικών
Ελάχιστη τιμή μεγίστων τιμών	Μέγιστη πιθανότητα κλάσης
Ελάχιστη τιμή λοξοτήτων	Μέση τιμή πιθανοτήτων κλάσης
Κυρτότητα ελαχίστων τιμών	Ποσοστό PC για 95% διακύμανση
Κυρτότητα μεγίστων τιμών	Κυρτότητα πρώτης PC
Λοξότητα λοξοτήτων	Λοξότητα PC
Άθροισμα επιπέδων	

- Εύρεση των βέλτιστων υπερ-παραμέτρων για κάθε αλγόριθμο. Προς αυτό το σκοπό χρησιμοποιήθηκε η βιβλιοθήκη HPOlib, την οποία έχουμε περιγράψει στην Ενότητα 2.3.4. Ο αλγόριθμος που επιλέχθηκε ήταν ο Tree Parzen Estimator, καθώς είναι σημαντικά ταχύτερος από τους υπόλοιπους. Από τη πλευρά μας ήταν απαραίτητος ο ορισμός του χώρου αναζήτησης υπερ-παραμέτρων και της συνάρτησης κόστους για κάθε αλγόριθμο, η οποία ορίστηκε ως $Cost = 1 - Accuracy$. Στον Πίνακα 4.3 μπορούμε να δούμε τους αλγορίθμους μάθησης με τους οποίους ασχοληθήκαμε, καθώς και τις υπερ-παραμέτρους τους.

Πίνακας 4.3: Οι αλγόριθμοι που χρησιμοποιεί το σύστημα Automated Data Scientist και οι υπερ-παράμετροί τους, όπως τις ορίζει το πακέτο caret. knn: κ-κοντινότερος γείτονας, rpart: δέντρο ταξινόμησης και παλινδρόμησης (CART), nnet: ANN, svmRadial: SVM με χρήση γκαουσιανού πυρήνα.

knn	rpart	nnet	svmRadial
k	cp	size	C
		decay	sigma

Στα πειράματα που ακολουθούν έχουμε χρησιμοποιήσει τη τεχνική leave one out για την αξιολόγηση των μοντέλων, 10-fold cross-validation για τη ρύθμιση και ως κριτήριο της απόδοσης των μοντέλων παλινδρόμησης τη ρίζα του μέσου τετραγωνικού σφάλματος (root mean squared error).

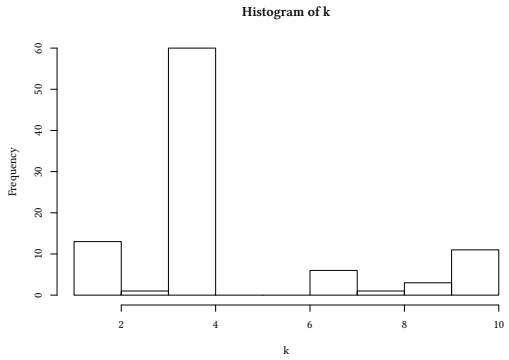
Όσο αφορά τις τεχνικές προ-επεξεργασίας που χρησιμοποιήθηκαν: ως μέθοδο φιλτραρίσματος για επιλογή χαρακτηριστικών ορίζουμε την επιλογή των λιγότερο συσχετισμένων χαρακτηριστικών με βάση τη γραμμική συσχέτιση. Η προς-τα-εμπρός επιλογή χαρακτηριστικών γίνεται με χρήση του πακέτου Boruta [3] και της συνάρτησης rfe του πακέτου caret. Τέλος, ο υπολογισμός

[3] `https://cran.r-project.org/web/packages/Boruta/Boruta.pdf`

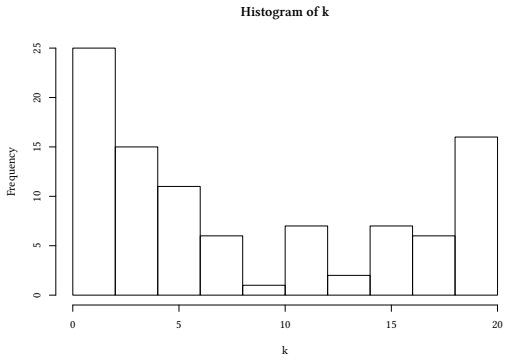
των διαστημάτων πρόβλεψης γίνεται με τη τεχνική του bootstrapping, η οποία περιγράφεται στο Παράρτημα Η΄.

4.3.1 Πρόβλεψη υπερ-παραμέτρου πλήθους γειτόνων για αλγόριθμο k-κοντινότερου γείτονα

Περιγραφή προβλήματος Η υπερ-παράμετρος κ στον αλγόριθμο κ-κοντινότερου γείτονα ορίζει πόσα από τα κοντινότερα παραδείγματα θα ληφθούν υπόψιν κατά την πρόβλεψη. Πρόκειται για μία ακέραια και θετική τιμή, το ιστόγραμμα της οποίας, μετά από τη βελτιστοποίησή της στο σετ δεδομένων εκπαίδευσης φαίνεται στο σχήμα 4.8 για βελτιστοποίηση με χρήση του αλγορίθμου TPE και στο 4.9 για βελτιστοποίηση με πλεγματική αναζήτηση.



Σχήμα 4.8: Ιστόγραμμα υπερ-παραμέτρου k για βελτιστοποίηση με TPE.



Σχήμα 4.9: Ιστόγραμμα υπερ-παραμέτρου k για βελτιστοποίηση με πλεγματική αναζήτηση.

Παρατηρούμε πως στο σχήμα 4.8 το δείγμα μας είναι συγκεντρωμένο στην τιμή 4, με αποτέλεσμα οι υπόλοιπες να είναι εξωκείμενες. Καθώς η ιδιότητα αυτή καθιστά την εκπαίδευση του HPP μοντέλου ιδιαίτερη αναπτύχθηκε μία νέα τεχνική, αυτή της πρόβλεψης με χρήση General-inflated Generalized Poisson μοντέλου, την οποία αναλύουμε στην επόμενη ενότητα. Επίσης, στην Ενότητα 4.3.1 θα εκπαιδεύσουμε ένα μοντέλο παλινδρόμησης με χρήση των τιμών που προέκυψαν από την πλεγματική αναζήτηση, θεωρώντας το κ συνεχές.

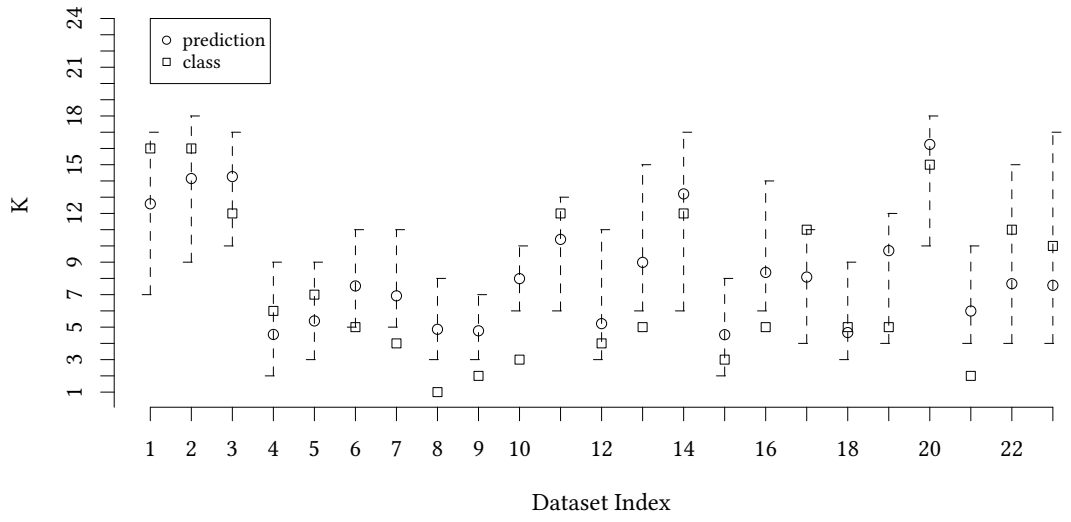
Εκπαίδευση μοντέλου παλινδρόμησης για k του k-κοντινότερου γείτονα

Πίνακας 4.4: Επιλογή αλγορίθμου για την υπερ-παράμετρο k του κ-κοντινότερου γείτονα

	Χωρίς προ-επεξεργασία	Με επιλογή χαρακτηριστικών	Με επιλογή χαρακτηριστικών και κανονικοποίηση
lm	—	$14 \cdot e + 15$	—
lm + log	—	7.946205* [4]	
svmRadial	7.387213	5.813214	—
svmRadial + log	6.255705	5.89	6.096127
ranger + log	5.794678	5.39141	**5.227026**

[4] Η συνοδεία μιας μέτρησης με * υποδηλώνει ότι κρίθηκε χρήσιμο να αφαιρεθούν κάποια παραδείγματα από το σετ ελέγχου, καθώς την επηρέαζαν υπερβολικά. Αποτελεί ικανότητα του συστήματος η αναγνώριση τέτοιων παραδειγμάτων και η αποδοχή αδυναμίας εκπαίδευσης για αυτά.

Στη συνέχεια εκπαιδεύουμε ένα μοντέλο παλινδρόμησης με χρήση του αλγορίθμου randomForest με λογαριθμικό μετασχηματισμό και εφαρμογή φιλτραρίσματος, προς-τα-εμπρός επιλογής χαρακτηριστικών και κανονικοποίησης κατά την προ-επεξεργασία.



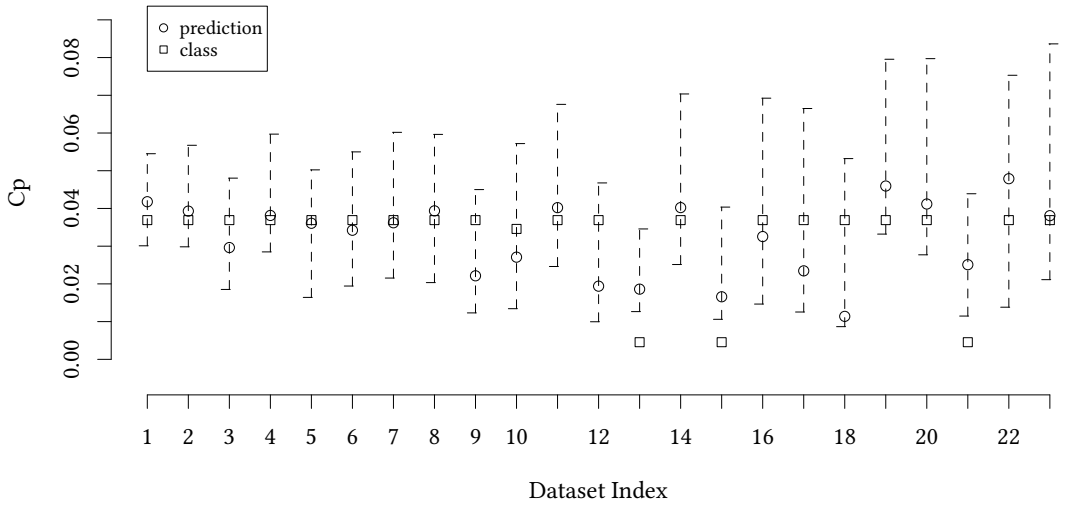
Σχήμα 4.10: Διάγραμμα διαστημάτων πρόβλεψης για υπερ-παράμετρο k

4.3.2 Πρόβλεψη υπερ-παραμέτρου πολυπλοκότητας για αλγόριθμο δέντρου ταξινόμησης

Πίνακας 4.5: Επιλογή αλγορίθμου για την υπερ-παράμετρο cp του δέντρου ταξινόμησης

	Με κανονικοποίηση	Με επιλογή χαρακτηριστικών, κανονικοποίηση
lm	—	1.542080
lm + log	—	0.78553
svmRadial	—	0.9514741
svmRadial + log	0.767217	0.768084
ranger + log	0.755421	**0.692675**
rpart + log	—	0.8007
blackboost + log	—	0.800744
nnet + log	—	1.3621
cubist + log	—	**0.692675**
xgbTree + log	—	1.204

Στη συνέχεια εκπαιδεύουμε ένα μοντέλο παλινδρόμησης με χρήση του αλγορίθμου randomForest με λογαριθμικό μετασχηματισμό και εφαρμογή φιλτραρίσματος, προς-τα-εμπρός επιλογής χαρακτηριστικών και κανονικοποίησης κατά την προ-επεξεργασία.



Σχήμα 4.11: Διάγραμμα διαστημάτων πρόβλεψης για υπερ-παράμετρο cp

4.3.3 Πρόβλεψη υπερ-παραμέτρου πλάτους πυρήνα για τον αλγόριθμο SVM

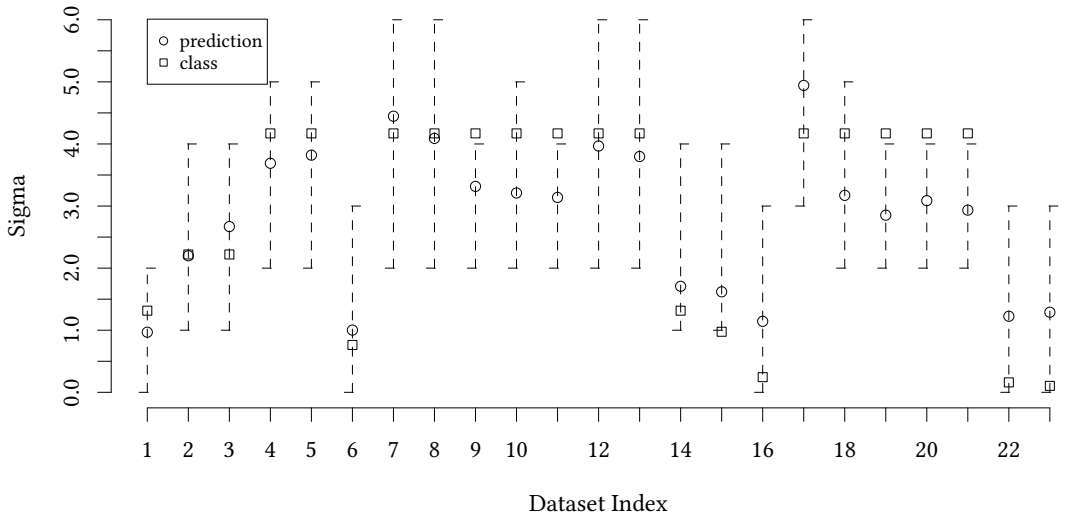
Πίνακας 4.6: Επιλογή αλγορίθμου για την υπερ-παράμετρο sigma του SVM

	Χωρίς προ-επεξεργασία	Με αφαίρεση ακραίων τιμών	Με προς-τα-εμπρός επιλογή χαρακτηριστικών
lm	$15 \cdot e + 13$	6.296997*	$7 \cdot e + 9$
lm + log	$4.2 \cdot e + 10$*	4.239327	2.563388*
svmRadial	2.682720	2.609875	2.685054
svmRadial + log	2.7075885	2.702844	
ranger	—	—	—
ranger + log	—	**2.453491**	2.679798

Πίνακας 4.7: Επιλογή αλγορίθμου για την υπερ-παράμετρο sigma του SVM

	Με φιλτράρισμα χαρακτηριστικών	Με φιλτράρισμα χαρακτηριστικών, αφαίρεση ακραίων τιμών	Με φιλτράρισμα χαρακτηριστικών, αφαίρεση ακραίων τιμών, κανονικοποίηση
lm	$10 \cdot e + 10$	69993	3.076596*
lm + log	Inf	4.055437*	3.423858*
svmRadial	2.588112	—	2.585245
svmRadial + log	2.720634	2.664860	2.690744
ranger	—	2.576414	2.671640
ranger + log	2.6248730	2.696213	**2.530286**

Στη συνέχεια εκπαιδεύουμε ένα μοντέλο παλινδρόμησης με χρήση του αλγορίθμου randomForest με λογαριθμικό μετασχηματισμό και εφαρμογή αφαίρεσης ακραίων τιμών, φιλτραρίσματος, προς-τα-εμπρός επιλογής χαρακτηριστικών κανονικοποίησης κατά την προ-επεξεργασία.



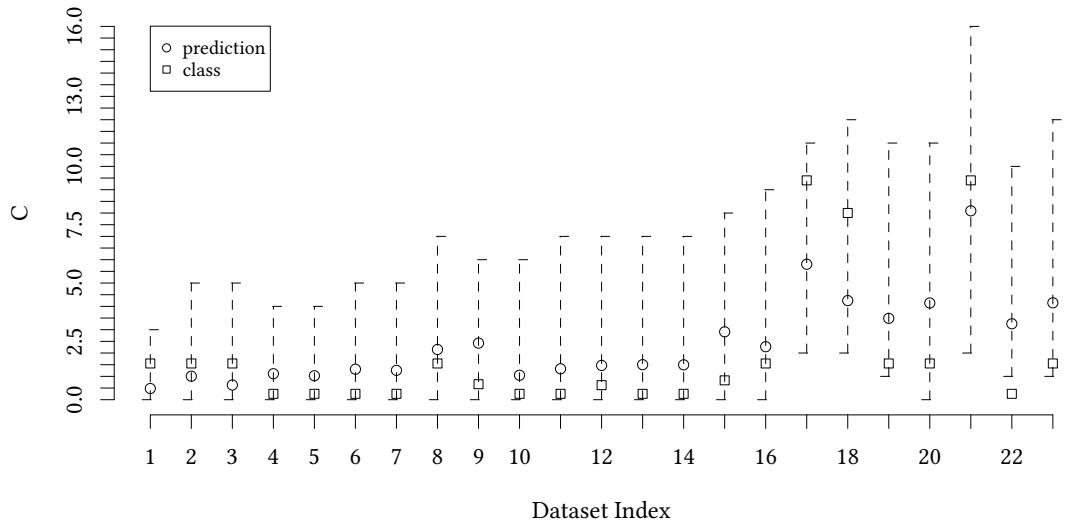
Σχήμα 4.12: Διάγραμμα διαστημάτων πρόβλεψης για υπερ-παράμετρο sigma

4.3.4 Πρόβλεψη υπερ-παραμέτρου κόστους για τον αλγόριθμο SVM

Πίνακας 4.8: Επιλογή αλγορίθμου για την υπερ-παράμετρο C του SVM

	Με κανονικοποίηση	Με επιλογή χαρακτηριστικών	Με επιλογή χαρακτηριστικών, κανονικοποίηση
lm	12.91	12.8225	12.82250
lm + log	15.618852*	9.991954*	9.991954*
svmRadial	10.401787	10.507888	10.21597
svmRadial + log	11.136914	10.676247	10.840548
ranger + log	9.997852	9.319905	**9.294558**

Στη συνέχεια εκπαιδεύουμε ένα μοντέλο παλινδρόμησης με χρήση του αλγορίθμου randomForest με λογαριθμικό μετασχηματισμό και εφαρμογή φιλτραρίσματος, προς-τα-εμπρός επιλογής χαρακτηριστικών και κανονικοποίησης κατά την προ-επεξεργασία.



Σχήμα 4.13: Διάγραμμα διαστημάτων πρόβλεψης για υπερ-παράμετρο C

4.3.5 Πρόβλεψη υπερ-παραμέτρου μεγέθους για τον αλγόριθμο ANN

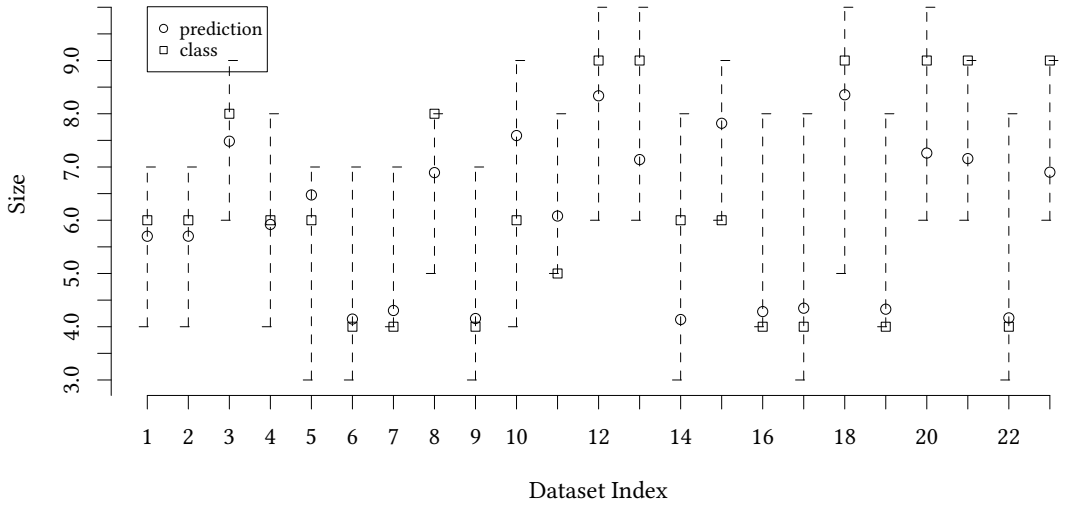
Πίνακας 4.9: Επιλογή αλγορίθμου για την υπερ-παράμετρο size του ANN

	Χωρίς προ-επεξεργασία	Με αφαίρεση ακραίων τιμών, κανονικοποίηση	Με φιλτράρισμα, προς τα εμπρός επιλογή χαρακτηριστικών (Boruta), αφαίρεση ακραίων τιμών
lm + log	4.8964*	4.042651	2.534214
svmRadial	2.354932	2.358656	2.164228
svmRadial + log	2.344243	2.320754	2.113830
ranger	—	—	2.131324
ranger + log	2.286583	2.339490	2.367264
glm	—	—	2.367264
glm + log	—	—	2.534214

Στη συνέχεια εκπαιδεύουμε ένα μοντέλο παλινδρόμησης με χρήση του αλγορίθμου svmRadial με λογαριθμικό μετασχηματισμό και εφαρμογή αφαίρεσης ακραίων τιμών, φιλτραρίσματος, προς-τα-εμπρός επιλογής χαρακτηριστικών και κανονικοποίησης κατά την προ-επεξεργασία.

Πίνακας 4.10: Επιλογή αλγορίθμου για την υπερ-παράμετρο size του ANN

	Με φιλτράρισμα, προς τα εμπρός επιλογή χαρακτηριστικών (Boruta), αφαίρεση ακραίων τιμών, κανονικοποίηση	Με φιλτράρισμα, προς τα εμπρός επιλογή χαρακτηριστικών (rfe), κανονικοποίηση, αφαίρεση ακραίων τιμών
lm + log	2.534214	—
svmRadial	2.113806	2.208724
svmRadial + log	**2.045345**	2.180961
ranger	2.115296	—
ranger + log	2.135809	
glm	2.367264	—
glm + log	2.53421417	—



Σχήμα 4.14: Διάγραμμα διαστημάτων πρόβλεψης για υπερ-παράμετρο size

4.3.6 Πρόβλεψη υπερ-παραμέτρου φθοράς για τον αλγόριθμο ANN

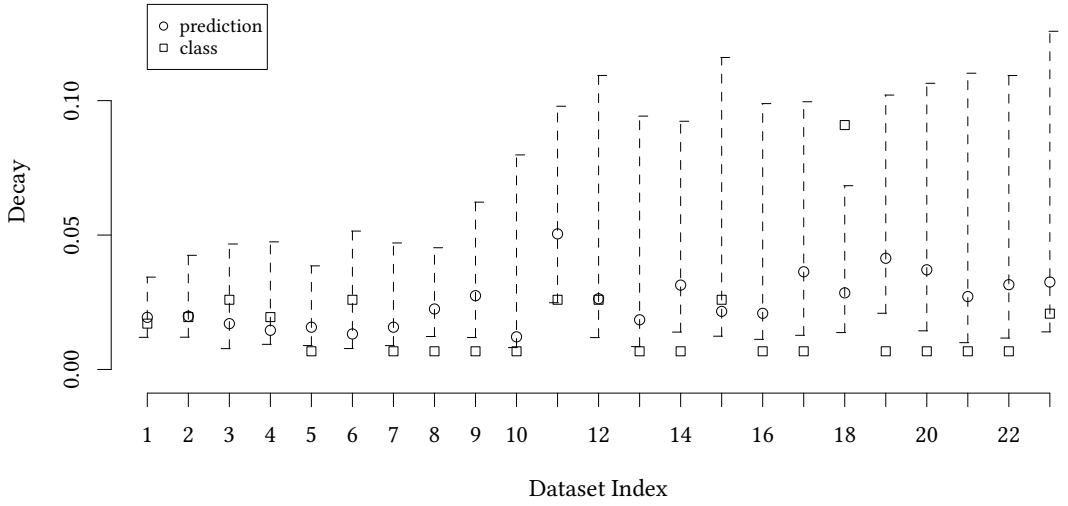
Πίνακας 4.11: Επιλογή αλγορίθμου για την υπερ-παράμετρο decay του ANN

	Με αφαίρεση ακραίων τιμών	Με αφαίρεση ακραίων τιμών κανονικοποίηση	Με φιλτράρισμα, προς τα εμπρός επιλογή χαρακτηριστικών (Boruta) και αφαίρεση ακραίων τιμών
lm	1.3346*	1.340122	0.2674278
lm + log	2.08406*	2.084079	0.277538
svmRadial	0.279624	0.279403	0.270774
svmRadial + log	0.252120	0.253401	0.251316
ranger	0.276386	0.285601	0.244
ranger + log	0.250204	0.248379	0.24289

Πίνακας 4.12: Επιλογή αλγορίθμου για την υπερ-παράμετρο decay του ANN

	Με φιλτράρισμα, προς τα εμπρός επιλογή χαρακτηριστικών (Boruta), αφαίρεση ακραίων τιμών και κανονικοποίηση	Με φιλτράρισμα, προς τα εμπρός επιλογή χαρακτηριστικών (rfe), κανονικοποίηση και αφαίρεση ακραίων τιμών
lm	0.267428	—
lm + log	0.277538	—
svmRadial	0.268606	0.267425
svmRadial + log	0.252016	0.255515
ranger	0.247057	—
ranger + log	**0.240159**	—

Στη συνέχεια εκπαιδεύουμε ένα μοντέλο παλινδρόμησης με χρήση του αλγορίθμου randomforest με λογαριθμικό μετασχηματισμό και εφαρμογή αφαίρεσης ακραίων τιμών, φιλτραρίσματος, προς-τα-εμπρός επιλογής χαρακτηριστικών και κανονικοποίησης κατά την προ-επεξεργασία.



Σχήμα 4.15: Διάγραμμα διαστημάτων πρόβλεψης για υπερ-παράμετρο decay

Συμπεράσματα Τα μοντέλα HPP που εκπαιδεύσαμε δεν καταφέρνουν να προβλέψουν επακριβώς τις υπερ-παραμέτρους. Το γεγονός αυτό μάλλον οφείλεται στα μετα-χαρακτηριστικά και συγκεκριμένα την αδυναμία τους να περιγράψουν τις συναρτήσεις-στόχους που θέσαμε. Τα πειράματά μας ωστόσο αποκαλύπτουν την ύπαρξη κάποια συσχέτισης μεταξύ αυτών και των μετα-χαρακτηριστικών.

Η προσθήκη των διαστημάτων πρόβλεψης αποδεικνύεται ότι αναιρεί την αδυναμία των μοντέλων HPP, καθώς η βέλτιστη τιμή βρίσκεται συνήθως μέσα στο διάστημα πρόβλεψης. Συγκεκριμένα το ποσοστό των σετ δεδομένων ελέγχου για το οποίο η σωστή τιμή της υπερ-παραμέτρου βρίσκεται μέσα στο διάστημα πρόβλεψης είναι 0.7826 για το k, 0.8695 για το cp, 0.8695 για το sigma, 0.9565 για το C, 1 για το size και 0.6086 για το decay. Βέβαια, ακόμη και στις περιπτώσεις που δε βρίσκεται μέσα, παρατηρούμε πως δεν απέχει πολύ. Η επίτευξη αυτή προσδίδει βαρύτητα στην αξιολόγηση του ensemble, η οποία ακολουθεί.

4.4 Αξιολόγηση της τεχνικής σχηματισμού ensemble με προς τα εμπρός επιλογή μοντέλων

Η αξιολόγηση της τεχνικής ensemble που χρησιμοποιήσαμε επιχειρεί να επιβεβαιώσει την προσδοκία ότι ο ensemble παρουσιάζει τουλάχιστον το ίδιο καλή απόδοση με το καλύτερο μοντέλο, το οποίο βρίσκεται στην αποθήκη βελτιστοποιημένων μοντέλων.

Για τα πειράματά μας εκπαιδεύουμε τα μοντέλα στο 80% των σετ δεδομένων και κρατάμε τα υπόλοιπα για την αξιολόγηση του ensemble, η οποία γίνεται ως εξής: εξάγονται τα μεταχαρακτηριστικά των σετ δεδομένων, προβλέπονται οι βέλτιστες υπερ-παράμετροι για κάθε αλγόριθμο μάθησης, εκπαιδεύονται τα μοντέλα και τέλος σχηματίζεται ο ensemble. Για κάθε σετ δεδομένων καταγράφεται η απόδοση του ensemble και του βέλτιστου μοντέλου ως η ακρίβεια (accuracy) που επιτεύχθηκε με 10-fold cross-validation.

Για την αξιολόγηση του συστήματος θα χρησιμοποιηθούν δύο τεχνικές της σύγχρονης βιβλιογραφίας: στατισικά τεστ για τη διαπίστωση σημαντικής διαφοράς στην απόδοση των αλγορίθμων και διαγράμματα προφίλ απόδοσης για την οπτικοποίηση της απόδοσης των αλγορίθμων στα διαφορετικά σετ δεδομένων, τα οποία θα αναλύσουμε στη συνέχεια.

Διαγράμματα προφίλ απόδοσης Τα διαγράμματα προφίλ απόδοσης (performance profile plots) [34] αποτελούν ένα εργαλείο αξιολόγησης και σύγκρισης της απόδοσης εργαλείων βελτιστοποίησης. Χρησιμοποιούνται σε περιπτώσεις εφαρμογής διαφορετικών τεχνικών βελτιστοποίησης σε ένα σύνολο προβλημάτων ως εναλλακτική απεικόνιση εκτενών πινάκων, μιας συνηθισμένης και προβληματικής λύσης. Το προφίλ απόδοσης είναι η αθροιστική συνάρτηση κατανομής μιας τεχνικής για μία μετρική απόδοσης.

Ως μετρική απόδοσης ορίζουμε το λόγο της απόδοσης της τρέχουσας τεχνικής προς τη μεγαλύτερη απόδοση που επιτεύχθηκε από οποιαδήποτε τεχνική για ένα συγκεκριμένο σετ δεδομένων, δηλαδή

$$r_{p,s} = \frac{t_{p,s}}{\max\{t_{p,s} : s \in S\}} \tag{4.1}$$

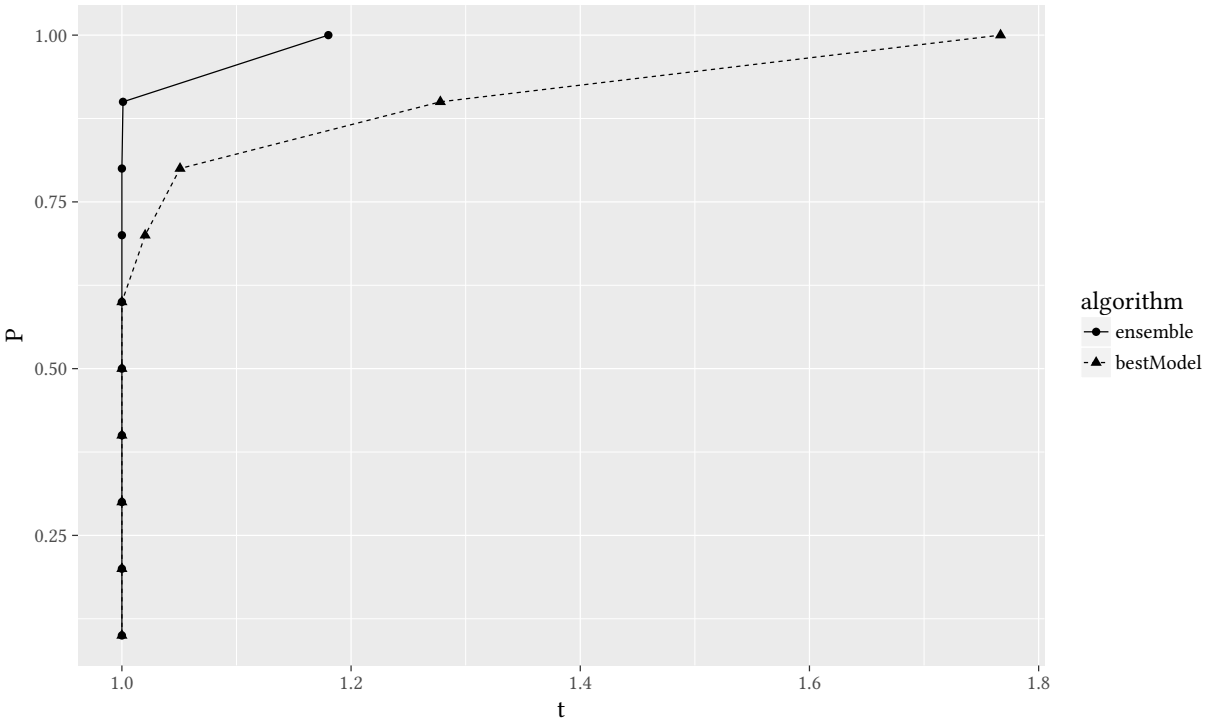
όπου r ο λόγος απόδοσης, t η απόδοση, p το σετ δεδομένων και s η τεχνική.

Το διάγραμμα απεικονίζει τη τιμή

$$\rho_\tau = \frac{size\{p \in P : r_{p,s} \geq \tau\}}{n_p} \tag{4.2}$$

όπου n_p το πλήθος των σετ δεδομένων. Η τιμή αυτή εκφράζει την πιθανότητα μία τεχνική να βρίσκεται σε απόσταση τ από τον καλύτερο λόγο απόδοσης. Επομένως το σημείο $\tau = 1$ εκφράζει τη πιθανότητα μία τεχνική να είναι η βέλτιστη.

Η εφαρμογή του Wilcoxon rank-sum τεστ με επίπεδο εμπιστοσύνης 95%, εφαρμογή διόρθωσης συνέχειας (Παράρτημα ΣΤ′) και συνυπολογισμό του ότι τα πειράματα είναι ζευγαρωμένα (paired) δίνει p-value 0.8336.



Σχήμα 4.16: Διάγραμμα προφίλ απόδοσης για τη σύγκριση του ensemble με το βέλτιστο μοντέλο: Παρατηρούμε πως ο ensemble υπερισχύει του βέλτιστου μοντέλου σε όλα τα σετ δεδομένων.

Συμπεράσματα Το διάγραμμα απόδοσης του Σχήματος 4.16 και το στατιστικό τεστ που εφαρμόσαμε συμφωνούν ότι ο ensemble έχει απόδοση ισότιμη του καλύτερου μοντέλου της βιβλιοθήκης και δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ τους. Το διάγραμμα απόδοσης επίσης δίνει μεγαλύτερη πιθανότητα στον ensemble να έχει απόδοση πλησιέστερη στη βέλτιστη.

4.5 Αξιολόγηση συστήματος Automated Data Scientist

Η αξιολόγηση του Automated Data Scientist στοχεύει να αποδείξει ότι το σύστημα που έχουμε σχεδιάσει έχει απόδοση συγκρίσιμη με τεχνικές της σύγχρονης βιβλιογραφίας. Καθώς η ουσιαστική πρωτοτυπία του συστήματος βρίσκεται στον τρόπο με τον οποίο γίνεται η βελτιστοποίηση των υπερ-παραμέτρων για τα μοντέλα μηχανικής μάθησης που χρησιμοποιούμε θα συγκρίνουμε το σύστημά μας με δύο τεχνικές βελτιστοποίησης:

- **πλεγματική αναζήτηση** Πρόκειται για τη συνηθέστερη τεχνική αναζήτησης υπερπαραμέτρων μέχρι και σήμερα.
- **Tree Parzen Estimator** Η τεχνική αυτή, που έχει περιγραφεί στην ενότητα 2.3.2 αποτελεί state of the art στο χώρο του AutoML.

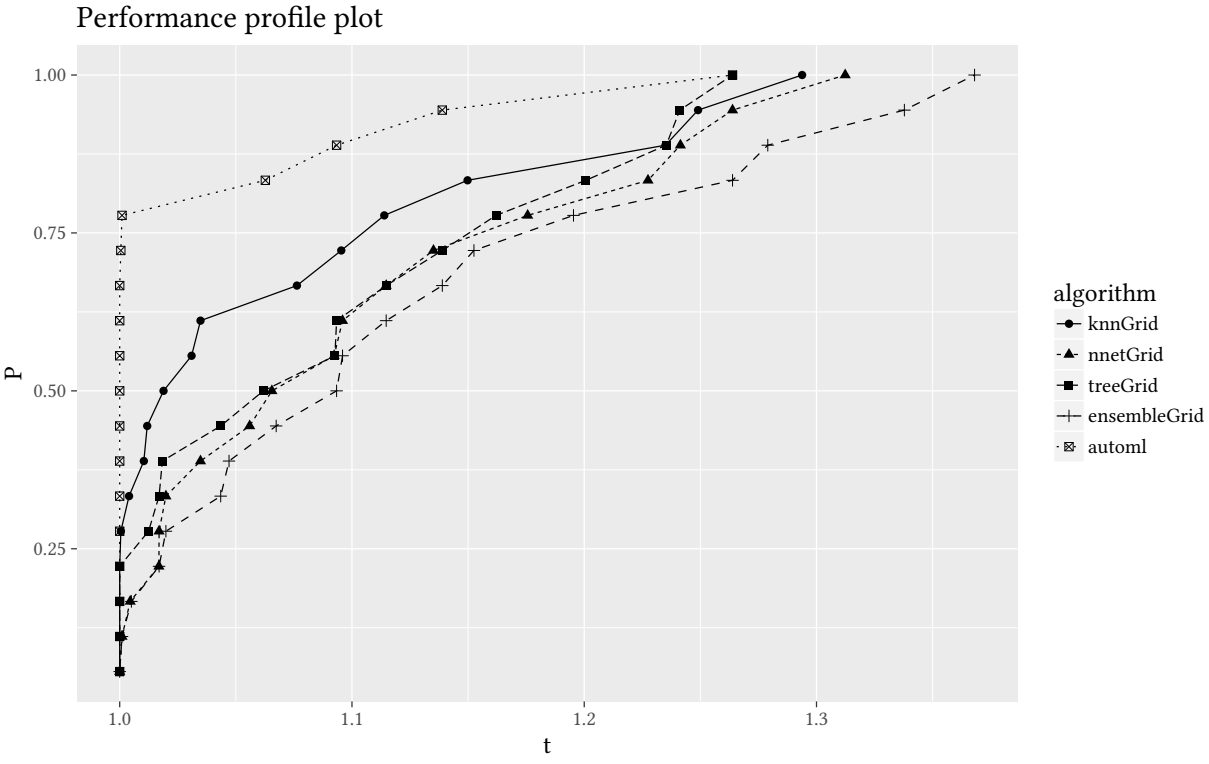
Επίσης, προκειμένου να αξιολογηθεί η συνεισφορά της τεχνικής σχηματισμού ensemble με προς-τα-εμπρός επιλογής μοντέλων πραγματοποιούμε 4 διαφορετικά πειράματα: ένα για κάθε αλγόριθμο μάθησης, όπου εκπαιδεύονται μοντέλα μόνο με το συγκεκριμένο και ένα συνολικό, όπου ο ensemble χρησιμοποιεί όλους τους αλγορίθμους.

Για την αξιολόγηση του συστήματος θα χρησιμοποιηθούν δύο τεχνικές της σύγχρονης βιβλιογραφίας: στατισικά τεστ για τη διαπίστωση σημαντικής διαφοράς στην απόδοση των αλγορίθμων και διαγράμματα προφίλ απόδοσης για την οπτικοποίηση της απόδοσης των αλγορίθμων στα διαφορετικά σετ δεδομένων.

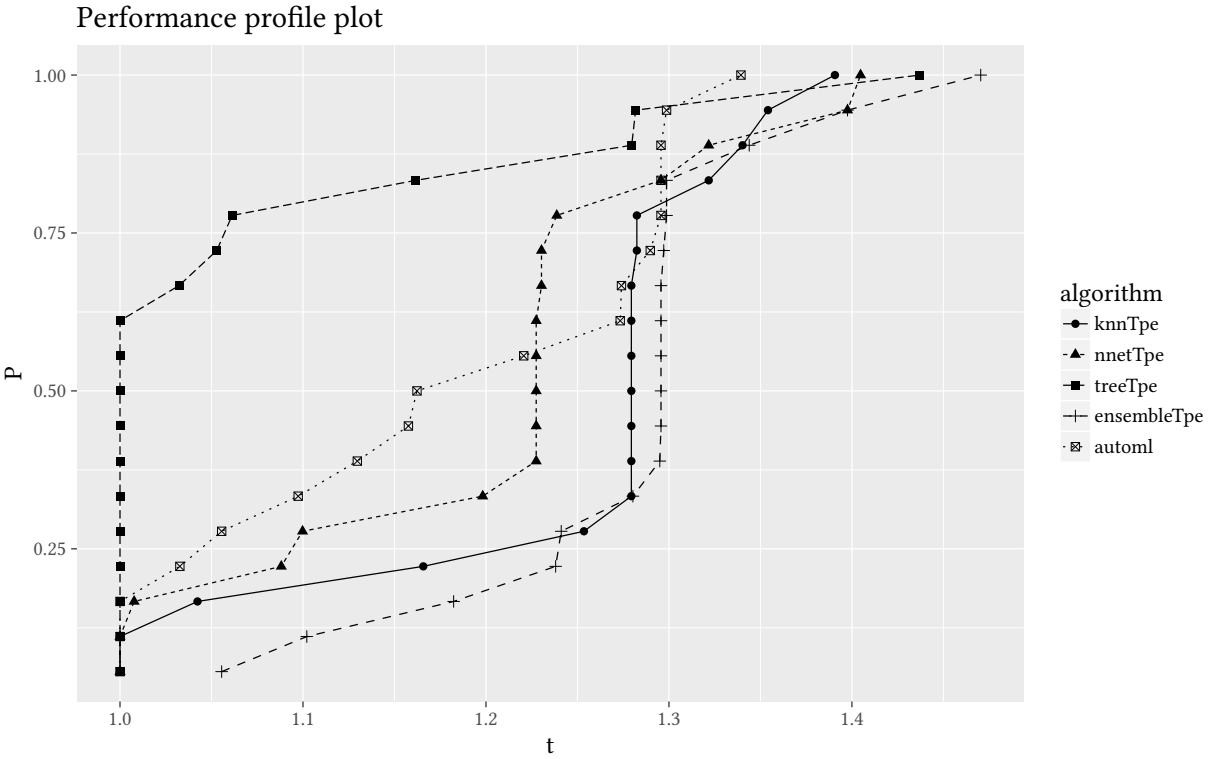
Εφαρμόζωντας το στατιστικό τεστ Friedman rank sum για σύγκριση μεταξύ πολλαπλών αλγορίθμων διαπιστώνουμε σημαντική στατιστική διαφορά καθώς το p-value είναι $6.285 \cdot 10^{-5}$ για επίπεδο εμπιστοσύνης 0.95. Προκειμένου να εντοπίσουμε τα ζεύγη των αλγορίθμων τα οποία προκαλούν τη σημαντική στατιστική διαφορά θα εφαρμόσουμε το Nemenyi post-hoc τεστ.

Πίνακας 4.13: Ο πίνακας αυτός περιέχει τα p-values του post-hoc Nemenyi τεστ για τη διαπίστωση στατιστικής διαφοράς μεταξύ των διαφορετικών μεθόδων ανά ζεύγη. Το τεστ αυτό εφαρμόστηκε μετά από τη διαπίστωση στατιστικής διαφοράς με το Friedman rank sum τεστ. Με έντονη γραφή παρουσιάζονται τα ζεύγη τα οποία εμφανίζουν στατιστικά σημαντική διαφορά.

	automl	ensembleGrid	ensembleTpe	knnGrid	nknnTpe	nnetGrid	nnetTpe	treeGrid
ensembleGrid	**0.0021**	—	—	—	—	—	—	—
ensembleTpe	**0.0021**	1	—	—	—	—	—	—
knnGrid	0.9213	0.1593	0.1593	—	—	—	—	—
knnTpe	0.9213	0.1593	0.1593	1	—	—	—	—
nnetGrid	0.2337	0.8531	0.8531	0.9643	0.9643	-	—	—
nnetTpe	0.7854	0.3024	0.3024	1	1	0.9949	—	—
treeGrid	0.8321	0.2554	0.2554	1	1	0.9902	1	—
treeTpe	1	**0.0067**	**0.0067**	0.9828	0.9828	0.409	0.9213	0.9457



Σχήμα 4.17: Διάγραμμα προφίλ απόδοσης συνολικού συστήματος: σύγκριση του συστήματός μας με τη μέθοδο της πλεγματικής αναζήτησης



Σχήμα 4.18: Διάγραμμα προφίλ απόδοσης συνολικού συστήματος: σύγκριση του συστήματός μας με τη μέθοδο της TPE βελτιστοποίησης.

Συμπεράσματα Με τη βοήθεια του Πίνακα 4.13 μπορούμε να αναγνωρίσουμε τα ζεύγη αλγορίθμων που παρουσιάζουν σημαντική διαφορετική απόδοση στα σετ δεδομένων ελέγχου. Το σύστημά μας διαφέρει σημαντικά μόνο με τις μεθόδους ensembleGrid και ensembleTpe.

Σύμφωνα με το Σχήμα 4.17 το σύστημά μας είναι κατά 77% πιθανότερο να έχει τη βέλτιστη απόδοση σε όλα τα σετ δεδομένων, ενώ παραμένει συνεχώς υψηλότερα από τις υπόλοιπες υπό σύγκριση μεθόδους, για τις οποίες η επιλογή των υπερ-παραμέτρων έγινε με πλεγματική αναζήτηση.

Σύμφωνα με το Σχήμα 4.18, στο οποίο το σύστημά μας συγκρίνεται με μεθόδους όπου χρησιμοποιήθηκε TPE βελτιστοποίηση, καλύτερη μέθοδος στο σύνολο των σετ δεδομένων αποδεικνύεται η treeTpe, με πιθανότητα 67.5%. Το σύστημά μας ωστόσο είναι το δεύτερο αποδοτικότερο και για $\tau = 1.33$ είναι με βεβαιότητα η καλύτερη μέθοδος, δηλαδή για όλα τα σετ δεδομένων βρίσκεται σε απόσταση 1.33 από τη βέλτιστη μέθοδο. Συμπεραίνουμε λοιπόν πως δεν είναι απόλυτα βέλτιστο για όλα τα σετ δεδομένων, αλλά έχει σταθερή καλή απόδοση.

Το σύστημά μας είναι εξίσου αποδοτικό με τις τρέχουσες τεχνικές σχεδόν σε όλα τα σετ εκπαίδευσης ελέγχου. Επίσης, έχει επιτευχθεί η αντικατάσταση της χρονοβόρας βελτιστοποίησης υπερ-παραμέτρων με απλή πρόβλεψη και η δημιουργία μετα-γνώσης, χαρακτηριστικά που δε συνδέονται άμεσα με την απόδοση του συστήματος, ωστόσο επιφέρουν υπολογιστικά οφέλη και δυνατότητες εκμετάλλευσης. Συμπεραίνουμε επομένως πως το σύστημά μας αποτελεί χρήσιμη προσθήκη στο σύνολο των AutoML εργαλείων.

5

ΒΙΒΛΙΟΓΡΑΦΙΑ

Η βιβλιογραφία αντικατοπτρίζει την προσπάθεια της κοινότητας να αυτοματοποιήσει τη διαδικασία της μηχανικής μάθησης προσφέροντας πακέτα λογισμικού που την υλοποιούν και έρευνες που επιχειρούν να επεκτείνουν την τρέχουσα κατάσταση. Οι εξελίξεις εντοπίζονται σε τομείς όπως η βελτιστοποίηση υπερ-παραμέτρων, η εισαγωγή μετα-μάθησης και η ανάπτυξη καλύτερων τεχνικών σχηματισμού πολύπλοκων μοντέλων.

Η ανατροπή του τοπίου για τη βελτιστοποίηση υπερ-παραμέτρων συνέβη όταν οι Bergstra and Bengio [35] απέδειξαν πως η τεχνική της πλεγματικής αναζήτησης επιφέρει αποτέλεσμα χειρότερο από την τυχαία αναζήτηση. Έκτοτε έχουν δοκιμαστεί γενετικοί αλγόριθμοι [14], αναζήτηση κλίσης [15] και η bayesian βελτιστοποίηση, η οποία φαίνεται να έχει επικρατήσει με τη μορφή της τεχνικής SMBO [26]. Οι Hutter et al. [36] εισάγουν την έννοια των χρονικών ορίων στη διαδικασία της βελτιστοποίησης λαμβάνοντας υπόψιν ως κόστος, τόσο την ποιότητ,α όσο και το χρόνο.

Προσπάθειες εισαγωγής μετα-μάθησης κατέβαλαν οι Feurer, Springenberg, and Hutter [27], οι οποίοι χρησιμοποίησαν μετα-χαρακτηριστικά των σετ δεδομένων, ώστε να προβλέψουν τιμές των υπερ-παραμέτρων που, με βάση παλαιότερα πειράματα, πιθανώς να οδηγούν σε καλύτερα μοντέλα. Ωστόσο η διαπίστωση αδυναμίας ικανοποιητικής πρόβλεψης τους οδήγησε σε χρήση των τιμών αυτών για αρχικοποίηση του SMBO αλγόριθμου αναζήτησης που χρησιμοποιούν.

Οι Kuba et al. [37] χρησιμοποιούν δέντρα παλινδρόμησης για να προβλέψουν τις παράμετρους ϵ και σ ενός SVM. Έπειτα από μία διεξοδική ανάλυση των μετα-χαρακτηριστικών καταλήγουν σε μη ικανοποιητικά μοντέλα πρόβλεψης, πρόβλημα που προτείνουν να διορθώσουν εισάγοντας ένα τελικό στάδιο τοπικής αναζήτησης γύρω από τις προβλέψεις τους.

Οι Soares, Brazdil, and Kuba [25] ασχολούνται με τη πρόβλεψη της υπερ-παραμέτρου σ ενός SVM, που καθορίζει το πλάτος του γκαουσιανού πυρήνα. Χρησιμοποιώντας μετα-χαρακτηριστικά των σετ δεδομένων και ένα μοντέλο k-κοντινότερου γείτονα προβλέπουν τη διάταξη προκαθορισμένων τιμών της υπερ-παραμέτρου και με τη τεχνική της Top-N αξιολόγησης επιλέγουν τις βέλτιστες τιμές. Το σύστημά τους δε προβλέπει άμεσα τη βέλτιστη υπερ-παράμετρο, αλλά κατατάσσει ένα προκαθορισμένο σετ ως προς την απόδοσή του στο νέο σετ δεδομένων. Η μεθοδολογία τους απαιτεί τον προ-υπολογισμό της απόδοσης του SVM στα σετ δεδομένων εκπαίδευσης για τις διαθέσιμες τιμές, προσέγγιση απαγορευτική για πολυδιάστατους αλγορίθμους μάθησης. Επίσης, η ελευθερία επιλογής του N αυξάνει την απόδοση, αλλά καθιστά μια σχεδιαστική επιλογή, η οποία μειώνει τον αυτοματισμό της διαδικασίας. Τέλος, η μέθοδός τους εξασφαλίζει χειρότερο αποτέλεσμα από αυτό που επιτυγχάνεται με τη τεχνική cross-validation, ωστόσο κρίνεται ικανοποιητική καθώς επιφέρει χρονική και υπολογιστική βελτίωση.

Ενδιαφέρον παρουσιάζουν οι προσπάθειες των ερευνητών να αναλύσουν τη διαδικασία της μετα-μάθησης, ώστε να ανακαλύψουν τους μηχανισμούς που τη διέπουν με στόχο την αναγνώριση χρήσιμων χαρακτηριστικών, κατάλληλων αλγορίθμων μετα-μάθησης και γενικότερα την

παραγωγή μετα-γνώσης. Οι Rostislav [23] τοποθετούν τη μετα-μάθηση μέσα στον τομέα της Εξόρυξης Δεδομένων, την προσδιορίζουν ως την ικανότητα προσαρμογής με βάση προϋπάρχουσα εμπειρία, παραθέτουν την ιστορική της εξέλιξη και παρουσιάζουν συστήματα που τη χρησιμοποιούν.

Εκτεταμένη έρευνα πάνω στη χρήση IBL αλγορίθμων για πρόβλεψη υπερ-παραμέτρων πραγματοποιούν οι Abdulrahman and Brazdil [38]. Αποδίδουν την καταλληλότητα των αλγορίθμων αυτών για μοντέλα μετα-μάθησης στην εκ φύσεως αδυναμία του προβλήματος για δημιουργία γενικών μοντέλων λόγω των περιορισμένων δεδομένων και της ιδιαιτερότητας κάποιων υπερπαραμέτρων και τη δυνατότητά τους να ενημερώνονται χωρίς εκπαίδευση. Επίσης, παροτρύνουν προς την επιλογή μετα-χαρακτηριστικών με βάση τη σημασιολογική τους συνεισφορά στο πρόβλημα και την αξιολόγησή τους ως προς τη συσχέτισή τους με την υπερ-παράμετρο υπό πρόβλεψη.

Οι Reif, Shafait, and Dengel [33] αναγνωρίζουν το πρόβλημα που προκύπτει στη προσπάθεια εφαρμογής μετα-μάθησης όταν τα μετα-χαρακτηριστικά έχουν διαφορετικό πλήθος για τα σετ δεδομένων και εισάγουν την έννοια των μετα-μετα-χαρακτηριστικών. Πρόκειται για μια προσπάθεια στατιστικής περιγραφής των μετα-χαρακτηριστικών μέσω στατιστικών δεικτών όπως η μέση τιμή, η διακύμανση και η κυρτότητα. Η τεχνική τους, πέρα από τη λύση του προηγούμενου προβλήματος, εμπλουτίζει τη πληροφορία που διατίθεται για μετα-μάθηση. Οι Bensusan and Kalousis [39] εισάγουν τη χρήση πληροφορίας προερχόμενης από τα ιστογράμματα των μετα-χαρακτηριστικών προκειμένου να αποτυπώσουν πληρέστερα την κατανομή τους.

Η συνειδητοποίηση ότι η επίλυση προβλημάτων με χρήση αυτοματοποιημένων συστημάτων απαιτεί τη χρήση πληθώρας τεχνικών και αλγορίθμων οδήγησε στην αναζήτηση βελτιστοποιημένων μεθόδων συνδυασμού μοντέλων μάθησης. Οι Caruana et al. [5] εισάγουν τη μέθοδο σχηματισμού ensembles μοντέλων με την τεχνική της προς-τα-εμπρός επιλογής αναπτύσσοντας τεχνικές για την εξασφάλιση της καλής ποιότητας και την αποφυγή υπερ-προσαρμογής του τελικού ensemble. Διαφορετική προσέγγιση ακολοθούν οι Levésque, Gagné, and Sabourin [40], οι οποίοι απορρίπτουν το σχηματισμό του ensemble ως τελικό στάδιο. Η προσέγγισή τους ενσωματώνει το σχηματισμό στη διαδικασία της βελτιστοποίησης των υπερ-παραμέτρων καθώς η συνάρτηση κόστους υπό βελτιστοποίηση αφορά τη συνεισφορά ενός μοντέλου στον ensemble και η εισαγωγή ενός νέου μοντέλου στη συλλογή γίνεται με μία round robin στρατηγική.

6

ΣΥΝΟΨΗ

Στη διάρκεια της εκπόνησης της διπλωματικής εργασίας φτάσαμε σε συμπεράσματα, τα οποία εκφράζονται με ουσιώδεις διαπιστώσεις για το αντικείμενο, εγγενείς δυσκολίες του προς επίλυση προβλήματος και μία κριτική ενατένιση της σύγχρονης βιβλιογραφίας.

Διαπιστώσαμε πως η σχεδίαση ενός ολοκληρωμένου εργαλείου AutoML απαιτεί πολύπλευρη προσέγγιση, καθώς οφείλει να επιτελεί λειτουργίες για:

- **Την αυτοματοποίηση τετριμμένων καθηκόντων**
- **Την προσαρμοζόμενη αντιμετώπιση κάθε προβλήματος**
- **Την ενσωμάτωση ποικίλλων τεχνικών μηχανικής μάθησης, γεγονός που απαιτεί**
- **Τη χρήση ensembles με ισχυρούς μηχανισμούς επιλογής και αξιολόγησης μοντέλων**

Εντοπίσαμε δυσκολίες κατά την εφαρμογή μετα-μάθησης, καθώς τα μετα-μοντέλα αποδείχτηκαν ανίσχυρα και η ενίσχυσή τους με διαστήματα πρόβλεψης χρονικά και υπολογιστικά απαιτητική. Το γεγονός αυτό ήταν βέβαια αναμενόμενο δεδομένης της νεότητας αυτού του αντικειμένου και των διαπιστωμένων, εγγενών δυσκολιών στη σχετική βιβλιογραφία [27, 37, 25].

Η επαφή με το αντικείμενο του AutoML μας αποκάλυψε τις αδυναμίες της σύγχρονης εφαρμογής μηχανικής μάθησης, που αιτιολογούν, αλλά δε δικαιολογούν, τη στασιμότητα του αντικειμένου: αφέλεια, εμμονή σε παρελθοντικές πανάκειες, αδυναμία μεταφερσιμότητας εφαρμοσμένης γνώσης, έλλειψη σχεδιασμού επαναπαράξιμων πειραμάτων και η αναμενόμενη απαξίωσή τους από την επιστημονική κοινότητα.

Θεωρούμε πως το AutoML αποτελεί προϊόν της συνειδητοποίησης της κοινότητας και το πέρασμα σε ένα στάδιο μάθησης, όπου με ωριμότητα, επαγωγική σκέψη και διερευνητική διάθεση θα προσεγγίσουμε την εκμάθηση, και όχι εφαρμογή, της (μηχανικής) μάθησης.

7

ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Στο σημείο αυτό θα παρουσιάσουμε ιδέες που προέκυψαν κατά την εκπόνηση της διπλωματικής, οι οποίες βελτιώνουν και επεκτείνουν το σύστημα που υλοποιήσαμε. Οι ιδέες αυτές ποικίλλουν ως προς την πολυπλοκότητα, την απαιτητικότητα και τη χρησιμότητά τους, αποτελούν ωστόσο προσθήκες χρήσιμες στο σύνολό τους, καθιστώντας το σύστημά μας πιο αποδοτικό, αυτόνομο και εύχρηστο.

Η θεωρία της μετα-μάθησης έχει δυνατότητα εφαρμογής σε οποιαδήποτε απόφαση επωφελείται από παρελθοντική γνώση. Το σύστημά μας ενσωματώνει τεχνικές μετα-μάθησης για την επιλογή βέλτιστων υπερ-παραμέτρων, ένα απαιτητικό, πολυδιάστατο πρόβλημα βελτιστοποίησης. Αναγνωρίζουμε τη δυνατότητα εφαρμογής μετα-μάθησης σε στάδια όπως η επιλογή αλγορίθμου μάθησης και τεχνικών προ-επεξεργασίας, λειτουργικότητες που θα προσέθεταν επιπλέον εμπειρία στο σύστημά μας.

Επιπλέον επέκταση στο κομμάτι της μετα-μάθησης θα επιτευχθεί με την ενδεχόμενη βελτίωση των υπαρχόντων μετα-μοντέλων μέσω επιπλέον πειραμάτων. Τα νέα πειράματα μπορούν να εξερευνήσουν τη συνεισφορά:

- **Επιπλέον μετα-χαρακτηριστικών** Προϋπόθεση αποτελεί η προσφορά νέων μετα-χαρακτηριστικών από τη σύγχρονη βιβλιογραφία και η ανακάλυψη μετα-χαρακτηριστικών που δεν ανέδειξε η εκτεταμένη έρευνά μας.

- **Βελτίωση των διαστημάτων πρόβλεψης μέσω**:

  - *Πειραματισμού με διάφορα επίπεδα εμπιστοσύνης* Το τρέχον σύστημα χρησιμοποιεί επίπεδο εμπιστοσύνης 95%, αύξηση του οποίου θα οδηγήσει πιθανότερα σε αποδοτικότερο μετα-μοντέλο, αλλά υψηλότερους χρόνους εκπαίδευσης.

  - *Πειραματισμού με διάφορα μεγέθη bootstrap δειγμάτων* Το ανώτερο όριο, 50, έχει τεθεί λόγω χρονικής πολυπλοκότητας. Η διερεύνηση της βελτίωσης που θα προσέφερε η αύξηση των δειγμάτων για τον προσδιορισμό του ανώτατου ορίου, πέρα από το οποίο δε προσφέρεται βελτίωση, θα είναι χρήσιμη στην περίπτωση που η χρονική πολυπλοκότητα δεν αποτελεί κριτήριο.

Ένα πείραμα μηχανικής μάθησης διαθέτει στάδια, τα οποία επιδέχονται παραλληλοποίηση, καθώς διασπώνται σε επιμέρους, ανεξάρτητα καθήκοντα. Παρόλο που ο χρονοβόρος σχηματισμός του τελικού ensemble παρουσιάζει συνέχεια που επιτάσσει την ολική αντιμετώπιση της αποθήκης μοντέλων σε κάθε επανάληψη, υπάρχουν στάδια ευκόλως παραλληλοποιήσιμα (embarassingly parallel), όπως η πρόβλεψη υπερ-παραμέτρων με χρήση μετα-μοντέλων, η αξιολόγηση με τις τεχνικές k-fold cross-validation και Leave one out και η αποθήκευση των εκπαιδευμένων μοντέλων.

Ενδιαφέρουσα επέκταση του συστήματός μας αποτελεί η τροφοδότησή του με επιπλέον σετ

δεδομένων . Προς αυτό το σκοπό δεν απαιτείται επέκταση της λειτουργικότητας του εργαλείου, αλλά συλλογή των σετ δεδομένων και επανεκπαίδευση των μετα-μοντέλων. Έτσι, το σύστημά μας θα αποκτήσει μεγαλύτερο πεδίο εφαρμογής.

Ευκολότερη επίτευξη του προ-αναφερόμενου στόχου και γενικότερη βελτίωση της ευχρηστίας του συστήματος θα επιφέρει η υλοποίηση και ενσωμάτωση λειτουργιών για την εκπαίδευση του συστήματος. Υποψήφιες λειτουργίες αποτελούν:

- **Ένα εργαλείο αυτόματης συλλογής σετ δεδομένων** Κατά τη συλλογή των απαραίτητων μοντέλων παρατηρήσαμε πως στη διαδικασία εμπλέκονται διαφορετικοί πάροχοι, οι οποίοι προσφέρουν σετ δεδομένων σε ποικίλες μορφές αρχείων και διαθέτουν περιγραφές τους που συχνά στερούνται χρήσιμης πληροφορίας. Ως αποτέλεσμα η συλλογή απαιτεί εκτεταμένη αναζήτηση, αξιολόγηση της πληροφορίας και καθαρισμό των αρχείων, στάδια που την καθιστούν χρονικά και νοητικά απαιτητική. Θεωρούμε πως ένα εργαλείο που θα λειτουργεί ως διεπαφή μεταξύ του συστήματός μας και των διαδικτυακών αποθηκών σετ δεδομένων θα προσφέρει ευχρηστία και θα βοηθήσει στην επέκταση του συστήματος.

- **Διεπαφή για εκπαίδευση μετα-μοντέλων** Η εκπαίδευση του συστήματος σε νέα σετ δεδομένων μέσω μιας εύχρηστης διεπαφής θα διευκολύνει την ανανέωση των μετα-μοντέλων και σε συνδυασμό με την προηγούμενη λειτουργικότητα θα διευκολύνει τη βελτίωση του συστήματος.

- **Διεπαφή για ενσωμάτωση ευριστικών κανόνων** Οι ευριστικοί κανόνες αποτελούν σημαντική πηγή λήψης αποφάσεων. Μέσω αυτών ενσωματώνεται στο πείραμα η εμπειρία της βιβλιογραφίας. Είναι λοιπόν επιθυμητό οι ευριστικοί κανόνες του συστήματος να ανανεώνονται τόσο με βάση τη βιβλιογραφία όσο και με τις επιθυμίες του χρήστη, λειτουργικότητα που θα οδηγήσει σε ένα πιο ρυθμιζόμενο σύστημα. Σημαντική είναι η γλώσσα στην οποία θα συντάσσονται οι ευριστικοί κανόνες και ο τρόπος με τον οποίο θα αποθηκεύονται σε μία βάση.

Τέλος, ενδιαφέρον παρουσιάζει η δυνατότητα αυτόματης παραγωγής ευριστικών κανόνων. Αναγνωρίζουμε πως οι ευριστικοί κανόνες αποτελούν μετα-γνώση, η οποία δεν έχει μοντέλο παραγωγής, αλλά διατυπώνεται σε μορφή ποσοτικών κανόνων που προκύπτουν από πειραματική εμπειρία. Καθώς διαθέτουμε ένα σύστημα εκτέλεσης πειραμάτων αναγνωρίζουμε τη δυνατότητα ανατροφοδότησης του συστήματος με την απόδοση των επιλογών του. Όπως λοιπόν η κοινότητα των αναλυτών δεδομένων πειραματίζεται, παρατηρεί και συμπεραίνει για τη δημιουργία ευριστικών κανόνων, έτσι και το σύστημά μας, με χρήση Αναγνώρισης Προτύπων, μπορεί να συλλάβει τους δικούς του ευριστικούς κανόνες εκ των οποίων θα επωφελείται το ίδιο για τη λήψη αποφάσεων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] S. Craw et al. “CONSULTANT: providing advice for the machine learning toolbox”. In: *Research and Development in Expert Systems IX:* Cambridge: Cambridge University Press, Feb. 1993, pp. 5–24. DOI: 10.1017/CBO9780511569944.002. URL: https://www.cambridge.org/core/books/research-and-development-in-expert-systems-ix/consultant-providing-advice-for-the-machine-learning-toolbox/26A8CCC0F5F5E4ECC6CB0B20E2C27083.

[2] Pavel Brazdil, João Gama, and Bob Henery. “Characterizing the applicability of classification algorithms using meta-level learning”. In: *Machine Learning: ECML-94: European Conference on Machine Learning Catania, Italy, April 6–8, 1994 Proceedings*. Ed. by Francesco Bergadano and Luc De Raedt. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994, pp. 83–102. ISBN: 978-3-540-48365-6. DOI: 10.1007/3-540-57868-4_52. URL: http://dx.doi.org/10.1007/3-540-57868-4_52.

[3] D. H. Wolpert and W. G. Macready. “No Free Lunch Theorems for Optimization”. In: *Trans. Evol. Comp* 1.1 (Apr. 1997), pp. 67–82. ISSN: 1089-778X. DOI: 10.1109/4235.585893. URL: http://dx.doi.org/10.1109/4235.585893.

[4] David H. Wolpert. “The Lack of a Priori Distinctions Between Learning Algorithms”. In: *Neural Comput.* 8.7 (Oct. 1996), pp. 1341–1390. ISSN: 0899-7667. DOI: 10.1162/neco.1996.8.7.1341. URL: http://dx.doi.org/10.1162/neco.1996.8.7.1341.

[5] Rich Caruana et al. “Ensemble Selection from Libraries of Models”. In: *Proceedings of the Twenty-first International Conference on Machine Learning*. ICML ’04. Banff, Alberta, Canada: ACM, 2004, pp. 18–. ISBN: 1-58113-838-5. DOI: 10.1145/1015330.1015432. URL: http://doi.acm.org/10.1145/1015330.1015432.

[6] A. M. Turing. “Computers &Amp; Thought”. In: ed. by Edward A. Feigenbaum and Julian Feldman. Cambridge, MA, USA: MIT Press, 1995. Chap. Computing Machinery and Intelligence, pp. 11–35. ISBN: 0-262-56092-5. URL: http://dl.acm.org/citation.cfm?id=216408.216410.

[7] Thomas M. Mitchell. *Machine Learning*. 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN: 0070428077, 9780070428072.

[8] S. B. Kotsiantis. “Supervised Machine Learning: A Review of Classification Techniques”. In: *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2007, pp. 3–24. ISBN: 978-1-58603-780-2. URL: http://dl.acm.org/citation.cfm?id=1566770.1566773.

[9] G. E. P. Box and D. R. Cox. “An Analysis of Transformations”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 26.2 (1964), pp. 211–252. ISSN: 00359246. URL: http://www.jstor.org/stable/2984418.

[10] Christophe Ambroise and Geoffrey J. McLachlan. “Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data”. In: *Proceedings of the National Academy of Sciences of the United States of America* 99.10 (2002), pp. 6562–6566. ISSN: 00278424. URL: http://www.jstor.org/stable/3058705.

[11] Foster Provost and Ron Kohavi. “On Applied Research in Machine Learning”. In: *Machine learning*. 1998, pp. 127–132.

[12] Brett Presnell. “An introduction to Categorical Data Analysis Using R”. In: ().

[13] Ilya Loshchilov and Frank Hutter. “CMA-ES for Hyperparameter Optimization of Deep Neural Networks”. In: *CoRR* abs/1604.07269 (2016). URL: http://arxiv.org/abs/1604.07269.

[14] S. A. Rojas and D. Fernandez-Reyes. “Adapting multiple kernel parameters for support vector machines using genetic algorithms”. In: *2005 IEEE Congress on Evolutionary Computation*. Vol. 1. Sept. 2005, 626–631 Vol.1. DOI: 10.1109/CEC.2005.1554741.

[15] Fabian Pedregosa. *Hyperparameter optimization with approximate gradient*. Version 1. arXiv: 1602.02355v5 [cs.LG].

[16] J. A. Nelder and R. Mead. “A Simplex Method for Function Minimization”. In: *The Computer Journal* 7.4 (1965), pp. 308–313. DOI: 10.1093/comjnl/7.4.308. eprint: http://comjnl.oxfordjournals.org/content/7/4/308.full.pdf+html. URL: http://comjnl.oxfordjournals.org/content/7/4/308.abstract.

[17] D. Huang et al. “Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models”. In: *Journal of Global Optimization* 34.3 (2006), pp. 441–466. ISSN: 1573-2916. DOI: 10.1007/s10898-005-2454-3. URL: http://dx.doi.org/10.1007/s10898-005-2454-3.

[18] J. B. Mockus and L. J. Mockus. “Bayesian approach to global optimization and application to multiobjective and constrained problems”. In: *Journal of Optimization Theory and Applications* 70.1 (1991), pp. 157–172. ISSN: 1573-2878. DOI: 10.1007/BF00940509. URL: http://dx.doi.org/10.1007/BF00940509.

[19] H. J. Kushner and L. J. Mockus. “A New Method of Locating the Maximum Point of an Arbitrary Mult”. In: *Journal of Basic Engineering* 86.1 (1964), pp. 97–106. ISSN: 1573-2878. DOI: 10.1115/1.3653121. URL: http://dx.doi.org/10.1115/1.3653121.

[20] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. “Sequential Model-Based Optimization for General Algorithm Configuration”. In: *Learning and Intelligent Optimization: 5th International Conference, LION 5, Rome, Italy, January 17-21, 2011. Selected Papers*. Ed. by Carlos A. Coello Coello. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 507–523. ISBN: 978-3-642-25566-3. DOI: 10.1007/978-3-642-25566-3_40. URL: http://dx.doi.org/10.1007/978-3-642-25566-3_40.

[21] M. Wistuba, N. Schilling, and L.Schmidt-Thieme. “Sequential Model-Free Hyperparameter Tuning”. In: *2015 IEEE International Conference on Data Mining*. Nov. 2015, pp. 1033–1038. DOI: 10.1109/ICDM.2015.20.

[22] James Bergstra et al. “Algorithms for Hyper-parameter Optimization”. In: *Proceedings of the 24th International Conference on Neural Information Processing Systems*. NIPS’11. Granada, Spain: Curran Associates Inc., 2011, pp. 2546–2554. ISBN: 978-1-61839-599-3. URL: http://dl.acm.org/citation.cfm?id=2986459.2986743.

[23] Striz Rostislav. “Extending Metalearning to Data Mining and KDD”. In: *Metalearning: Applications to Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 61–72. ISBN: 978-3-540-73263-1. DOI: 10.1007/978-3-540-73263-1_4. URL: http://dx.doi.org/10.1007/978-3-540-73263-1_4.

[24] Matthias Feurer, Jost Springenberg, and Frank Hutter. *Initializing Bayesian Hyperparameter Optimization via Meta-Learning*. 2015. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10029.

[25] Carlos Soares, Pavel B. Brazdil, and Petr Kuba. “A Meta-Learning Method to Select the Kernel Width in Support Vector Regression”. In: *Machine Learning* 54.3 (2004), pp. 195–209. ISSN: 1573-0565. DOI: 10.1023/B:MACH.0000015879.28004.9b. URL: http://dx.doi.org/10.1023/B:MACH.0000015879.28004.9b.

[26] Chris Thornton et al. “Auto-WEKA: Automated Selection and Hyper-Parameter Optimization of Classification Algorithms”. In: *CoRR* abs/1208.3719 (2012). URL: http://arxiv.org/abs/1208.3719.

[27] Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. “Using Meta-learning to Initialize Bayesian Optimization of Hyperparameters”. In: *Proceedings of the 2014 International Conference on Meta-learning and Algorithm Selection - Volume 1201*. MLAS’14. Prague, Czech Republic: CEUR-WS.org, 2014, pp. 3–10. ISBN: 1613-0073. URL: http://dl.acm.org/citation.cfm?id=3015544.3015549.

[28] Thomas G. Dietterich. “Ensemble Methods in Machine Learning”. In: *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. ISBN: 978-3-540-45014-6. DOI: 10.1007/3-540-45014-9_1. URL: http://dx.doi.org/10.1007/3-540-45014-9_1.

[29] F. E Grubbs. “Procedures for detecting outlying observations in samples”. In: *Technometrics, Volume 11 (1969) - Volume 1201*. 1969. URL: http://www.citeulike.org/user/gatsoulis/article/7096000.

[30] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[31] Edwin M. Knorr, Raymond T. Ng, and Vladimir Tucakov. “Distance-based Outliers: Algorithms and Applications”. In: *The VLDB Journal* 8.3-4 (Feb. 2000), pp. 237–253. ISSN: 1066-8888. DOI: 10.1007/s007780050006. URL: http://dx.doi.org/10.1007/s007780050006.

[32] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000. ISBN: 0471056693.

[33] Matthias Reif, Faisal Shafait, and Andreas Dengel. *Meta 2-Features: Providing Meta-Learners More Information*.

[34] Elizabeth D. Dolan and Jorge J. Moré. “Benchmarking optimization software with performance profiles”. In: *Mathematical Programming* 91.2 (2002), pp. 201–213. ISSN: 1436-4646. DOI: 10.1007/s101070100263. URL: http://dx.doi.org/10.1007/s101070100263.

[35] James Bergstra and Yoshua Bengio. “Random Search for Hyper-parameter Optimization”. In: *J. Mach. Learn. Res.* 13 (Feb. 2012), pp. 281–305. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2188385.2188395.

[36] F. Hutter et al. “Time-Bounded Sequential Parameter Optimization”. In: *Proceedings of the conference on Learning and Intelligent OptimizatioN (LION 4)*. Jan. 2010.

[37] Petr Kuba et al. “Exploiting sampling and meta-learning for parameter setting support vector machines”. In: *Proceedings of the IBERAMIA*. Vol. 2002. 2002, pp. 217–225.

[38] Salisu Mamman Abdulrahman and Pavel Brazdil. “Measures for Combining Accuracy and Time for Meta-learning”. In: *Proceedings of the 2014 International Conference on Meta-learning and Algorithm Selection - Volume 1201*. MLAS’14. Prague, Czech Republic: CEUR-WS.org, 2014, pp. 49–50. ISBN: 1613-0073. URL: http://dl.acm.org/citation.cfm?id=3015544.3015557.

[39] Hilan Bensusan and Alexandros Kalousis. “Estimating the Predictive Accuracy of a Classifier”. In: *Proceedings of the 12th European Conference on Machine Learning*. EMCL ’01. London, UK, UK: Springer-Verlag, 2001, pp. 25–36. ISBN: 3-540-42536-5. URL: http://dl.acm.org/citation.cfm?id=645328.650030.

[40] Julien-Charles Levésque, Christian Gagné, and Robert Sabourin. *Bayesian Hyperparameter Optimization for Ensemble Learning*. Version 1. arXiv: 1605.06394 [cs.LG].

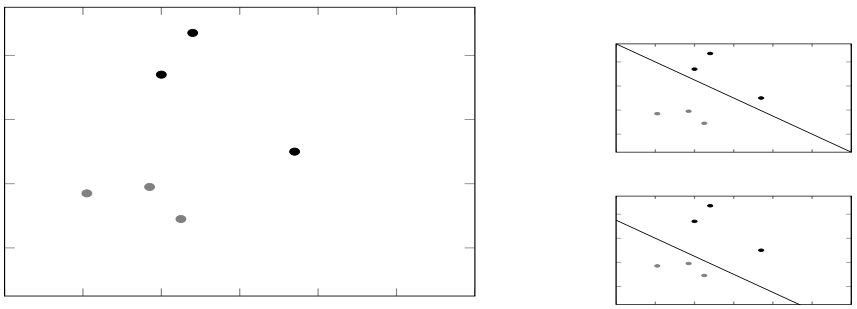
Παραρτήματα

Α΄

ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΙΚΗΣ ΣΤΗΡΙΞΗΣ

Πρόκειται για μία από τις πιο πρόσφατες τεχνικές στον τομέα της επιβλεπόμενης μάθησης, που χρησιμοποιείται ευρέως τόσο σε προβλήματα ταξινόμησης, όσο και σε προβλήματα παλινδρόμησης. Εισήχθησαν από τους Vapnik and Lerner [1] το 1963.

Έστω ότι βρισκόμαστε μπροστά από ένα πρόβλημα ταξινόμησης, με την κλάση να παίρνει 2 τιμές και τα παραδείγματα να έχουν 2 χαρακτηριστικά. Τότε ο χώρος μας έχει τη μορφή του σχήματος Α΄.1, όπου διαπιστώνουμε πως υπάρχουν διαφορετικές υποθέσεις που διαχωρίζουν σωστά τα δεδομένα.



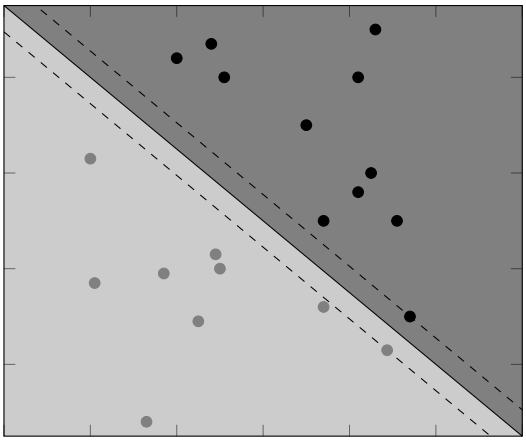
Σχήμα Α΄.1: Τα δεδομένα που προσφέρονται στον αλγόριθμο (αριστερά) οφείλουν να διαχωριστούν ως προς την κλάση μέσω της υπόθεσης. Όπως φαίνεται στα δύο σχήματα δεξιά το πρόβλημα δεν έχει μοναδική λύση, επομένως ο SVM το αναδιατυπώνει συγκεκριμενοποιώντας το στόχο της εκπαίδευσης.

Οι μηχανές διανυσματικής στήριξης μπορούν να απαντήσουν στο εύλογο ερώτημα: "Ποια από τις παραπάνω υποθέσεις είναι η καλύτερη;" Λαμβάνοντας υπόψιν πως η ποιότητα μιας υπόθεσης καθορίζεται βασικά από την ικανότητά της να γενικεύει, οι αλγόριθμοι αυτοί επιλέγουν την υπόθεση έτσι, ώστε τα πιο κοντινά σημεία που ταξινομούνται σε διαφορετικές κατηγορίες να χωρίζονται από όσο το δυνατόν μεγαλύτερο κενό. Τα σημεία αυτά ονομάζονται διανύσματα στήριξης.

Θεωρητική θεμελίωση Έστω πως τα δεδομένα μας είναι δισδιάστατα και επιχειρούμε να ορίσουμε την ευθεία που εξασφαλίζει μεγαλύτερο κενό μεταξύ των κοντινότερων σημείων που ανήκουν σε διαφορετική κλάση. Η ευθεία που αναζητούμε φαίνεται στο παρακάτω σχήμα και δίνεται από τον τύπο $w^T x = 0$ και οι ευθείες που περνούν από τα διανύσματα στήριξης ορίζονται ως $w^T x = 1$ και$w^T x = -1$

Πριν συνεχίσουμε θα χρειαστεί να ορίσουμε δύο τεχνικές παραδοχές:

- Όπως είναι γνωστό, ένα επίπεδο είναι αμετάβλητο ως προς την κλιμάκωση, δηλαδή με όποια σταθερά και να το πολλαπλασιάσω θα συνεχίσω να έχω το ίδιο επίπεδο. Για αυτό θα κανονικοποιούμε ώστε $\left\|w^T x\right\| = 1$



Σχήμα Α΄.2: Ο SVM σχηματίζει την υπόθεση (συνεχής γραμμή) έτσι ώστε τα σημεία που βρίσκονται πιο κοντά σε αυτήν (τα σημεία που τέμνονται από τις διακεκομμένες γραμμές) να απέχουν όσο το δυνατόν περισσοτερο. Τα σημεία αυτά ονομάζονται διανύσματα στήριξης.

- Μας βολεύει να βγάλουμε τον σταθερό όρο w_0 από το διάνυσμα w και να ορίσουμε την επιφάνεια ως $w^T x + b = 0$, όπου προφανώς το b αντιστοιχεί στο w_0.

Πώς υπολογίζουμε την απόσταση ενός σημείου από ένα υπερεπίπεδο; Αρχικά παρατηρώ πως το w είναι κάθετο στο υπερεπίδεδο. Αυτό αποδεικνύεται πολύ εύκολα ως εξής: Έστω δύο σημεία x' και x'' πάνω στο υπερεπίπεδο. Τότε ισχύει $w^T x' + b = 0$ και $w^T x'' + b = 0$. Επομένως $w^T(x' - x'') = 0$, δηλαδή το w είναι κάθετο σε οποιαδήποτε ευθεία ενώνει δύο σημεία του υπερεπιπέδου.

Η απόσταση του σημείου x_n από το υπερεπίπεδο υπολογίζεται ως εξής: παίρνω οποιοδήποτε σημείο x στο υπερεπίπεδο και προβάλλω το διάνυσμα $x_n - x$ στο w. Η πράξη αυτή, με το κανονικοποιημένο w να ορίζεται ως $\bar{w} = \frac{w}{\|w\|}$, δίνεται από τον τύπο:

$$distance = |\bar{w}(x_n - x)| = \frac{1}{\|w\|}|w^T x_n - w^T x| = \frac{1}{\|w\|}|w^T x_n + b - w^T x - b| = \frac{1}{\|w\|} \qquad (\text{Α}'.1)$$

Η προσθαφαίρεση του b μας βοήθησε να παρατηρήσουμε πως το πρώτο άθροισμα ισούται με 1, λόγω της πρώτης παραδοχής, και το δεύτερο άθροισμα δίνει 0, καθώς αποτελεί την εξίσωση του υπερεπιπέδου.

Στη συνέχεια θα προσπαθήσουμε να ορίσουμε το πρόβλημα που προσπαθούν να επιλύσουν οι μηχανές διανυσματικής στήριξης και να το φέρουμε σε τέτοια μορφή, ώστε η επίλυσή του να είναι εύκολη και αυτοματοποιημένη.

Το πρόβλημα που θέλουμε να βελτιστοποιήσουμε είναι το εξής: θέλουμε να μεγιστοποιήσουμε την απόσταση ενός οποιουδήποτε σημείου από το υπερεπίπεδο υπό τον περιορισμό ότι για το κοντινότερο σημείο, έχουμε κανονικοποιήσει ώστε να ισχύει η εξίσωση $w^T x_n = 1$. Η μαθηματική διατύπωση αυτού του προβλήματος είναι η εξής:

$$\text{Μεγιστοποίηση} \qquad \frac{1}{\|w\|} \qquad (\text{Α}'.2)$$

$$\text{υπό τον περιορισμό ότι} \qquad \min_{n=1,2,\ldots,N} |w^T x + b| = 1 \qquad (\text{Α}'.3)$$

Η παραπάνω διατύπωση δεν είναι φιλική προς επίλυση, κυρίως λόγω της μορφής του περιορι-

σμού, για αυτό θα την αναδιατυπώσουμε ως εξής:

$$\text{Ελαχιστοποίηση} \qquad \frac{1}{2}w^T w \tag{Α΄.4}$$

$$\text{υπό τον περιορισμό ότι} \qquad y_n(w^T x_n + b) \geq 1, n = 1, ..., N \tag{Α΄.5}$$

Πολλαπλασιαστές Lagrange

Πρόκειται για μία μέθοδο εύρεσης τοπικών μεγίστων ή ελαχίστων μιας συνάρτησης που υπακούει σε κάποιον περιορισμό ισότητας. Αν ο σκοπός μου είναι να μεγιστοποιήσω μια συνάρτηση $f(x, y)$ υπό τον περιορισμό ότι $g(x, y) = 0$, τότε αυτή η μέθοδος ορίζει και επιλύει τη συνάρτηση Lagrange $L(x, y, \lambda) = f(x, y) - \lambda g(x, y)$, εισάγοντας μια θετική μεταβλητή χαλαρότητας λ. Οι προϋποθέσεις Karush–Kuhn–Tucker [α΄], επεκτείνουν την εφαρμογή των πολλαπλασιαστών Lagrange, επιτρέποντας τη βελτιστοποίηση προβλημάτων υπό περιορισμούς σε μορφή ανισοτήτων.

[α΄] https://en.wikipedia.org/wiki/Karush%E2%80%93Kuhn%E2%80%93Tucker_conditions

Η εξίσωση Lagrange, που προκύπτει από το παραπάνω πρόβλημα με τη βοήθεια των προϋποθέσεων Karush–Kuhn–Tucker, είναι η εξής:

$$\text{Ελαχιστοποίηση} \qquad L(w, b, a) = \frac{1}{2}w^T w - \sum_{n=1}^{N} a_n(y_n(w^T x_n + b) - 1) \tag{Α΄.6}$$

όπου a είναι η θετική μεταβλητή χαλαρότητας που εισήγαγαν οι πολλαπλασιαστές Lagrange. Για να ελαχιστοποιήσω ως προς τα w και b, αρκεί να βρω τις μερικές παραγώγους και να τις μηδενίσω:

$$\text{Άρα} \qquad \nabla_w L = w - \sum_{n=1}^{N} a_n y_n x_n = 0 \tag{Α΄.7}$$

$$\text{και} \qquad \frac{\partial L}{\partial b} = \sum_{n=1}^{N} a_n y_n = 0 \tag{Α΄.8}$$

Αντικαθιστώντας στην αρχική εξίσωση, το πρόβλημα βελτιστοποίησης διατυπώνεται ως εξής:

$$\text{Ελαχιστοποίηση} \qquad L(a) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m a_n a_m x_n^T x_m \tag{Α΄.9}$$

$$\text{υπό τη συνθήκη} \qquad \sum_{n=1}^{N} a_n y_n = 0 \tag{Α΄.10}$$

$$\text{και} \qquad a_n \geq 0 \tag{Α΄.11}$$

Τετραγωνικός Προγραμματισμός

Είναι μια ειδική υποκατηγορία μαθηματικής βελτιστοποίησης, που ασχολείται με τη βελτιστοποίηση τετραγωνικών συναρτήσεων μεταβλητών που υπόκεινται σε γραμμικούς περιορισμούς. Στόχος του είναι να βρουν το n-διάστατο διάνυσμα x που ελαχιστοποιεί τη συνάρτηση $\frac{1}{2}x^T Qx + c^T x$ υπό τον περιορισμό $x \leq b$

Η λύση του παραπάνω προβλήματος δίνεται από κάποιο πακέτο τετραγωνικού περιορισμού,

όπου το Q διατυπώνεται ως εξής:

$$\begin{bmatrix} y_1 y_1 x_1^T x_1 & y_1 y_2 x_1^T x_2 & \dots & y_1 y_N x_1^T x_N \\ y_2 y_1 x_2^T x_1 & y_2 y_2 x_2^T x_2 & \dots & y_2 y_N x_2^T x_N \\ \vdots & \vdots & \ddots & \vdots \\ y_N y_1 x_N^T x_1 & y_N y_2 x_N^T x_2 & \dots & y_N y_N x_N^T x_N \end{bmatrix} \tag{A′.12}$$

Καταφέραμε να διατυπώσουμε το πρόβλημα που επιλύουν οι μηχανές διανυσματικής στήριξης σε όρους προβλήματος βελτιστοποίησης που επιλύεται σχετικά εύκολα. Πρόβλημα θα συναντήσουμε όταν το πλήθος των παρατηρήσεων N είναι τόσο μεγάλο ώστε να δίνει στον πίνακα Q απαγορευτικό μέγεθος.

Μη γραμμικά διαχωρίσιμες κλάσεις Μέχρι τώρα είδαμε πως οι αλγόριθμοι αυτοί σχηματίζουν υπερεπίπεδα, επομένω κάποιος θα μπορούσε να συμπεράνει πως λειτουργούν μόνο για γραμμικά διαχωρίσιμα προβλήματα. Ωστόσο, αν καταφέρω να μετασχηματίσω τα δεδομένα μου σε κάποιο χώρο μεγαλύτερων διαστάσεων, όπου είναι γραμμικά διαχωρίσιμα, και βρω τα διανύσματα στήριξης εκεί, τότε μπορώ με τον αντίστροφο μετασχηματισμό να βρω τα διανύσματα στήριξης στον αρχικό μου χώρο.

Έστω πως εκτελώ τον εξής μετασχηματισμό:

$$X \to Z \tag{A′.13}$$

Αν παρατηρήσω την τελική διατύπωση του προβλήματος που επιλύουν αυτοί οι αλγόριθμοι, θα δω πως η μόνη επίδραση αυτού του μετασχηματισμού είναι πως στη θέση των εσωτερικών γινομένων μεταξύ των x, πλέον πρέπει να υπολογίζω εσωτερικά γινόμενα μεταξύ των z σημείων.

Η παραπάνω διαπίστωση μπορεί με μια πρώτη ματιά να μην προκαλεί ενδιαφέρον, αποτέλεσε ωστόσο τον ακρογωνιαίο λίθο στον οποίο βασίζεται η ανωτερότητα αυτής της οικογένειας αλγορίθμων. Ας θεωρήσουμε ένα πρόβλημα ταξινόμησης, όπου τα δεδομένα είναι τόσο περίπλοκα, που προκειμένου να γίνει ο γραμμικός διαχωρισμός τους, να απαιτείται η μεταφορά τους σε κάποιο χώρο τεραστίων, δυνητικά άπειρων διαστάσεων. Εκεί που οι περισσότεροι αλγόριθμοι σηκώνουν τα χέρια ψηλά, οι μηχανές διανυσματικές στήριξης κάνουν την εξής σχεδιαστική επιλογή: αντί να μεταφέρουν τα χαρακτηριστικά σε έναν άπειρο χώρο και να επιλύσουν εκεί το πρόβλημα, ορίζουν μόνο αυτό που χρειάζονται, δηλαδή το εσωτερικό γινόμενο μεταξύ διανυσμάτων στον καινούριο χώρο. Το γινόμενο αυτό αποτελεί μία συνάρτηση που ονομάζεται πυρήνας και συμβολίζεται ως εξής:

$$K(x, x') = z \cdot z' \tag{A′.14}$$

Σε αυτό το σημείο, μπορεί να αναρωτηθεί κάποιος πώς μπορεί να ορίσει έναν πυρήνα, χωρίς να έχει αντίληψη του χώρου, στον οποίο θα μεταφερθεί. Η λογική είναι κάπως ανάποδη: αρκεί να ορίσω μια κάποια συνάρτηση και στη συνέχεια να μπορώ να αποδείξω ότι μπορεί να προκύψει ως εσωτερικό γινόμενο δύο μετασχηματισμένων διανυσμάτων. Υπάρχει μάλιστα η συνθήκη του Mercer [2], που εξασφαλίζει πως οποιαδήποτε συνάρτηση πυρήνα

$$K(x, x') \tag{A′.15}$$

είναι έγκυρη, αρκεί να είναι συμμετρική και ο πίνακας που ακολουθεί να είναι θετικά ημιορισμένος:

$$\begin{bmatrix} (x_1, x_1) & (x_1, x_2) & \dots & (x_1, x_N) \\ (x_2, x_1) & (x_2 x_2) & \dots & (x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ (x_N, x_1) & (x_N, x_2) & \dots & (x_N, x_N) \end{bmatrix} \tag{A′.16}$$

Κατά κανόνα η επιλογή του πυρήνα γίνεται από μια λίστα συχνά χρησιμοποιούμενων συναρτήσεων:

- *Πολυωνυμικός.* Δίνεται από τον τύπο:

$$K(x, x') = (x^T x' + c)^d \tag{Α΄.17}$$

όπου d είναι η διάσταση του νέου χώρου και c μία παράμετρος που καθορίζει την επιρροή που έχουν οι όροι μεγαλύτερης τάξης σε σχέση με τους όρους μικρότερης τάξης.

- *Γκαουσιανός (Radial basia function).* Δίνεται από τον τύπο:

$$K(x, x') = e^{(-\frac{\|x-x'\|^2}{2\sigma^2})} \tag{Α΄.18}$$

Ο πυρήνας αυτός μας μεταφέρει σε ένα χώρο άπειρων διαστάσεων. Ο αριθμητής του εκθέτη υπολογίζει την ευκλείδεια απόσταση μεταξύ των 2 σημείων, οπότε μπορούμε να τον αντιληφθούμε ως ένα μέτρο ομοιότητας.

Μηχανές διανυσματικής στήριξης χαλαρού περιθωρίου Η εφαρμογή μετασχηματισμού δεν αποτελεί μονόδρομο κατά την αντιμετώπιση μη-γραμμικών δεδομένων. Σε περιπτώσεις που η μη-γραμμικότητα δεν είναι έντονη συχνά επιτρέπουμε τη λανθασμένη κατηγοριοποίηση ενός μικρού μέρους των δεδομένων εξασφαλίζοντας απλούστερη υπόθεση και επομένως μειώνοντας τη πιθανότητα υπερ-προσαρμογής.

Η απαίτηση δυνατότητας λανθασμένης κατηγοριοποίησης υλοποιείται με μια ειδική κατηγορία των μηχανών διανυσματικής στήριξης: τις μηχανές χαλαρού περιθωρίου. Στους αλγόριθμους αυτούς το υπερεπίπεδο ορίζεται κανονικά, ώστε να μεγιστοποιείται το χάσμα, ωστόσο επιτρέπεται σε κάποια σημεία να το παραβιάσουν, δηλαδή να βρεθούν πέρα από τη νοητή γραμμή του περιθωρίου που ορίζεται από τα διανύσματα στήριξης της κατηγορίας τους.

Μαθηματικά, οι αλγόριθμοι αυτοί διατυπώνονται ως εξής: Η εξίσωση που ορίζει το περιθώριο εκατέρωθεν του υπερεπιπέδου διαχωρισμού $y_n(w^T x_n + b) \geq 1, n = 1, ..., N$ πλέον παραβιάζεται, οπότε εισάγουμε μια μεταβλητή χαλαρότητας, την ξ_n, ώστε :

$$y_n(w^T x_n + b) \geq 1 - \xi_n, n = 1, ..., N \tag{Α΄.19}$$

και η εξίσωση που βελτιστοποιεί πλέον ο αλγόριθμος είναι:

$$\text{Ελαχιστοποίηση} \qquad \frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n \tag{Α΄.20}$$

$$\text{υπό τον περιορισμό ότι} \qquad y_n(w^T x_n + b) \geq 1 - \xi_n, n = 1, ..., N \tag{Α΄.21}$$

$$\text{και} \qquad \xi_n \geq 0 \tag{Α΄.22}$$

Ο παράγοντας C προσδιορίζει πόσο αυστηρός είναι ο αλγόριθμος ως προς τη παραβίαση του περιθωρίου: μία μεγάλη τιμή του C δηλώνει πως επιθυμώ πολύ μικρή παραβίαση.

Β΄

NAIVE BAYES

Θεώρημα Bayes. Μία ακόμη πηγή έμπνευσης για το πρόβλημα της ταξινόμησης βρίσκεται στην επιστήμη των πιθανοτήτων. Ο αλγόριθμος Naive Bayes δίνει απάντηση στο ερώτημα:"Δεδομένων των παραδειγμάτων που έχω, ποια είναι η πιθανότερη υπόθεση;" κάνοντας χρήση του θεωρήματος Bayes, το οποίο στην περίπτωσή μας διατυπώνεται ως εξής:

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)} \tag{Β΄.1}$$

όπου h είναι η υπόθεση και d τα παραδείγματα.

- $P(h \mid d)$ Η πιθανότητα μιας υπόθεσης δεδομένων των παραδειγμάτων. Την αποκαλούμε εκ των υστέρων πιθανότητα, καθώς την υπολογίζουμε αφού έχουμε δει τα δεδομένα.
- $P(d \mid h)$ Η πιθανότητα να έχω τα παραδείγματα d, δεδομένου του ότι η υπόθεση h είναι σωστή.
- $P(h)$ Η πιθανότητα η υπόθεση h να είναι σωστή. Ονομάζεται εκ των προτέρων πιθανότητα, αφού την υπολογίζουμε βασιζόμενοι σε κάποια πεποίθηση και χωρίς κάποια γνώση για τα δεδομένα.
- $P(d)$ Η πιθανότητα των δεδομένων. Θα δούμε στη συνέχεια πως δεν χρειάζεται να ασχοληθούμε μαζί της.

Υπολογισμός Μοντέλου Η διαδικασία εφαρμογής του αλγορίθμου αυτού είναι η εξής: αρχικά, έχοντας τα χαρακτηριστικά και την κλάση κάθε παραδείγματος στο σετ εκπαίδευσης, υπολογίζουμε την πιθανότητα κάθε κλάσης, ως τη συχνότητα εμφάνισής της. Στη συνέχεια, υπολογίζουμε τις πιθανότητες κάθε τιμής ενός χαρακτηριστικού. Αν για παράδειγμα προβλέπουμε την πιθανότητα να βρέξει ($rain = yes$) με βάση την ύπαρξη σύννεφων($cloudy = yes$), τότε υπολογίζουμε:

$$P(cloudy = yes \mid rain = yes) = \frac{count(cloudy = yes, rain = yes)}{count(rain = yes)} \tag{Β΄.2}$$

Πρόβλεψη Όταν φτάσει κάποιο στοιχείο για το οποίο θέλουμε να προβλέψουμε την κλάση του, τότε χρησιμοποιούμε το Θεώρημα Bayes για να υπολογίσουμε την πιθανότητα κάθε κλάσης και να διαλέξουμε την μεγαλύτερη. Σε αυτό το σημείο παρατηρούμε πως η ποσότητα $P(d)$ στον παρονομαστή είναι σταθερή για κάθε κλάση και επομένως δεν συνεισφέρει στον υπολογισμό της πιθανότερης υπόθεσης. Άρα αρκεί να μεγιστοποιήσουμε την ποσότητα:

$$MAP(h) = P(d \mid h)P(h) \tag{Β΄.3}$$

Μερικές παρατηρήσεις σχετικά με αυτόν τον αλγόριθμο:

- ο υποτιμητικός χαρακτηρισμός του ως “απλοϊκό”, οφείλεται στην υπόθεση του θεωρήματος Bayes για στατιστική ανεξαρτησία των γεγονότων. Αν και στα περισσότερα πραγματικά προβλήματα δεν ικανοποιείται μια τέτοια απαίτηση για τα χαρακτηριστικά των δεδομένων, ο αλγόριθμος αυτός συνεχίζει να δίνει καλά αποτελέσματα, διαψεύδοντας το όνομά του.

- καθώς στον υπολογισμό κάποιων πιθανοτήτων εμπλέκεται πολλαπλασιασμός πολλών και δυνητικά μικρών πιθανοτήτων, υπάρχει ο κίνδυνος μαθηματικής υποροής (underflow) στο λογισμικό που τις εκτελεί. Για αυτό το λόγο συνηθίζουμε να δουλεύουμε με τους λογαρίθμους των πιθανοτήτων και όχι απευθείας με τις πιθανότητες.

Γ′

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η λογιστική συνάρτηση. Η συνάρτηση αυτή παίρνει τιμές από $-\infty$ μέχρι $+\infty$ και δίνει έξοδο μεταξύ 0 και 1, άρα μπορούμε να την ερμηνεύσουμε ως πιθανότητα. Δίνεται από τον τύπο:

$$\theta(s) = \frac{e^s}{1+e^s} \tag{Γ′.1}$$

Ερμηνεία Καθώς θέλουμε να κάνουμε μια πρόβλεψη για κάποιο άγνωστο χαρακτηριστικό με βάση κάποια άλλα χαρακτηριστικά-προβλέπτες που το αφορούν κινούμαστε στα εξής πλαίσια: αρχικά υπολογίζουμε πώς επηρεάζει κάθε χαρακτηριστικό-προβλέπτης την άγνωστη ποσότητα, δηλαδή του δίνουμε κάποιο βάρος. Στη συνέχεια με βάση τα χαρακτηριστικά ενός δεδομένου παίρνουμε μία τιμή για αυτό, την οποία θα μπορούσαμε να ερμηνεύσουμε ως το βαθμό που εμφανίζει το δεδομένο ως προς το χαρακτηριστικό που προβλέψουμε. Στη συνέχεια περνάμε αυτή τη τιμή από ένα κατώφλι, ώστε να δούμε πού θα την κατατάξουμε. Η μαθηματική μετάφραση της παραπάνω διαδικασίας είναι η εξής:

$$s = w^T x \rightarrow h(x) = \theta(s) = \frac{e^{w^T x}}{1+e^{w^T x}} \tag{Γ′.2}$$

Η πραγματική πιθανότητα, που προσπαθούμε να προσεγγίσουμε ορίζεται ως εξής:

$$P(y \mid x) = \begin{cases} f(x) & \text{if } y = +1 \\ 1 - f(x) & \text{if } y = -1 \end{cases} \tag{Γ′.3}$$

Αν υποθέσουμε πως η υπόθεσή μας είναι σωστή, δηλαδή $h = f$, τότε η πιθανότητα να πάρουμε έξοδο y για ένα δεδομένο με χαρακτηριστικά x είναι:

$$P(y \mid x) = \begin{cases} h(x) & \text{if } y = +1 \\ 1 - h(x) & \text{if } y = -1 \end{cases} \tag{Γ′.4}$$

Αν αντικαταστήσουμε με $h(x) = \theta(w^T x)$ και λαμβάνοντας υπόψιν πως $\theta(-s) = 1 - \theta(s)$, τότε η πιθανότητα προκύπτει:

$$P(y \mid x) = \theta(w^T x) \tag{Γ′.5}$$

Ο παραπάνω τύπος λαμβάνει υπόψιν του μόνο ένα σημείο. Αν έχω N δεδομένα στο σετ εκπαίδευσης τότε η υπόθεσή μου γίνεται:

$$\prod_{n=1}^{N} P(y \mid x) = \prod_{n=1}^{N} \theta(w^t x_n) \tag{Γ′.6}$$

Η ερώτηση την οποία οφείλουμε να απαντήσουμε τώρα είναι : "Δεδομένου του σετ εκπαίδευσης, ποιά είναι η πιθανότερη υπόθεση;" Η συνάρτηση, την οποία θέλουμε να ελαχιστοποιήσουμε, είναι η εξής:

$$Error(w) = \frac{1}{N}\sum_{n=1}^{N} \ln(1 + e^{-w^T x_n}) \tag{Γ'.7}$$

Gradient descent Πρόκειται για έναν αλγόριθμο βελτιστοποίησης που προσπαθεί να βρει το τοπικό ελάχιστο μιας κυρτής συνάρτησης. Η διαδικασία είναι επαναληπτική και σε κάθε βήμα ο αλγόριθμος επιλέγει άπληστα να κινηθεί προς την πιο απότομη κατεύθυνση, που αντιστοιχεί στην αντίθετη κατεύθυνση της κλίσης στο συγκεκριμένο σημείο.

Εκτός από την κατεύθυνση προς την οποία θα κινηθεί ο αλγόριθμος πρέπει να επιλέξει και το μέγεθος του βήματος που θα εκτελέσει. Η παράμετρος αυτή επηρεάζει τόσο την ταχύτητα εκτέλεσης του αλγορίθμου, όσο και την επιτυχία του: αν τα βήματα που κάνει είναι σταθερά και μικρά, τότε θα φτάσει εγγυημένα σε κάποιο ελάχιστο, αλλά πολύ αργά, μειώνοντας την ταχύτητα του συστήματος. Αντιθέτως αν το βήμα είναι πολύ μεγάλο, μπορεί να υπερπηδά το ελάχιστο κάθε φορά που το πλησιάζει και ο αλγόριθμος να μην συγκλίνει ποτέ. Συνήθως υιοθετούμε μια πιο σύνθετη προσέγγιση: επιλέγουμε αρχικά μεγάλο βήμα, ώστε να πλησιάσουμε γρήγορα στη λύση και το μειώνουμε μόλις φτάσουμε κοντά.

Δ΄

Κ-ΚΟΝΤΙΝΟΤΕΡΟΣ ΓΕΙΤΟΝΑΣ

Ο αλγόριθμος αυτός ανήκει στην κατηγορία των αλγορίθμων βασισμένων σε παραδείγματα (instance-based), δηλαδή οι προβλέψεις του βασίζονται εξ ολοκλήρου στα παραδείγματα (instances) και δε λαμβάνει χώρα κάποια εκπαίδευση. Οι αλγόριθμοι αυτοί είναι πολύ χρήσιμοι σε εφαρμογές που απαιτούν online μάθηση, επειδή τα δεδομένα έρχονται σειριακά και δεν είναι διαθέσιμα σε ομάδες για εκπαίδευση, όπως σε προβλέψεις μετοχών στο χρηματιστήριο.

Η ταξινόμηση ενός σημείου με άγνωστη κλάση γίνεται ως εξής: βρίσκουμε τους k κοντινότερους γείτονές του και του αναθέτουμε την κλάση της πλειοψηφίας.

Μία σημαντική παράμετρος, που εξαρτάται από το πεδίο εφαρμογής, είναι ο τρόπος με τον οποίο υπολογίζεται η απόσταση μεταξύ των σημείων. Διάφορες επιλογές είναι:

- *Ευκλείδεια απόσταση.* Πρόκειται για το συνηθέστερο τρόπο υπολογισμού απόστασης και δίνεται από τον τύπο:
$$\sqrt{\sum_{i=1}(x_i - y_i)^2} \tag{Δ΄.1}$$

- *Απόσταση Hamming.* Χρησιμοποιείται για κατηγορικά δεδομένα και κυρίως σε εφαρμογές επεξεργασίας κειμένου. Η απόσταση μεταξύ δύο παραδειγμάτων είναι το άθροισμα της απόστασης μεταξύ των χαρακτηριστικών τους, που ισούται με μηδέν για τα χαρακτηριστικά που συμπίπτουν και ένα για τα υπόλοιπα.

- *Απόσταση Manhattan.* Εμπνευσμένη από την οργάνωση του Manhattan σε οικοδομικά τετράγωνα, για τον υπολογισμό της απόστασης μεταξύ δύο σημείων μπορούμε να κινηθούμε μόνο οριζοντίως ή καθέτως. Ορίζεται ως εξής:
$$\sum_{i=1}|x_i - y_i| \tag{Δ΄.2}$$

Τέλος, η επιλογή του k οφείλει να γίνει με προσοχή. Αν είναι πολύ μεγάλο υπάρχει ο κίνδυνος κατά την ταξινόμηση να λαμβάνουμε υπόψιν πολύ μακρινά παραδείγματα, ενώ αν είναι πολύ μικρό η ταξινόμηση θα επηρεάζεται εύκολα από ενδεχόμενο θόρυβο στα δεδομένα.

Συνάρτηση Ακτινικής βάσης Η συνάρτηση αυτή σχετίζεται με πολλές έννοιες της μηχανικής μάθησης. Σε αυτό το σημείο θα ορίσουμε το βασικό της μοντέλο και θα δούμε τη λειτουργία της ως τεχνική βασισμένη σε παραδείγματα.

Η λογική του μοντέλου αυτού είναι η εξής: η υπόθεση σε ένα σημείο επηρεάζεται από την απόστασή του από κάθε παράδειγμα του σετ εκπαίδευσης. Πιο συγκεκριμένα, η μαθηματική διατύ-

πωση της υπόθεσης, που έχει και τη μορφή του σχήματος που ακολουθεί, είναι η εξής:

$$h(x) = sign(\sum_{n=1}^{N} w_n e^{-\gamma \|x - x_n\|^2}) \quad (\Delta'.3)$$

Η επιλογή της βέλτιστης υπόθεσης έγκειται σε αυτήν που προβλέπει σωστά όλα τα παραδείγματα του σετ εκπαίδευσης. Αν και συνήθως προσπαθούμε να ελαχιστοποιήσουμε το σφάλμα, εδώ είμαστε σίγουροι πως θα καταφέρουμε να το μηδενίσουμε, καθώς το μοντέλο μας έχει στη διάθεση του πάρα πολλές παραμέτρους (όσα είναι και τα παραδείγματα). Επομένως το πρόβλημα βελτιστοποίησης ορίζεται ως:

$$E_{in} = 0 \rightarrow \sum_{n=1}^{N} w_n e^{-\gamma \|x_n - x_m\|^2} = y_n \forall n \in D_N \quad (\Delta'.4)$$

όπου E_{in} είναι το σφάλμα στα παραδείγματα εκπαίδευσης και D_N το σετ εκπαίδευσης.

Ο παραπάνω τύπος δίνει ένα σύστημα N γραμμικών εξισώσεων με N αγνώστους που διατυπώνεται εύκολα ως εξής:

$$\underbrace{\begin{bmatrix} e^{-\gamma\|x_1 - x_1\|^2} & \dots & e^{-\gamma\|x_1 - x_N\|^2} \\ e^{-\gamma\|x_2 - x_1\|^2} & \dots & e^{-\gamma\|x_2 - x_N\|^2} \\ \vdots & \vdots & \vdots \\ e^{-\gamma\|x_N - x_1\|^2} & \dots & e^{-\gamma\|x_N - x_N\|^2} \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}}_{W} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{Y}$$

Η λύση αυτού του συστήματος δίνεται από τον τύπο:

$$w = \Phi^{-1} y \quad (\Delta'.5)$$

όπου ο πίνακας Φ πρέπει να είναι αντιστρέψιμος

Επίδραση της μεταβλητής γ Η μεταβλητή αυτή ορίζει πόσο απλωμένη είναι η καμπάνα γύρω από κάθε σημείο του σετ εκπαίδευσης και άρα πόσο επίδραση έχει αυτό στη γειτονιά του.

Η συνάρτηση RBF ως μοντέλο βασισμένο σε παραδείγματα. Όπως είδαμε η ταξινόμηση ενός σημείου εξαρτάται από την απόστασή του από τα υπόλοιπα του σετ εκπαίδευσης, τεχνική που παραπέμπει άμεσα στη λογική της κατηγοριοποίησης με βάση τα παραδείγματα. Μέχρι τώρα έχουμε θεωρήσει ως συνάρτηση βάσης την γκαουσιανή, αν όμως τοποθετήσουμε έναν απλό κύλινδρο γύρω από κάθε σημείο, η τεχνική αυτή ταυτίζεται με το μοντέλο k-NN.

Επιλογή K κέντρων. Η χρήση τόσων παραμέτρων όσων είναι και τα στοιχεία του σετ εκπαίδευσης κάνει τη διαδικασία της εκπαίδευσης χρονοβόρα και ενέχει κινδύνους υπερπροσαρμογής. Συνήθως λοιπόν χρησιμοποιούμε μια τροποποίηση της τεχνικής που έχουμε περιγράψει, όπου αντί να υπολογίζουμε την απόσταση από όλα τα σημεία, επιλέγουμε K αντιπροσωπευτικά

σημεία του χώρου και τους αναθέτουμε κάποια από τα σημεία του σετ εκπαίδευσης. Έτσι, σχηματίζονται ομάδες σημείων που αντιπροσωπεύονται από το κέντρο τους και απαιτείται πλέον ο καθορισμός K και όχι N παραμέτρων.

Πλέον η υπόθεση δίνεται από τον τύπο:

$$h(x) = sign(\sum_{k=1}^{K} w_n e^{-\gamma \|x-\mu_k\|^2}) \quad (\Delta'.6)$$

όπου μ_k είναι το κέντρο μιας ομάδας. Η επιλογή των βαρών w_n είναι παρόμοια: έχω N εξισώσεις και K παραμέτρους, οπότε το σύστημα λύνεται με τη χρήση του ψευδοαντίστροφου πίνακα:

$$w = \Phi^T \Phi^{-1} \Phi^T y \quad (\Delta'.7)$$

Επίλυση υπερ-ορισμένων συστημάτων με χρήση ψευδοαντίστροφου

Ένα γραμμικό σύστημα $y = Ax$ χαρακτηρίζεται ως υπερ-ορισμένο όταν έχει περισσότερες εξισώσεις από αγνώστους. Σε αυτή τη περίπτωση ο πίνακας A είναι μη τετραγωνικός και επομένως μη αντιστρέψιμος, οπότε η λύση δεν μπορεί να δοθεί ως συνήθως από $x = A^{-1}y$. Μία συνήθης λύση είναι η χρήση του ψευδοαντίστροφου Moore-Penrose [α΄], που ορίζεται ως $A^{\dagger} = (A^T A)^{-1} A^T$, ώστε να ισχύει $A^{\dagger} = I$, αλλά όχι $A^{\dagger}$. Τότε, ο πολλαπλασιασμός και των δύο μερών της εξίσωσης με $A^{\dagger}$ δεν εγγυάται ισότητα, αλλά προσέγγιση ελαχίστων τετραγώνων και η λύση είναι $x \approx A^{\dagger} y$

[α΄] https://en.wikipedia.org/wiki/Moore%E2%80%93Penrose_pseudoinverse

Το νέο πρόβλημα που αναδύεται είναι αυτό της επιλογής των βέλτιστων μ_k. Το πρόβλημα αυτό επιλύεται με την τεχνική της k-means ομαδοποίησης και διατυπώνεται ως εξής: Πρέπει να διαχωρίσουμε τα σημεία $x_1, ..., x_n$ σε k ομάδες $S_1, ..., S_k$, ώστε να ελαχιστοποιήσουμε το μέγεθος:

$$\sum_{k=1}^{K} \sum_{x_n \in S_k} \left\| x_n - \mu_k \right\|^2 \quad (\Delta'.8)$$

που δίνει το άθροισμα των αποστάσεων κάθε σημείου από το κέντρο της ομάδας στην οποία ανήκει.

Η συνηθέστερη λύση του δίνεται από τον αλγόριθμο του Lloyd [3], που εντοπίζει ένα τοπικό ελάχιστο επαναληπτικά σπάζοντας τη διαδικασία σε δύο ανεξάρτητα στάδια:

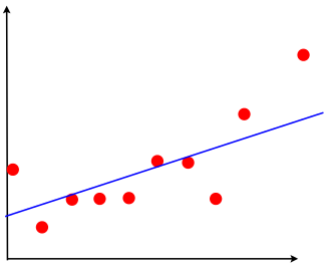
- *υπολογισμός κέντρων.* Δεδομένων των ομάδων, το κέντρο κάθε ομάδας παίρνει την μέση τιμή των σημείων που της ανήκουν.
- *υπολογισμός ομάδων.* Για κάθε σημείο του σετ εκπαίδευσης υπολογίζουμε την απόστασή του από το κέντρο κάθε ομάδας και το αναθέτουμε στην κοντινότερη ομάδα.

Η διαδικασία επαναλαμβάνεται μέχρι να συγκλίνουμε σε μια ομαδοποίηση των σημείων, δηλαδή οι ομάδες να μη μεταβάλλονται σε μια επανάληψη του αλγορίθμου.

Ε΄

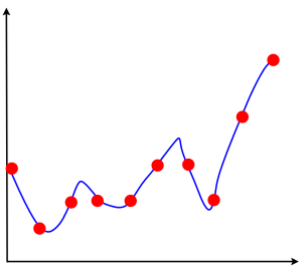
ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ

Διακύμανση - Πόλωση Μοντέλου Κατά την επιλογή του μοντέλου που θα χρησιμοποιήσουμε για μια εφαρμογή μηχανικής μάθησης, το βασικό δίλημμα μπροστά στο οποίο βρισκόμαστε είναι αυτό της πολυπλοκότητας της υπόθεσης. Το γεγονός πως προσπαθούμε να προσεγγίσουμε μία άγνωστη συνάρτηση και η αμφιβολία που νιώθουμε για την αξιοπιστία των δεδομένων μας καθιστούν τη λήψη της απόφασης ενστικτώδη και ριψοκίνδυνη. Στο παρακάτω παράδειγμα, έχουμε κάποια δισδιάστατα δεδομένα για ένα πρόβλημα παλινδρόμησης και θέλουμε να επιλέξουμε μεταξύ δύο μοντέλων:



Σχήμα Ε΄.1: Μοντέλο υψηλής πόλωσης

Η πρώτη επιλογή μας συνιστά μία ευθεία, δηλαδή ένα πολυώνυμο πρώτης τάξης, το οποίο καθορίζεται από δύο παραμέτρους. Παρατηρούμε πως το μοντέλο αυτό είναι τόσο απλό, που όσο και να προσπαθήσουμε δε θα καταφέρει να προβλέψει καλά τα δεδομένα μας. Θα ήταν ουτοπικό να χαρακτηρίζεται κάποιο πραγματικό φαινόμενο από μια τόσο απλή συνάρτηση, οπότε θα απορρίπταμε αυτό το μοντέλο ως υψηλά πολωμένο, καθώς κάνει μια σημαντική υπόθεση απλότητας.



Σχήμα Ε΄.2: Μοντέλο υψηλής διακύμανσης

Η δεύτερη επιλογή μας είναι ένα πολυώνυμο υψηλής τάξης, το οποίο μπορεί εύκολα να επιδείξει μηδενικό σφάλμα στα δεδομένα εκπαίδευσης που διαθέτουμε. Το μοντέλο αυτό είναι φαινομενικά τέλειο και αναπόφευκτα προκαλεί αμφιβολίες: είναι όλα τα δεδομένα μας τόσο αξιόπιστα και χαρακτηριστικά για το φαινόμενο που προβλέπουμε, ώστε να αξίζει να τα προσεγγίσουμε

τέλεια; Γενικά, δεδομένα πολύ υψηλής διακύμανσης είναι ύποπτα για θόρυβο, ο οποίος ως τυχαίος είναι συχνά υψίσυχνος, σε αντίθεση με φυσικά δεδομένα που υπακούν σε κάποιες συνθήκες ομαλότητας. Μήπως λοιπόν στην προσπάθειά μας να προβλέψουμε καλά την άγνωστη συνάρτηση παρασυρθήκαμε και μοντελοποιήσαμε το θόρυβο; Ένα τέτοιο μοντέλο είναι καταδικασμένο να αποτύχει σε καινούρια δεδομένα και χαρακτηρίζεται ως μοντέλο υψηλής διακύμανσης. Το παραπάνω πρόβλημα είναι αυτό που έχουμε αποκαλέσει υπερ-προσαρμογή.

Ο θόρυβος, που μας παρασύρει σε υπερ-προσαρμογή, αποτελείται από δύο συνιστώσες: το στοχαστικό θόρυβο, ο οποίος εμφανίζεται τυχαία στα δεδομένα μας και χαρακτηρίζεται από μία κατανομή $\epsilon(x)$ και τον ντετερμινιστικό. Ο τελευταίος οφείλεται στην πολυπλοκότητα της συνάρτηση-στόχου: το μοντέλο ερμηνεύει ως θόρυβο οποιαδήποτε διακύμανση δεν μπορεί να μοντελοποιήσει, καθώς είναι πολύ πολύπλοκη για αυτό, είτε αυτή γεννήθηκε τυχαία είτε προήλθε από τη συνάρτηση-στόχο.

Η μαθηματική αποτύπωση του παραπάνω προβλήματος έχει ως εξής: αν επιλέξουμε ένα μοντέλο $g(x)$ για να προβλέψουμε μια συνάρτηση $f(x)$ και ορίσουμε ως $\bar{g}(x)$ την καλύτερη δυνατή πρόβλεψη που μπορεί να κάνει το μοντέλο δεδομένων των παραμέτρων που περιέχει και της πολυπλοκότητας της $f(x)$ και $g_D(x)$ όλες τις πιθανές υποθέσεις που είναι σε θέση να κάνει το μοντέλο, ρυθμίζοντας τις παραμέτρους του, τότε το σφάλμα στο σετ εκπαίδευσης μπορεί να χαρακτηρισθεί ως εξής:

$$E_{error} = \underbrace{E_x(g_D(x) - g(x))^2}_{\text{σφάλμα λόγω διακύμανσης}} + \underbrace{E_x(\bar{g}(x) - f(x))^2}_{\text{σφάλμα λόγω πόλωσης}} + E_{\epsilon,x}(\epsilon(x^2)) \tag{Ε′.1}$$

Η ιδέα της κανονικοποίησης Η τεχνική της κανονικοποίησης στοχεύει στη μείωση της διακύμανσης με ταυτόχρονη διατήρηση χαμηλής πόλωσης. Ουσιαστικά προσπαθεί να διατηρήσει την πολυπλοκότητα της υπόθεσης, διατηρώντας το ίδιο πλήθος παραμέτρων, ώστε να έχει τη δυνατότητα να προβλέψει μια πολύπλοκη συνάρτηση-στόχο, περιορίζοντας ωστόσο την επιλογή των τιμών των παραμέτρων, ώστε να δυσκολεύεται να προβλέψει το θόρυβο.

ΣΤ΄

ΣΤΑΤΙΣΤΙΚΑ ΤΕΣΤ ΥΠΟΘΕΣΗΣ

Γενικά χαρακτηριστικά των στατιστικών τεστ υπόθεσης έχουν περιγραφεί στην Ενότητα **??**. Σε αυτό το σημείο θα αναφέρουμε περιληπτικά μερικά είδη τέτοιων τεστ, τα οποία διαφοροποιούνται κυρίως ως προς:

- τις υποθέσεις που κάνουν για τους πληθυσμούς.
- τη στατιστική που χρησιμοποιούν για να περιγράψουν το δείγμα.

Pearson's Chi-squared τεστ Πρόκειται για ένα στατιστικό τεστ μεταξύ δύο συνόλων κατηγορικών δεδομένων που εξετάζει αν οι διαφορές τους προκλήθηκαν τυχαία. Είναι κατάλληλο για μη-ζευγαρωμένα (unpaired) δεδομένα από μεγάλα δείγματα. Προέρχεται από την ευρύτερη οικογένεια των τεστ που αξιολογούνται με αναφορά στην κατανομή chi-squared, για την οποία όταν η μηδενική υπόθεση είναι αληθής η κατανομή του test statistic είναι chi-squared [1].

Yate's correction for continuity Η τεχνική αυτή χρησιμοποιείται για διόρθωση του εξής προβλήματος: κατά την εφαρμογή του Pearson's chi-squared τεστ γίνεται η υπόθεση πως η διακριτή πιθανότητα των παρατηρούμενων συχνοτήτων στον πίνακα ενδεχομένων μπορεί να προσεγγιστεί από μία συνεχή chi-squared κατανομή.

ANOVA τεστ Εισήχθη στο "Statistical methods for research workers. By Sir Ronald A. Fisher. Edinburgh (Oliver and Boyd), 12th Ed., 1954. Pp. xv, 356; 12 Figs., 74 Tables. 16s" [4] ως μία τεχνική ανάλυσης των διαφορών που παρουσιάζονται στις μέσες τιμές διαφορετικών ομάδων. Στην περίπτωση που οι ομάδες είναι ανεξάρτητες χρησιμοποιείται η one-way εκδοχή, ενώ όταν υπάρχει κάποια συσχέτιση μεταξύ τους η repeated-measures. Το τεστ αυτό χρησιμοποιείται για την περίπτωση σύγκρισης περισσότερων των τριών πληθυσμών, καθώς πολλαπλά t-tests θα οδηγούσαν σε μη αποδεκτό σφάλμα τύπου I.

Προκειμένου να ορίσει το F-statistic η τεχνική αυτή αναλύει τη διακύμανση που εμφανίζεται στο πληθυσμό σε αυτή που οφείλεται σε διαφορές μεταξύ των διαφορετικών ομάδων και διαφορές εντός των ομάδων, δηλαδή διαχωρίζει τις πηγές διακύμανσης. Οι υποθέσεις που κάνει αυτό το τεστ είναι:

- Κανονική κατανομή της εξαρτημένης μεταβλητής για κάθε ομάδα.
- Υπάρχει ομοιογένεια στις διακυμάνσεις, δηλαδή είναι ίσες για κάθε ομάδα.
- Οι παρατηρήσεις είναι ανεξάρτητες, γεγονός που καθορίζεται κατά τη συλλογή των δε-

[1] https://en.wikipedia.org/wiki/Chi-squared_test

δομένων.

Friedman τεστ Πρόκειται για ένα μη-παραμετρικό τεστ για την ανίχνευση διαφορών μεταξύ πολλών αλγορίθμων σε πολλά σετ δεδομένων. Θεωρείται μια μη-παραμετρική εκδοχή της ANOVA, με απόρροια την απεμπλοκή από τις υποθέσεις της κανονικής κατανομής, των ίσων διακυμάνσεων των residuals και την απώλεια ισχύος.

Σημαντική προσθήκη αποτελεί η εναλλακτική test statistic που εισήγαγαν οι Iman and Davenport [5], καθώς διαπίστωσαν ότι η βασική ήταν ανεπιθύμητα συντηρητική.

Σε περίπτωση διαπίστωσης σημαντικής στατιστικής διαφοράς στην απόδοση πολλών αλγορίθμων προκύπτει η ανάγκη εξακρίβωσης των ζευγαριών που οδήγησαν σε αυτό το αποτέλεσμα. Προς αυτό το σκοπό μπορούν να χρησιμοποιηθούν τα εξής post-hoc τεστ: η διαδικασία Tukey, το Dunnett τεστ, η διόρθωση Bonferroni, το τεστ Nemenyi, η προς-τα-κάτω διαδικασία του Holm, η διαδικασία του Hommel [6, 7, 8].

Fisher's exact τεστ Εισήχθη από τον Fisher μέσω του παραδείγματος της Κυρίας που δοκιμάζει Τσάι [2] ως ένα τεστ για κατηγορικά δεδομένα, τα οποία κατατάσσονται μεταξύ δύο κατηγοριών. Σύμφωνα με την ανάλυση η μηδενική υπόθεση αντιστοιχεί σε υπερ-γεωμετρική κατανομή των δεδομένων.

Χρησιμοποιείται κυρίως στην περίπτωση μικρών δειγμάτων. Λέγεται ακριβές επειδή για μικρά δείγματα η σημασία της διακύμανσης από τη μηδενική υπόθεση (p-value) μπορεί να υπολογιστεί ακριβώς αντί να βασίζεται σε μια προσέγγιση που γίνεται ακριβής καθώς το μέγεθος του δείγματος πλησιάζει το άπειρο.

Mann–Whitney U τεστ (Wilcoxon rank-sum) Εισήχθη από τον Wilcoxon [9] και αναλύθηκε διεξοδικά από τους Mann and Whitney [10]. Πρόκειται για ένα μη-παραμετρικό τεστ της μηδενικής υπόθεσης ότι είναι εξίσου πιθανό μία τυχαία επιλεγμένη τιμή από ένα δείγμα να είναι μικρότερη ή μεγαλύτερη από μία επιλεγμένη τιμή από ένα άλλο δείγμα. Σε αντίθεση με το t-test δεν απαιτεί κανονικότητα των πληθυσμών.

McNemar Εισήχθη από τον McNemar [11] ως ένα τεστ για ζευγαρωμένα ονομαστικά δεδομένα, δηλαδή δεδομένα που διαφοροποιούνται μόνο από το όνομά τους και υπάρχει ένα-προς-ένα συσχέτιση μεταξύ τους.

Cochran-Mantel-Haenszel Συνδιαμορφώθηκε από τους Cochran [12] and Mantel and Haenszel [13] και αποτελεί γενίκευση του McNemar, καθώς υποστηρίζει διαστρωμάτωση των δεδομένων σε αυθαίρετο πλήθος ομάδων.

[2] https://en.wikipedia.org/wiki/Lady_tasting_tea

Ζ΄

ΑΛΓΟΡΙΘΜΟΙ

for *n=1 μέχρι N* **do**
    βρες το x_{n+1} ως το σημείο που μεγιστοποιεί τη συνάρτηση απόκτησης ;
    υπολόγισε το αντίστοιχο y_{n+1} από τη συνάρτηση κόστους;
    ανανέωσε το στατιστικό μοντέλο με το νέο σημείο (x_{n+1}, y_{n+1});
end

Αλγόριθμος 1: Ψευδοκώδικας Bayesian Βελτιστοποίησης

H=0;
for *t=1 μέχρι T* **do**
    βρες το $\theta^* = argmin S(\theta, M_{t-1})$;
    υπολόγισε το $f(\theta^*)$;
    Ανανέωσε το σετ εκπαίδευσης $H = H \cup (\theta^*, f(\theta^*))$;
    Εκπαίδευσε ένα νέο μοντέλο M_t στο H ;
end

Αλγόριθμος 2: Ψευδοκώδικας SMBO

Η΄

ΕΞΑΓΩΓΗ ΔΙΑΣΤΗΜΑΤΩΝ ΠΡΟΒΛΕΨΗΣ ΑΠΟ ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Το διάστημα πρόβλεψης αποτελεί μια εκτίμηση για το διάστημα στο οποίο θα βρεθούν μελλοντικές παρατηρήσεις ενός πληθυσμού με μία συγκεκριμένη πιθανότητα.

Αν θεωρήσουμε μία κανονική κατανομή $\mathcal{N}(\mu, \sigma)$, τότε το διάστημα πρόβλεψης για πιθανότητα γ προκύπτει με τη βοήθεια της τυπικής κανονικής κατανομής Z, για την οποία τα τεταρτημόρια είναι προ-υπολογισμένα.

$$\gamma = P\Big(l < X < u\Big) = P\left(\frac{l-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{u-\mu}{\sigma}\right) = P\left(\frac{l-\mu}{\sigma} < Z < \frac{u-\mu}{\sigma}\right) \tag{Η΄.1}$$

Επομένως

$$\frac{l-\mu}{\sigma} = -z \qquad \text{και} \qquad \frac{u-\mu}{\sigma} = -z \tag{Η΄.2}$$

και το διάστημα πρόβλεψης ορίζεται ως:

$$[\mu - z\sigma, \mu + z\sigma] \tag{Η΄.3}$$

Από γραμμικό μοντέλο Κατά την εκπαίδευση ενός γραμμικού μοντέλου συνίσταται η επίδειξη κανονικότητας των residuals του μοντέλου, προκειμένου να είναι δυνατός ο υπολογισμός των διαστημάτων πρόβλεψης μέσω της κανονικής κατανομής.

Αν θεωρήσουμε ότι έχουμε n παραδείγματα και s_y είναι η τυπική απόκλιση των residuals του μοντέλου, η οποία ορίζεται ως:

$$\sqrt{\frac{\sum(y_i - \hat{y_i})^2}{n-2}} \tag{Η΄.4}$$

τότε το διάστημα πρόβλεψης δίνεται από τον τύπο:

$$\hat{y} \pm t^*_{n-2} s_y \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}} \tag{Η΄.5}$$

όπου ο όρος t αντιστοιχεί στο δείκτη t value, το t-statistic της μηδενικής υπόθεσης ο συντελεστής του μοντέλου παλινδρόμησης να είναι μηδενικός.

Από SVM Ένα μοντέλο παλινδρόμησης παραγόμενο με SVM διαφέρει από ένα συμβατικό γραμμικό μοντέλο ως προς τον τρόπο διαχείρισης των δεδομένων, τα οποία μετασχηματίζει σε

ένα νέο χώρο σύμφωνα με τη συνάρτηση πυρήνα. Ως αποτέλεσμα οι τεχνικές εξαγωγής διαστημάτων πρόβλεψης που περιγράψαμε δεν είναι εφαρμόσιμες.

Η βιβλιογραφία περιέχει διάφορες προσπάθειες απόδοσης πιθανοτικής εξόδου στον αλγόριθμο SVM, όπως ο αλγόριθμος του Platt [14] για μοντέλα ταξινόμησης και η μέθοδος των Jiang, Zhang, and Cai [15] και Lin and Weng [16] για παλινδρόμησης. Εμείς βασίσαμε τα πειράματά μας στη δεύτερη μέθοδο, καθώς εφαρμόζεται από την επικρατέστερη βιβλιοθήκη εκπαίδευσης SVM μοντέλων, τη LIBSVM [1], η οποία επίσης αποτελεί τη βάση της βιβλιοθήκης kernlab που χρησιμοποιήσαμε.

Η μέθοδος αυτή μοντελοποιεί τα σφάλματα των προβλέψεων (residuals) ως

$$\zeta = y - f(\hat{x}) \tag{Η′.6}$$

όπου y η κλάση μίας παρατήρησης και $\hat{f}(x)$ η πρόβλεψη για αυτήν. Στόχος της ανάλυσης είναι η εύρεση της κατανομής της τυχαίας μεταβλητής ζ, ώστε η κατανομή του y να προκύψει από τη συνέλιξη των επιμέρους κατανομών.

Όπως περιγράφουν οι Chang and Lin [17] ,ο υπολογισμός της κατανομής γίνεται παράγοντας τα σφάλματα εκτός δείγματος (out-of-sample residuals) με τη χρήση cross-validation και αναγνωρίζοντας την κατανομή που τα περιγράφει. Σύμφωνα με τα πειράματα των Lin and Weng [16] καταλληλότερη κατανομή είναι η λαπλασιανή, η οποία για τυχαία μεταβλητή z περιγράφεται από τον τύπο

$$p(z) = \frac{1}{2\sigma} e^{-\frac{|z|}{\sigma}} \tag{Η′.7}$$

όπου σ η παράμετρος κλιμάκωσης (scale parameter), η τιμή της οποία δίνεται από τη τεχνική της μέγιστης πιθανοφάνειας ως:

$$\sigma = \frac{\sum_{i=1}^{l} |\zeta_i|}{l} \tag{Η′.8}$$

όπου l το πλήθος των residuals.

Για να υπολογίσουμε το διάστημα πρόβλεψης με βεβαιότητα $1 - 2\sigma$ θα χρειαστεί να προσδιορίσουμε το σ-μόριο της κατανομής, p_s, το οποίο στη γενική περίπτωση μιας συμμετρικής μεταβλητής Z δίνεται από τον τύπο

$$\int_{-\infty}^{p_z} p(z)dz = 1 - s \tag{Η′.9}$$

που με χρήση της σχέσης Η′.7 δίνει το διάστημα

$$(\sigma \ln 2s, -\sigma \ln 2s) \tag{Η′.10}$$

Τεχνική bootstrapping Η μέθοδος αυτή περιγράφεται από τους Davison and Hinkley [18] και μπορεί να εφαρμοστεί για οποιοδήποτε μοντέλο μηχανικής μάθησης. Στόχος της είναι η συλλογή ενός δείγματος από προβλέψεις ώστε με βάση αυτό να υπολογιστούν τα τεταρτημόρια του πληθυσμού των προβλέψεων και επομένως, τα διαστήματα πρόβλεψης. Το δείγμα συνθέτεται παίρνοντας διαδοχικά bootstrap σύνολα από το σετ δεδομένων με τυχαία δειγματοληψία με αντικατάσταση και εκπαιδεύοντας ένα μοντέλο για το καθένα. Στη συνέχεια γίνεται πρόβλεψη για το παράδειγμα, η οποία εισάγεται στο δείγμα.

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvm/

Θ΄

ΣΕΤ ΔΕΔΟΜΕΝΩΝ

	Όνομα	Πηγή	Πεδίο	Τύπος	Π.Π	Π.Χ	Κλάση	NAs
1	accident [19]	kaggle	N/A	csv	2001	51	continuous	Ναι
2	ad [20]	UCI	Computer	data	3279	1558	binary	Ναι
3	adni_demographic							
4	adult [21]	UCI	social	data	48842	14	binary	Ναι
5	adult-stretch [22]	UCI	social	data	16	4	binary	Όχι
6	adult+stretch [22]	UCI	social	data	16	4	binary	Όχι
7	AP_Endometrium_Breast [23]	openml	N/A	arff	405	10937	binary	Όχι
8	AP_Endometrium_Lung [24]	openml	N/A	arff	195	10937	binary	Όχι
9	AP_Endometrium_Omentum [25]	openml	N/A	arff	138	10937	binary	Όχι
10	arsenic-male-bladder [26]	openml	N/A	arff	559	5	binary	Όχι
11	Attribute_DataSet	UCI	Computer	xlsx	501	13	binary	Ναι
12	australian [27]	UCI	Financial	690	14	dat	binary	Ναι
13	baboon_mating [28]	kaggle	N/A	csv	12141	20	binary	Όχι
14	bands [29]	UCI	Physical	data	512	39	multiclass	Όχι
15	bank-additional	UCI	csv	business	2002	20	binary	Ναι
16	bank-additional-full	UCI	csv	business	45211	20	binary	Ναι
17	biodeg [30]	UCI	N/A	csv	1055	41	binary	Όχι
18	block_1							
19	block_2							
20	block_3							
21	block_4							
22	block_5							
23	block_6							
24	block_7							
25	block_8							
26	block_10							
27	car [31]	UCI	N/A	data	1728	6	binary	Όχι
28	chess	rel.	Sport	mysql	296	19	binary	Όχι
29	chscase_health [32]	openml	N/A	arff	50	5	binary	Όχι
30	cities_r2 [33]	kaggle	N/A	csv	494	21	continuous	Όχι
31	confidence [34]	openml	N/A	arff	72	4	2	Όχι
32	creditcard [35]	kaggle	N/A	csv	284808	30	binary	Όχι
33	crx [36]	UCI	N/A	data	125	15	binary	Ναι
34	data_banknote_authentication							
35	datatest	UCI	Computer	txt	20560	7	binary	Όχι
36	datatest2	UCI	Computer	txt	20560	7	binary	Όχι
37	datatraining	UCI	Computer	txt	20560	7	binary	Όχι
38	dbworld_bodies	UCI	Computer	arff	64	4702	binary	Όχι
39	dbworld_bodies_stemmed	UCI	Computer	arff	64	4702	binary	Όχι
40	dbworld_subjects	UCI	Computer	arff	64	4702	binary	Όχι
41	dbworld_subjects_stemmed	UCI	Computer	arff	64	4702	binary	Όχι
42	dcg [37]	rel.	Synthetic	mysql	7129	3		
43	default_of_credit_card_clients [38]	UCI	Business	xls	30000	24	binary	Όχι
44	dermatology [39]	UCI	Life	data	366	33	multi	Ναι
45	diagnosis							
46	fertility_Diagnosis	UCI	Life	txt	100	10	binary	Όχι
47	ftp [40]	rel.	Synthetic	mysql	29555	2	binary	Ναι
48	gym [41]	kaggle	N/A	csv	26067	6	continuous	Όχι
49	haberman [42]	UCI	Life	data	306	3	binary	Όχι

50	heart							
51	hepatitis [43]	UCI	Life	data	155	19	binary	Ναι
52	Hill_Valley_with_noise_Training	UCI	N/A	data	606	101	binary	Όχι
53	Hill_Valley_without_noise_Training	UCI	N/A	data	606	101	binary	Όχι
54	house-votes-84	UCI	Social	data	435	16	binary	Ναι
55	HR_comma_sep [44]	kaggle	N/A	csv	15000	9	multi	Όχι
56	imdb [45]	rel.	Real	mysql	986583	5	continuous	Όχι
57	indian_ilpd	UCI	Life	csv	583	10	binary	Όχι
58	ionosphere [46]	UCI	Physical	data	351	34	binary	Όχι
59	kohkiloyeh	UCI	computer	xlsx	100	6	binary	Όχι
60	krk [47]	rel.	Synthetic	mysql	1000	6	binary	Όχι
61	lupus [48]	openml	N/A	arff	87	4	binary	Όχι
62	lymphoma_2classes							
63	magic04	UCI	Physical	data	19020	11	binary	Όχι
64	mammographic_masses	UCI	Life	data	961	6	binary	Ναι
65	messidor_features							
66	monks-1.train [49]	UCI	N/A	txt	432	7	binary	Όχι
67	monks-2.train [49]	UCI	N/A	txt	432	7	binary	Όχι
68	monks-3.train [49]	UCI	N/A	txt	432	7	binary	Όχι
69	mushrooms [50]	kaggle	N/A	csv	8125	22	binary	Ναι
70	musk [51]	rel.	Real	mysql	6599	6598	binary	Όχι
71	mutagenesis [52]	rel.	Real	mysql	5244	16	binary	Όχι
72	numeric sequence [53]	kaggle	N/A	csv	2401	28	binary	Όχι
73	nursery [54]	UCI	social	data	12960	8	multi	Όχι
74	parkinsons [55]	UCI	Life	data	197	23	binary	Όχι
75	php3BOEY5 [56]	openml	N/A	arff	745	37	binary	Όχι
76	php4ylQmK [57]	UCI	Life	csv	195	23	binary	Όχι
77	php7E9bQN	openml	N/A	arff				
78	php9xWOpn	openml	N/A	arff				
79	phphHV8xl	openml	N/A	arff				
80	phpjG28NS	openml	N/A	arff				
81	phpLalDwz	openml	N/A	arff				
82	phplN67dW	openml	N/A	arff				
83	phpqZOQcc	openml	N/A	arff				
84	phps53v4E	openml	N/A	arff				
85	phpSRnbqC [58]	openml	N/A	arff	182	12	binary	Όχι
86	phpZeLjnh	openml	N/A	arff				
87	phpjG28NS	openml	N/A	arff				
88	phpR4hXE4	openml	N/A	arff	3772	29	binary	Όχι
89	pima-indians-diabetes [59]	UCI	Life	data	768	8	binary	Ναι
90	Pokemon	data.world	N/A	csv	800	13	binary	Ναι
91	Political-media-DFE	data.world	N/A	csv	5000	22	binary	Όχι
92	prostate_TumorVSNormal	openml	N/A	arff	136	12601	binary	Όχι
93	ptc [60]	rel.	Real	mysql	18313	6	binary	Όχι
94	Qualitative_Bankruptcy [61]	UCI	Computer	arff	250	7	binary	Όχι
95	rabe_97 [62]	openml	N/A	arff	46	5	binary	Όχι
96	reviews							
97	SalesKaggle3 [63]	kaggle	N/A	csv	198918	14	continuous	Όχι
98	seismic-bumps [64]	UCI	N/A	arff	2584	19	binary	Όχι
99	shuttle-landing-control [65]	UCI	Physical	data	15	6	binary	Όχι
100	sonar	UCI	Physical	data	208	60	binary	Όχι
101	spambase [66]	UCI	Computer	data	4601	57	binary	Ναι
102	SPECTF [67]	UCI	Life	data	267	44	binary	Όχι
103	student-mat [68]	kaggle	N/A	csv	396	32	multi	Όχι
104	testing [69]	UCI	Life	csv	500	5	binary	Όχι
105	ThoracricSurgery [70]	UCI	Life	arff	470	16	binary	Όχι
106	tic-tac-toe [71]	UCI	Game	data	958	9	binary	Όχι
107	trains [72]	rel.	Synthetic	mysql	64	7	binary	Όχι
108	training [69]	UCI	Life	csv	4339	5	binary	Όχι
109	transfusion	UCI	Business	data	748	5	binary	Όχι
110	UCI_Credit_Card	kaggle	N/A	csv	30000	25	binary	Όχι
111	university_grade [73]	rel.	Synthetic	mysql	93	5	continuous	Όχι
112	university_salary [73]	rel.	Synthetic	mysql	26	5	continuous	Όχι

113	utube [74]	rel.	Real	mysql	100001	4	continuous	Όχι
114	vgsales_EU [75]	kaggle	N/A	csv	16598	10	continuous	Όχι
115	vgsales_global [75]	kaggle	N/A	csv	16598	10	continuous	Όχι
116	vgsales_JP [75]	kaggle	N/A	csv	16598	10	continuous	Όχι
117	vgsales_NA [75]	kaggle	N/A	csv	16598	10	continuous	Όχι
118	vgsales_other [75]	kaggle	N/A	csv	16598	10	continuous	Όχι
119	Video_Games_Sales [76]	kaggle	N/A	csv	15850	15	continuous	Ναι
120	visualizing_soil [77]	openml	N/A	arff	8641	4	binary	Όχι
121	voice [78]	kaggle	N/A	csv	3168	20	binary	Όχι
122	winequality-red	UCI	business	csv	1599	11	continuous	Όχι
123	winequality-white	UCI	business	csv	4898	15	continuous	Όχι
124	world	rel.	Real	mysql	30671	15	continuous	Όχι
125	yellow-small [22]	UCI	social	data	19	4	binary	Όχι
126	yellow-small+adult-stretch [22]	UCI	social	data	19	4	binary	Όχι

Πίνακας Θ΄.1: Πληροφορίες για σετ δεδομένων

ΒΙΒΛΙΟΓΡΑΦΙΑ ΠΑΡΑΡΤΗΜΑΤΩΝ

[1] V. Vapnik and A. Lerner. “Pattern Recognition using Generalized Portrait Method”. In: *Automation and Remote Control* 24 (1963).

[2] J. Mercer. “Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations”. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 209.441-458 (1909), pp. 415–446. ISSN: 0264-3952. DOI: 10.1098/rsta.1909.0016. eprint: http://rsta.royalsocietypublishing.org/content/209/441-458/415.full.pdf. URL: http://rsta.royalsocietypublishing.org/content/209/441-458/415.

[3] S. Lloyd. “Least Squares Quantization in PCM”. In: *IEEE Trans. Inf. Theor.* 28.2 (Sept. 2006), pp. 129–137. ISSN: 0018-9448. DOI: 10.1109/TIT.1982.1056489. URL: http://dx.doi.org/10.1109/TIT.1982.1056489.

[4] “Statistical methods for research workers. By Sir Ronald A. Fisher. Edinburgh (Oliver and Boyd), 12th Ed., 1954. Pp. xv, 356; 12 Figs., 74 Tables. 16s”. In: *Quarterly Journal of the Royal Meteorological Society* 82.351 (1956), pp. 119–119. ISSN: 1477-870X. DOI: 10.1002/qj.49708235130. URL: http://dx.doi.org/10.1002/qj.49708235130.

[5] Ronald L. Iman and James M. Davenport. “Approximations of the critical region of the fbietkan statistic”. In: *Communications in Statistics - Theory and Methods* 9.6 (1980), pp. 571–595. DOI: 10.1080/03610928008827904. eprint: http://dx.doi.org/10.1080/03610928008827904. URL: http://dx.doi.org/10.1080/03610928008827904.

[6] James Jaccard, Michael A. Becker, and Gregory Wood. “Pairwise multiple comparison procedures: A review”. In: *The Psychological Bulletin* 96 (1984), pp. 589–596.

[7] S. Holm. “A simple sequentially rejective multiple test procedure”. In: *Scandinavian Journal of Statistics* 6 (1979), pp. 65–70.

[8] G. HOMMEL. “A stagewise rejective multiple test procedure based on a modified Bonferroni test”. In: *Biometrika* 75 (1988), p. 383. DOI: 10.1093/biomet/75.2.383. eprint: /oup/backfile/Content_public/Journal/biomet/75/2/10.1093/biomet/75.2.383/2/75-2-383.pdf. URL: +%20http://dx.doi.org/10.1093/biomet/75.2.383.

[9] Frank Wilcoxon. “Individual Comparisons by Ranking Methods”. In: *Biometrics Bulletin* 1.6 (Dec. 1945), pp. 80–83. ISSN: 00994987. DOI: 10.2307/3001968. URL: http://dx.doi.org/10.2307/3001968.

[10] H. B. Mann and D. R. Whitney. “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other”. In: *Ann. Math. Statist.* 18.1 (Mar. 1947), pp. 50–60. DOI: 10.1214/aoms/1177730491. URL: http://dx.doi.org/10.1214/aoms/1177730491.

[11] Quinn McNemar. “Note on the sampling error of the difference between correlated proportions or percentages”. In: *Psychometrika* 12.2 (1947), pp. 153–157. ISSN: 1860-0980. DOI: 10.1007/BF02295996. URL: http://dx.doi.org/10.1007/BF02295996.

[12] William G. Cochran. “Some Methods for Strengthening the Common χ² Tests”. In: *Biometrics* 10.4 (1954), pp. 417–451. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/3001616.

[13] Nathan Mantel and William Haenszel. “Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease”. In: *JNCI: Journal of the National Cancer Institute* 22.4 (1959), p. 719. DOI: 10.1093/jnci/22.4.719. eprint: /oup/backfile/Content_

public/Journal/jnci/22/4/10.1093/jnci/22.4.719/2/22-4-719.pdf. URL: +%20http://dx.doi.org/10.1093/jnci/22.4.719.

[14] John C. Platt. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press, 1999, pp. 61–74.

[15] Bo Jiang, Xuegong Zhang, and Tianxi Cai. "Estimating the Confidence Interval for Prediction Errors of Support Vector Machine Classifiers". In: *J. Mach. Learn. Res.* 9 (June 2008), pp. 521–540. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1390681.1390698.

[16] Chih-jen Lin and Ruby C. Weng. *Simple probabilistic predictions for support vector regression*. Tech. rep. 2004.

[17] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A Library for Support Vector Machines". In: *ACM Trans. Intell. Syst. Technol.* 2.3 (May 2011), 27:1–27:27. ISSN: 2157-6904. DOI: 10.1145/1961189.1961199. URL: http://doi.acm.org/10.1145/1961189.1961199.

[18] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. New York, NY, USA: Cambridge University Press, 2013. ISBN: 0511802846, 9780511802843.

[19] kaggle. " UK Car Accidents 2005-2015." In: (). URL: https://www.kaggle.com/silicon99/dft-accident-data.

[20] Nicholas Kushmerick. " Internet Advertisements Data Set". In: *UCI* (). URL: https://archive.ics.uci.edu/ml/datasets/internet+advertisements.

[21] M. Kohavi and B. Becker. *Census Income Data Set*. 2013. URL: http://archive.ics.uci.edu/ml/datasets/Census+Income.

[22] M. Pazzani. *Balloons Data Set*. 1996. URL: http://archive.ics.uci.edu/ml/datasets/Balloons.

[23] *AP_Endometrium_Breast*. URL: https://test.openml.org/d/1123.

[24] *AP_Prostate_Lung*. URL: https://www.openml.org/d/1155.

[25] G. Stiglic and P. Kokol. " Stability of Ranked Gene Lists in Large Microarray Analysis Studies". In: *Journal of biomedicine biotechnology*, (2010). URL: https://www.openml.org/d/1151.

[26] Joaquin Vanschore. " arsenic-male-bladder". In: (2010). URL: https://www.openml.org/d/482.

[27] (confidential). " Statlog (Australian Credit Approval) Data Set". In: *Journal of biomedicine biotechnology*, (). URL: https://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval).

[28] kaggle. " Baboon Mating and Genetic Admixture". In: (). URL: https://www.kaggle.com/dryad/baboon-mating.

[29] B. Evans. " Cylinder Bands Data Set". In: (). URL: https://archive.ics.uci.edu/ml/datasets/Cylinder+Bands.

[30] K. Mansouri et al. " Quantitative Structure - Activity Relationship models for ready biodegradability of chemicals." In: *Journal of Chemical Information and Modeling* (2013). URL: https://archive.ics.uci.edu/ml/datasets/QSAR+biodegradation.

[31] M. Bohanec. *Car Evaluation Data Set*. 2013. URL: http://archive.ics.uci.edu/ml/datasets/Car+Evaluation.

[32] Chatterjee Samprit, Mark S. Handcock, and Jeffrey S. Simonoffand John Wiley. *A Casebook for a First Course in Statistics and Data Analysis*. 1995. URL: https://www.openml.org/d/705.

[33] *Top 500 Indian Cities*. URL: https://www.kaggle.com/zed9941/top-500-indian-cities.

[34] " confidence". In: (). URL: https://www.openml.org/d/1015.

[35] *Credit Card Fraud Detection.* URL: https://www.kaggle.com/dalpozz/creditcardfraud.
[36] Quinlan. *Credit Approval Data Set.* 1987. URL: http://archive.ics.uci.edu/ml/datasets/Credit+Approval.
[37] Mathieu Bally. *DCG.* URL: https://relational.fit.cvut.cz/dataset/DCG.
[38] *Default of Credit Card Clients Dataset.* URL: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset.
[39] I. Nilsel and H. Guvenir Altay. *Dermatology Data Set.* URL: http://archive.ics.uci.edu/ml/datasets/Dermatology.
[40] Jan Motl. *FTP.* URL: https://relational.fit.cvut.cz/dataset/FTP.
[41] *Crowdedness at the campus gym.* URL: https://www.kaggle.com/nsrose7224/crowdedness-at-the-campus-gym.
[42] L Tjen-Sien. *Haberman's Survival Data Set.* URL: http://archive.ics.uci.edu/ml/datasets/Haberman%5C%27s+Survival.
[43] *Hepatitis Data Set.* URL: http://archive.ics.uci.edu/ml/datasets/Hepatitis.
[44] *Human Resources Analytics.* URL: https://www.kaggle.com/ludobenistant/hr-analytics.
[45] Janez Kranjc. *IMDb.* URL: https://relational.fit.cvut.cz/dataset/IMDb.
[46] V. Sigillito. *Ionosphere Data Set.* URL: http://archive.ics.uci.edu/ml/datasets/Ionosphere.
[47] Stephen Muggleton et al. "An Experimental Comparison of Human and Machine Learning Formalisms". In: *In Proceedings of the Sixth International Workshop on Machine Learning*. Vol. 53. 1989, pp. 113–118. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.52.9566.
[48] *lupus.* URL: https://www.openml.org/d/472.
[49] S. Thrun. *MONK's Problems Data Set.* URL: http://archive.ics.uci.edu/ml/datasets/MONK%5C%27s+Problems.
[50] *Mushroom Classification.* URL: https://www.kaggle.com/uciml/mushroom-classification.
[51] Arnaud Barragao. *Musk.* URL: https://relational.fit.cvut.cz/dataset/Musk.
[52] A. K. Debnath et al. "Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity." In: *Journal of medicinal chemistry* 34.2 (1991), pp. 786–797. ISSN: 0022-2623. DOI: 10.1021/jm00106a046.
[53] *SP1 factor binding sites on Chromosome1.* URL: https://www.kaggle.com/hobako1993/sp1-factor-binding-sites-on-chromosome1.
[54] V. Rajkovic. *Nursery Data Set.* URL: http://archive.ics.uci.edu/ml/datasets/Nursery.
[55] MA Little et al. "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection". In: *BioMedical Engineering OnLine* (2007), pp. 6–23.
[56] Hans Jesus Bauer and Deter Bergman. *PieChart2.* URL: https://www.openml.org/d/1452.
[57] Ross Quinlan. *Thyroid Disease Data Set.* URL: http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease.
[58] Rajen Bhatt. " Planning-Relax Dataset for Automatic Classification of EEG Signals." In: *UCI Machine Learning Repository* (). URL: https://archive.ics.uci.edu/ml/datasets/Planning+Relax.
[59] National Institute of Diabetes, Digestive, and Kidney Diseases. *Pima Indians Diabetes Data Set.* URL: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.

[60] Christoph Helma et al. “The Predictive Toxicology Challenge 2000-2001”. In: *Bioinformatics* 17.1 (Jan. 2001), pp. 107–108. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/17.1.107. URL: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/17.1.107.

[61] A. Martin, J. Uthayakumar, and M. Nadarajan. “ Qualitative_Bankruptcy Data Set ”. In: *UCI* (2014). URL: https://archive.ics.uci.edu/ml/datasets/qualitative%5C_bankruptcy.

[62] “ ”. In: *openml* ().

[63] *Historical Sales and Active Inventory*. URL: https://www.kaggle.com/flenderson/sales-analysis.

[64] M. Sikora and L. Wrobel. “ Application of rule induction algorithms for analysis of data collected by seismic hazard monitoring systems in coal mines”. In: *Archives of Mining Sciences* (2010). URL: https://archive.ics.uci.edu/ml/datasets/seismic-bumps.

[65] “ Shuttle Landing Control Data Set ”. In: (). URL: https://archive.ics.uci.edu/ml/datasets/Shuttle+Landing+Control.

[66] Mark Hopkins et al. *Spambase Data Set*. URL: http://archive.ics.uci.edu/ml/datasets/Spambase.

[67] “ SPECTF Heart Data Set ”. In: (). URL: https://archive.ics.uci.edu/ml/datasets/SPECTF+Heart.

[68] *Student Alcohol Consumption*. URL: https://www.kaggle.com/uciml/student-alcohol-consumption.

[69] B. Johnson, R. Tateishi, and N. Hoan. “ A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees.” In: *International Journal of Remote Sensing* (2013). URL: http://archive.ics.uci.edu/ml/datasets/Wilt.

[70] “ SPECTF Heart Data Set ”. In: (). URL: https://archive.ics.uci.edu/ml/datasets/SPECTF+Heart.

[71] David W. Aha. *Tic-Tac-Toe Endgame Data Set*. URL: http://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame.

[72] Donald Michie et al. *To the international computing community: A new east-west challenge*. Tech. rep. Oxford: Oxford University Computing laboratory, 1994.

[73] Oliver Schulte. *University*. URL: https://relational.fit.cvut.cz/dataset/University.

[74] Mathieu Bally. *Utube*. URL: https://relational.fit.cvut.cz/dataset/UTube.

[75] *Video Game Sales*. URL: https://www.kaggle.com/gregorut/videogamesales.

[76] *Video Game Sales*. URL: https://www.kaggle.com/gregorut/videogamesales.

[77] “ visualizing_soil ”. In: (). URL: https://www.openml.org/d/923.

[78] *Gender Recognition by Voice*. URL: https://www.kaggle.com/primaryobjects/voicegender.