Thesis Title

Institution Name

Νησιωτη Ελένη

Day Month Year

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Περίληψη

Abstract

Ευχαριστίες

Κεφάλαιο 1

Εισαγωγή

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 της και την κοινωνία με τη δυνατότητα, ή απειλή, 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 βελτιστοποίησης προβλημάτων με ιδιαιτερότητες, η επιτυχής επίλυση των οποίων ενδεχομένως επιφέρει εμπορικά, ανθρωπιστικά ή άλλου είδους οφέλη, που όμως δεν μπορούν να επεκταθούν πέρα από την προφανή επίλυση του προβλήματος. 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 άλλα προβλήματα 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 [16, 12], , 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 εφαρμογής μηχανικών μάθησης, χρησιμοποιώντας τεχνικές που πηγάζουν από αυτήν, αλλά και άλλες επιστήμες, προσπαθεί να αντιμετωπίσει τα απαιτητικά προβλήματα, που απασχολούν την τρέχουσα αγορά, με μία

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

νέα προσέγγιση: μαθαίνοντας στις μηχανές να μαθαίνουν, όχι πλέον την επίλυση μεμονωμένων προβλημάτων, αλλά *την ίδια τη διαδικασία της μάθησης.*

1.2 Μεθοδολογία και στόχοι

Μία σύντομη ματιά στη βιβλιογραφία του AutoML αποκαλύπτει την επιτακτικόττα της ανάγκης σχεδιασμού και υλοποίησης εργαλείων μηχανικής μάθησης, τα οποία υποβοηθούν τον αναλυτή δεδομένων αυτοματοποιώντας χρονοβόρες διαδικασίες και επιστρατεύοντας μηχανισμούς εξαγωγής μετα-γνώσης. 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 στόχο την προσομοίωση ενός πραγματικού αναλυτή δεδομένων, εξοπλισμένου με εργαλεία αυτοματοποίησης των τεχνικών μηχανικής μάθησης που χρησιμοποιεί.

Πρώτο βήμα στη προσέγγιση του προβλήματος αποτέλεσε ο εντοπισμός των σημείων στη διαδικασία της μηχανικής μάθησης που επιδέχονται και χρήζουν αυτοματοποίησης. Αναγνωριστικά 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 αλορίθμους βελτιστοποίησης και τρόπους ενσωμάτωσης μετα-γνώσης στη διαδικασία. Το σύστημά μας υποστηρίζει και πειραματίζεται με συμβατικές (πλεγματικής αναζήτηση) και πρωτοποριακές (bayesian βελτιστοποίηση) τεχνικές βελτιστοποίησης, τις οποίες αξιολογεί μέσω στατιστικών σετ υπόθεσης.

Η σημασία της μετα-γνώσης στη μηχανική μάθηση έχει αναγνωριστεί μέσα στη γενικότερη προσπάθεια αυτοματοποίησης, καθώς έχει αναδυθεί η ανάγκη και δυνατότητα εκμετάλλευσης της εμπειρίας που δημιουργείται με την επιτυχημένη αντιμετώπιση προβλημάτων. Αναγνωρίζουμε πως η υλοποίηση ενός λογισμικού ανάλυσης δεδομένων αποτελεί σημαντική ευκαιρία εκμετάλλευσης της θεωρίας της μετα-γνώσης για την επίτευξη ενός έμπειρου, εκπαιδευόμενου και επεκτάσιμου προγράμματος. Έτσι, το σύστημά μας ενσωματώνει τη χρήση μεταχαρακτηριστικών 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 αυτής της αρχής στο χώρο της μηχανικής μάθησης απαιτεί ιδιαίτερη προσοχή κατά την αξιολόγηση του μοντέλου, καθώς ενέχει ο κίνδυνος υπερ-προσαρμογής, ένα πρόβλημα που αναλύεται στο Παράρτημα **??**.

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. Δεδομένης της απαιτητικότητας που δημιουργεί η παρουσία πολλών, ενδεχομένως ποιοτικά αμφισβητήσιμων μοντέλων, ενσωματώσαμε την τεχνική του σχηματισμού συλλογών μοντέλων με προς τα εμπρός επιλογή **??**, μία προσέγγιση πο 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. Θεωρούμε πως η παράλειψη ενσωμάτωσης ευριστικών θα στερούσε από το σύστημά μας πρακτική γνώση, απαραίτητη για τη λήψη σχεδιαστικών αποφάσεων. Έχουμε επομένως αναζητήσει και συλλέξει ευριστικές από τη βιβλιογραφία, τις οποίες ενσωματώσαμε στο λογισμικό, παραμετροποιώντας σχεδιαστικές αποφάσεις που παίρνει ο αλγόριθμος.

1.3 Διάρθρωση Κειμένου

Η εργασία αποτελείται από 6 κεφάλαια, συμπεριλαμβανομένου και του παρόντος βιβλιογραφικού.

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 το τρέχων state of the art και αναλύοντας δύο εκφάνσεις αυτής της επιστήμης, που αξιολογήσαμε ως κυρίαρχες στη βιβλιογραφία: τη βελτιστοποίηση των υπερπαραμέτρων μοντέλων μηχανικής μάθησης και τη μετα-μάθηση.

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

Το Κεφάλαιο 4 επιχειρεί να αξιολογήσει το σύστημα ως ολότητα, καθώς και μεμονωμένες τεχνικές που χρησιμοποιεί, θεωρώντας τες ως υποσυστήματα. Αρχικά περιγράφουμε τη διαδικασία συλλογής των σετ δεδομένων που χρειαστήκαμε για την ανάλυση, καθώς και παραμετροποιήσεις εργαλείων που χρησιμοποιήσαμε, ώστε να είναι εφικτή η αναπαραγωγή των πειραμάτων. Στη συνέχεια 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, τοποθετούμε το σύστημα ως προς τη συνεισφορά του στη σύγχρονη βιβλιογραφία και, αφορμώμενοι από περιορισμούς, προβλήματα και ιδέες που προέκυψαν στη διάρκεια της εργασίας μας, παραθέτουμε μελλοντικές επεκτάσεις-βελτιώσεις του συστήματος.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Μηχανική Μάθηση

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* [77] 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 [52] κατέληξε στον ακόλουθο φορμαλιστικό ορισμό της μηχανικής μάθησης:

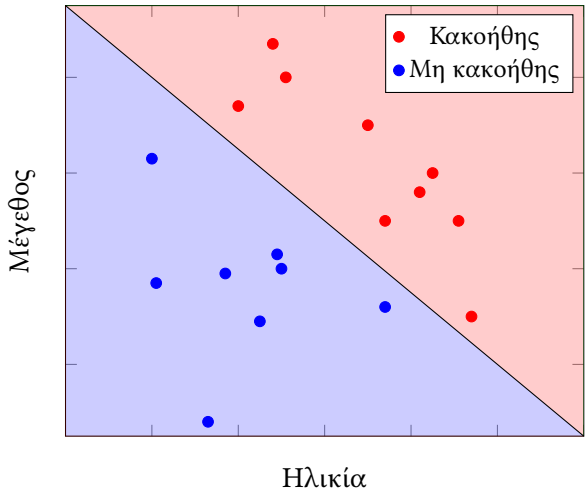
> Λέμε πως ένα πρόγραμμα υπολογιστή μαθαίνει από μια εμπειρία E, αναφερόμενοι σε ένα σύνολο καθηκόντων T και ένα μέτρο απόδοσης P, αν η απόδοση του στα καθήκοντα T, όπως μετράται από το P, βελτιώνεται καθώς αποκτά εμπειρία E.

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 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 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 Ούτε σε αυτή τη μορφή το πρόγραμμα διαθέτει πληροφορία για τη προβλεπόμενη συμπεριφορά, η προσέγγιση ωστόσο είναι διαφορετική: το πρόγραμμα δρα σε ένα



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.

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 Πρόκειται για προβλήματα πρόβλεψης μιας συνεχούς τιμής, όπως η τιμή πώλησης ακινήτων.

Προβλήματα ταξινόμησης 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.

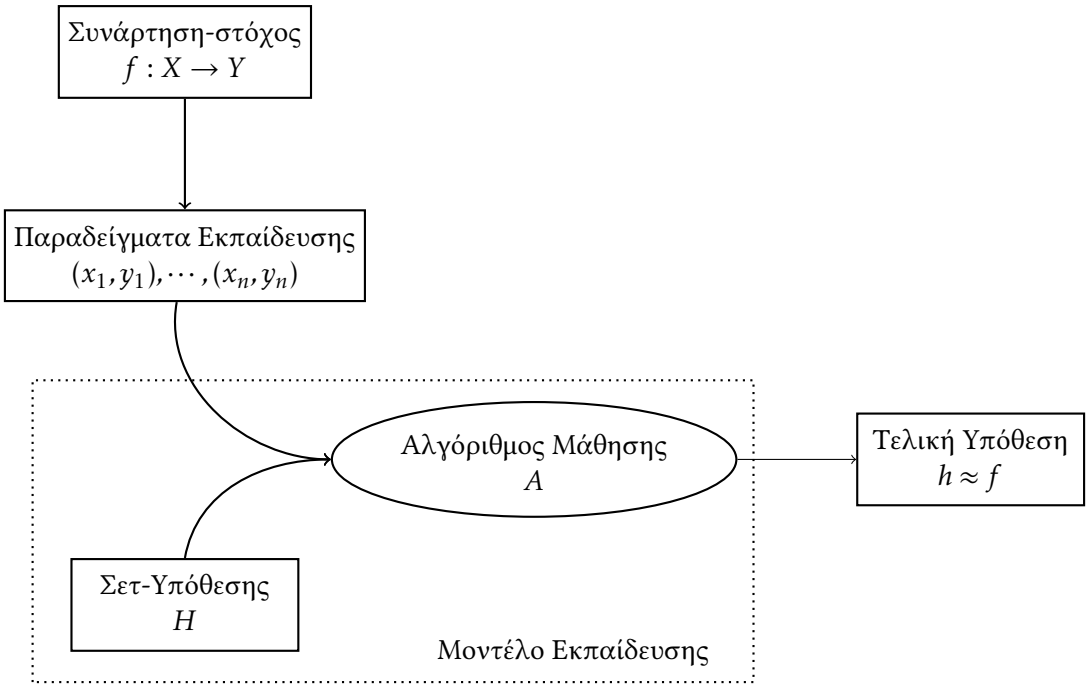
Προβλήματα ομαδοποίησης Σε αυτή τη περίπτωση γίνεται αναγνώριση ομάδων με βάση την ομοιογένειά τους, όπως στο παράδειγμα που χρησιμοποιήσαμε για τη περιγραφή της Μη Επιβλεπόμενης Μαθησης.

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:

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



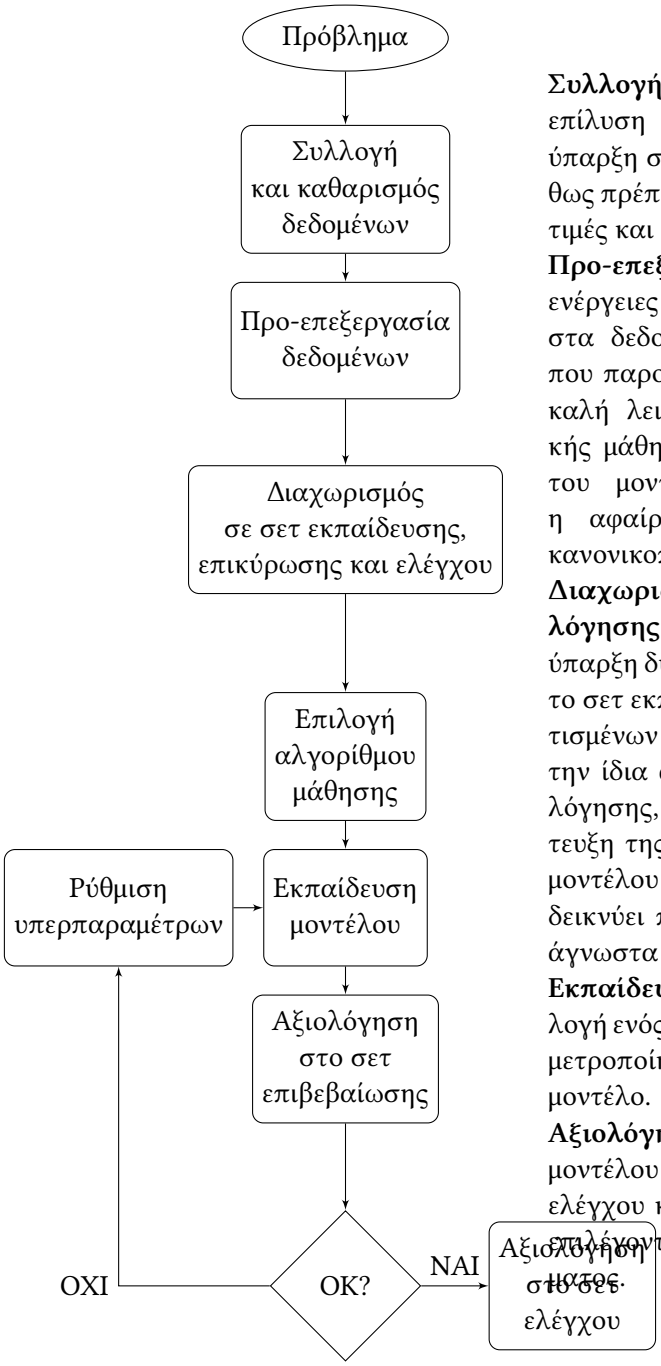
Σχήμα 2.2: Συστατικά Μηχανικής Μάθησης: Στόχος της παραγωγής ενός μοντέλου εκπαίδευσης είναι η προσέγγιση, μέσω της τελικής υπόθεσης, της συνάρτησης-στόχου , για την οποία το μοντέλο λαμβάνει πληροφορία μέσω των παραδειγμάτων. (Το σχήμα προέρχεται από τη σειρά διαλέξεων [1])

- Κλάση y_n. Πρόκειται για το στοιχείο που θέλουμε να προβλέψουμε, στη 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. 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-ANNs), οι μηχανές διανυσματικής στήριξης (Support Vector Machines-SVMs) 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 ANNs τον αλγόριθμο οπισθοδιάδοσης (backpropagation) κτλ.

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.



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

Προ-επεξεργασία 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 του 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

2.2.1 Κατά την προ-επεξεργασία

Ανάλυση κυρίαρχων συνιστωσών

Η τεχνική αυτή προέρχεται από τη γραμμική άλγεβρα και εφαρμόζεται με στόχο την εξαγωγή χρήσιμων χαρακτηριστικών σε προβλήματα μηχανικής μάθησης. 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:

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

Προκειμένου να ορίσουμε τον πίνακα P, οφείλουμε να αναλογιστούμε το σκοπό των ενεργειών μας: Τί είναι αυτό που μας ενοχλεί στην αρχική βάση και πώς ορίζουμε την καλύτερη έκφραση των δεδομένων; Σε αυτό το σημείο θυμόμαστε δύο παθογένειες των δεδομένων: το θόρυβο και τη περίσσεια πληροφορίας. Και τα δύο αυτά προβλήματα σχετίζονται άμεσα με την έννοια της αυτοσυσχέτισης: ο μεν θόρυβος είναι εξ ορισμού ασυσχέτιστος με όλα τα χαρακτηριστικά, η δε περίσσεια πληροφορίας ποσοτικοποιείται μέσω της αυτοσυσχέτισης μεταξύ των χαρακτηριστικών. Ο πίνακας αυτοσυσχέτισης 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 αυτοσυσχέτισής 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 πίνακς 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 τα κυρίαρχα συστατικά 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

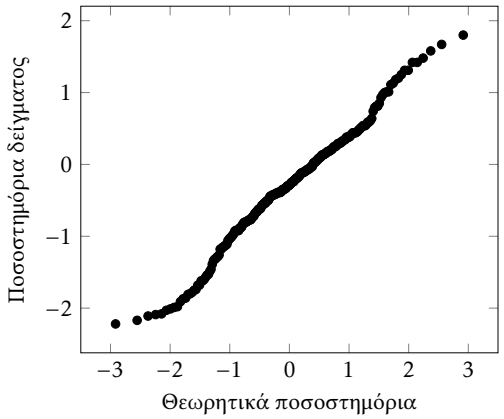
Στη διάρκεια ενός πειράματος μηχανικής μάθησης ανακύπτει συχνά η ανάγκη εξασφάλισης κανονικής κατανομής για ένα πληθυσμό. Παραδείγματος χάριν κατά την εκπαίδευση μοντέλων παλινδρόμησης η κανονικότητα των υπολειπόμενων τιμών (residuals) αποτελεί προϋπόθεση εγκυρότητας του μοντέλου. Την ανάγκη αυτή ικανοποιεί η οικογένεια των μετασχηματισμών ισχύος (power transformations), ένας εκ των οποίων είναι ο μετασχηματισμός Box-Cox. Εισήχθη από τους Box and Cox [11] και ορίζεται ως

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

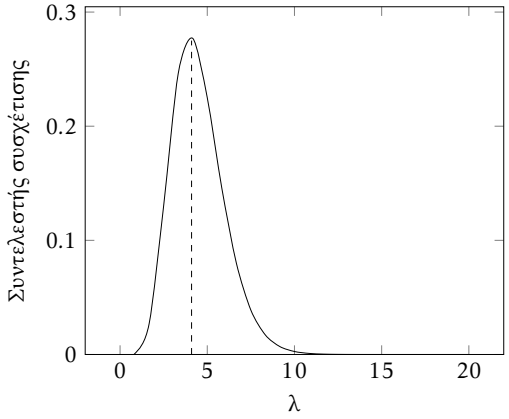
Προκειμένου να επιλεχθεί το λ, το οποίο οδηγεί στη βέλτιστα κανονική κατανομή γίνεται χρήση της ιδιότητας των Q-Q (Quartile-Quartile) διαγραμμάτων να απεικονίζουν την κανονικότητα ενός πληθυσμού. Πρόκειται για διαγράμματα διασποράς των σημείων:

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

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.



Σχήμα 2.3: 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 2.4: Η επιλογή του λ που βελτιστοποιεί ο συντελεστή συσχέτισης του διαγράμματος Q-Q συνεπάγεται το μετασχηματισμό στη βέλτιστα κανονική κατανομή.

2.2.2 Κατά την 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 (Support Vector Machines), ο απλοϊκός bayesian ταξινομητής (Naive Bayes), Λογιστική Παλινδρόμηση (Logistic Regression), οι οποίοι αναλύονται στα αντίστοιχα παραρτήματα. Μία τεχνική που ενσωματώνεται σε έναν αλγόριθμο μάθησης προκειμένου να αποφευχθεί το πρόβλημα της υπερ-προσαρμογής είναι αυτή της κανονικοποίησης (Παράρτημα Α΄). 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.

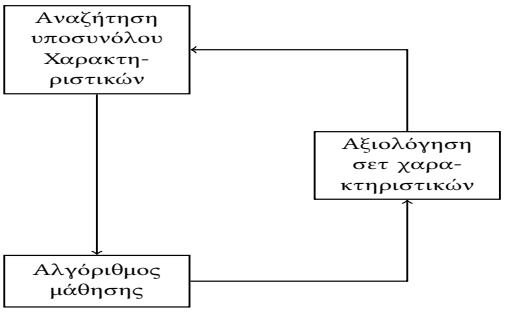
Επιλογή χαρακτηριστικών

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 σίγουρα στην πρόβλεψη. Οι μέθοδοι που το επιχειρούν αυτό λέγονται wrapper, επειδή εμπλέκουν τη διαδικασία της εκπαίδευσης και είναι πιο αποδοτικοί από τις μεθόδους επιλογής χαρακτηριστικών που είδαμε κατά την προ-επεξεργασία, αποκαλούμενες μέθοδοι φιλτραρίσματος, καθώς λαμβάνουν αποφάσεις πολύ πιο συνειδητά.

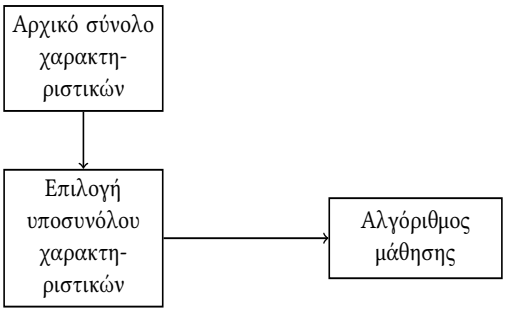
Στο παραπάνω σχήμα βλέπουμε πως οι μέθοδοι αυτοί διαλέγουν διαδοχικά ένα υποσύνολο των χαρακτηριστικών, εκπαιδεύουν ένα μοντέλο, το αξιολογούν και στη συνέχεια παραδίδουν το μοντέλο με την καλύτερη απόδοση και τα χαρακτηριστικά που επέλεξαν. Είναι σημαντικό πως κατά την επαναληπτική διαδικασία της επιλογής 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:

- *προς τα εμπρός επιλογή*. Ξεκινά με ένα άδειο σύνολο και προσθέτει διαδοχικά χαρακτηρι-



Σχήμα 2.5: 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 2.6: Μέθοδος φιλτραρίσματος για την επιλογή χαρακτηριστικών: γίνεται επιλογή ενός υποσυνόλου με βάση κάποιες ιδιότητες των χαρακτηριστικών, όπως η συσχέτισή τους στην Ανάλυση Κυρίαρχων Συνιστωσών.

στικά 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 υπερπροσαρμογή.

Συνάθροιση μοντέλων

Στο κεφάλαιο αυτό θα γνωρίσουμε μια οικογένεια τεχνικών, που στοχεύουν στη βελτίωση της μηχανικής μάθησης, καθώς αποτελούν προσθετικό 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 μοντέλου.

Μία από τις βασικές αρχές της μηχανικής μάθησης αποτελεί το ”ξυράφι του Όκαμ”: Δεδομένου ενός προβλήματος, μεταξύ ανταγωνιζομένων λύσεων επιλέγουμε αυτήν που κάνει τις λιγότερες υποθέσεις. Αυτή η αρχή ερμηνεύεται ως εξής: αν έχω καταφέρει να λύσω ένα πρόβλημα με διαφορετικούς, σωστούς τρόπους, τότε θα προτιμήσω τον απλούστερο. 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 προ επεξεργασία, 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 K υποσύνολα με n στοιχεία το καθένα, δειγματοληπτώντας με αντικατάσταση. Για κάθε διαφορετικό δείγμα εκπαιδεύουμε ένα μοντέλο με έναν αλγόριθμο ομαδοποίησης, για παράδειγμα με ένα δέντρο. Όταν θέλουμε να προβλέψουμε την κλάση ενός νέου στοιχείου, χρησιμοποιούμε τα K 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 K 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: πώς μπορεί κάποιος να συγκρίνει άμεσα την απόδοση ενός μοντέλου μηχανής διανυσματικής στήριξης με ενός γραμμικής παλινδρόμησης; Μήπως κάποιο είναι καταλληλότερο και οφείλω να το εμπιστευτώ περισσότερο; Η τεχνική αυτή, που αποκαλούμε εν συντομία stacking, δίνει λύση σε αυτό το πρόβλημα εισάγοντας την έννοια του μεταμοντέλου εκπαίδευσης. Σε πρώτο στάδιο τα μοντέλα εκπαιδεύονται και παράγεται η πρόβλεψη για κάθε παράδειγμα. Το δεύτερο στάδιο, που αποτελεί το μεταμοντέλο, 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. Είναι σημαντικό τα παραδείγματα που θα χρησιμοποιηθούν για την εκπαίδευση των διαφορετικών μοντέλων να είναι διαφορετικά από αυτά για τα οποία θα γίνει πρόβλεψη πριν την είσοδο στο μετα-μοντέλο, κάτι το οποίο μπορεί να γίνει με την τεχνική holdout ή cross validation.

2.2.3 Κατά την 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. Όπως έχει συνηθίσει, 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 Η τεχνική αυτή είναι πολύ απλή: αναθέτουμε ένα μέρος των δεδομένων για εκπαίδευση και τα υπόλοιπα τα αφήνουμε στην άκρη για αξιολόγηση. Οι συνήθεις αναλογίες είναι 80% – 20% και 75% – 25% 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 και επαναλαμβάνουμε τη διαδικασία όσες φορές είναι τα δεδομένα μας αφήνοντας κάθε φορά ένα διαφορετικό για αξιολόγηση.

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 μια 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:

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():

$$\bar{E}^{(0.632)} = 0.368 \cdot \bar{err} + 0.632 \cdot \bar{E}^{(1)} \tag{2.12}$$

όπου $\bar{err}$ 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).

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 [62], ως μια ειδική περίπτωση πίνακα ενδεχομένων, 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.

	Προβλεπόμενη κλάση	
Πραγματική κλάση	TP	FN
	FP	TN

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

Μετρικές

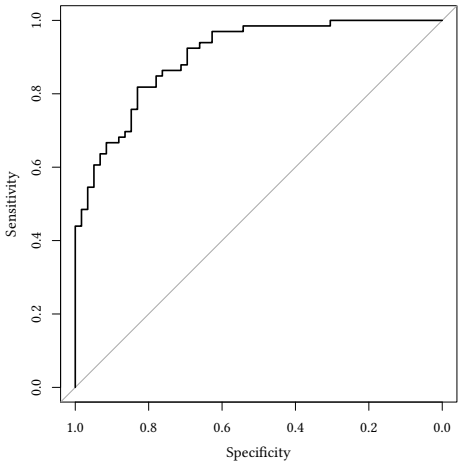
$$\text{Ακρίβεια (Accuracy)} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Ανάκληση (Recall)} = \frac{TP}{TP + FN}$$

$$\text{Ακρίβεια (Precision)} = \frac{TP}{TP + FN}$$

$$\text{F-μετρική (F-measure)} = \frac{2 * (Precision + Recall)}{Precision + Recall}$$

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 σχηματίζεται τοποθετώντας ένα σημείο για κάθε τιμή κατωφλίου με τετμημένη την Ειδικότητα (Specificity) και τεταγμένη την ευαισθησία (Sensitivity), οι οποίες ορίζονται ως



Σχήμα 2.7: Καμπύλη ROC: η μπλε γραμμή αντιστοιχεί σε τυχαίο ταξινομητή, αποτελεί δηλαδή το σημείο αναφοράς του διαγράμματος. Όσο πιο πάνω και δεξιά βρίσκεται η καμπύλη του τόσο καλύτερη η απόδοσή του, με το σημείο $(0, 1)$ να αντιστοιχεί σε τέλειο ταξινομητή.

Μετρικές

$$\text{Specificity} = \frac{TP}{TP + TN}$$

$$\text{Sensitivity} = \frac{FP}{FP + TN}$$

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 ερωτήματα που μπορούν να προκύψουν είναι: υπάρχει κάποια εξάρτηση μεταξύ της κλάσης και κάποιου κατηγορικού χαρακτηριστικού για ένα συγκεκριμένο σετ δεδομένων; Ποιος αλγόριθμος είναι γενικά καλύτερος;

Τα στατιστικά τεστ εφαρμόζονται σε πίνακες ενδεχομένων (*contigency tables*) 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. [61]

Μερικές έννοιες που σχετίζονται με τα στατιστικά τεστ είναι:

- υπόθεση. 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. 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.

- 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-value. Είναι η τιμή που οδηγεί στο αποτέλεσμα 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 μικρότερο από αυτό.
- σφάλματα τύπου I/II. Η απόρριψη μιας έγκυρης μηδενικής υπόθεσης χαρακτηρίζεται ως σφάλμα τύπου I, ενώ η αδυναμία απόρριψης μιας άκυρης σφάλμα τύπου II.
- ισχύς. 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 τα εργαλεία που προσφέρει η επιστήμη της μηχανικής μάθησης, μάλλον μας έχει δοθεί η εντύπωση πως απευθύνεται σε μια ελιτίστικη κοινωνία ειδικών, που με χρόνια εμπειρίας, εξειδικευμένη έρευνα και λίγη δόση τύχης, καταφέρνει να σχεδιάσει μοντέλα που έχουν μια κάποια χρησιμότητα στον πραγματικό κόσμο. Είναι αλήθεια πως σε κάθε στάδιο, οι σχεδιαστικές επιλογές που επηρεάζουν την απόδοση του μοντέλου φαίνονται να προέρχονται από έναν άπειρο χώρο και η καλύτερη επιλογή είναι 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 σή-

[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

μερα 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 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 με τα προιόντα 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 Υπερπαραμέτρων

Ένα βασικό στάδιο κατά την εκπαίδευση αλγορίθμων μηχανικής μάθησης είναι αυτό της επιλογής των υπερπαραμέτρων 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 $L(X^{(te)}; M)$ σε ένα δεδομένο σετ δεδομένων. Το μοντέλο κατασκευάζεται από έναν αλγόριθμο μάθησης , ο οποίος παραμετροποιείται από ένα σύνολο παραμέτρων λ.

Καταληγούμε λοιπόν στον μαθηματικό ορισμό της εύρεσης του συνόλου των υπερπαραμέτρων λ^*, που ορίζουν το βέλτιστο μοντέλο M^*

$$\lambda^* = \arg\min L(X^{(te)}; M(X^{(tr)}; \lambda)) = \arg\min_{\lambda} L(\lambda; M, X^{(tr)}, L) \tag{2.13}$$

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 2.13 μπορεί να αναδιατυπωθεί ώστε να περιλαμβάνει όλα τα M μοντέλα, καθένα εκ τω οποίων έχει διαφορετικές υπερ-παραμέτρους λ.

$$\lambda^* = \arg\min_{M^j \in M, \lambda \in \Lambda^j} L(X^{(te)}; M^j(X^{(tr)}; \lambda)) \tag{2.14}$$

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

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

Το πρόβλημα αυτό αναφέρεται ως Πρόβλημα Συνδυασμένης Επιλογής αλγορίθμου και Βελτιστοποίησης Υπερ-παραμέτρων (CASH) [46]

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 [66], κατάβαση κλήσης (gradient decent) [59], αλγόριθμοι [34] βασισμένοι σε ευριστικές [56]. Τα χαρακτηριστικά ωστόσο που αναφέρουμε προσδίδουν στη βελτιστοποίηση 2.13 ιδιαιτερότητες, που συγκεκριμενοποιούν τον κατάλληλο αλγόριθμο βελτιστοποίησης. Η Bayesian βελτιστοποίηση αποδεικνύει την καταλληλότητά της για το συγκεκριμένο πρόβλημα μέσω της παρουσίας της σε σύγχρονα εργαλεία AutoMl, όπως θα δούμε στην ενότητα **??**.

Bayesian Βελτιστοποίηση

Βελτιστοποίηση blackbox συναρτήσεων Η αναζήτηση των βέλτιστων παραμέτρων γίνεται με άξονα τη μεγιστοποίηση της γενικευμένης απόδοσης: ενός δείκτη που δηλώνει πόσο καλά δουλεύει 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 πραγματική συνάρτησης, που περιγράφει πώς προκύπτει η υπό μελέτη κλάση από τα χαρακτηριστικά. Η συνάρτησή αυτή μας είναι άγνωστη: δεν έχουμε λόγο να πιστεύουμε πως είναι γραμμική ή κυρτή, άρα η εύρεση ενός τοπικού μεγίστου δεν εξασφαλίζει και εύρεση ολικού. Επίσης, δεδομένου του ότι έχουμε να κάνουμε με πραγματικά προβλήματα, η συνάρτηση που προσπαθούμε να προσεγγίσουμε μάλλον είναι περίπλοκη. 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 δεδομένα που έχουμε 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:

$$\max_{x \in \subset R^d} f(x)$$

Ο όρος Bayesian βελτιστοποίηση εισήχθη από τους Mockus and Mockus [53] σε μια σειρά μελετών του 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:

$$P(M \mid E) \propto P(E \mid M) \cdot P(M) \tag{2.15}$$

Η παραπάνω πιθανότητα μας δίνει ένα μέσο για να απαντήσουμε στο πραγματικό ερώτημα: "Δεδομένων των παρατηρήσεων που έχω στη διάθεσή μου, ποιο μοντέλο είναι πιθανότερο να προσεγγίζει καλύτερα την πραγματική συνάρτηση;"

Η Γκαουσιανή διαδικασία ως εκ των προτέρων πιθανότητα Η εκ των προτέρων πιθανότητα αντικατοπτρίζει την πεποίθηση που έχουμε για την άγνωστη συνάρτηση. Αν πιστεύουμε για παράδειγμα πως ένα χαρακτηριστικό της είναι η ομαλότητα, αυτομάτως κάποιες συναρτήσεις γίνονται πιο πιθανές από άλλες.

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 Έχοντας κάποια στοιχεία της συνάρτησης και γνωρίζοντας την πιθανότητα ενός μοντέλου δεδομένων αυτών, πως θα επιλέξω τις επόμενες τιμές των παραμέτρων που θα δοκιμάσω; Προς αυτό το σκοπό χρησιμοποιούνται οι συναρτήσεις απόκτησης που μεγιστοποιούνται για σημεία που ενδεχομένως να μεγιστοποιούν και την άγνωστη συνάρτηση. Υπάρχουν διάφορες τεχνικές, ωστόσο η βασική αρχή επιλογής αποτελεί ένα συμβιβασμό μεταξύ "εξερεύνησης" και "εκμετάλλευσης": θέλουμε να κινηθούμε προς περιοχές για τις οποίες γνωρίζουμε ήδη πως το μοντέλο λειτουργεί καλά χωρίς ωστόσο να αποκλείουμε ανεξερεύνητες περιοχές, προκειμένου να αποφευχθεί ο κίνδυνος εμμονής σε κάποιο τοπικό μέγιστο.

- *Πιθανότητα βελτίωσης.* Μία από τις πρώτες τεχνικές, που εισήχθη από τον Kushner and Mockus [43], ήταν αυτή της επιλογής του σημείου x^+ με τη μεγαλύτερη πιθανότητα βελτίωσης:

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

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:

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

Η εισαγωγή της παραμέτρου ξ έλυσε μεν το πρόβλημα, δίνοντας ένα αντιληπτό κατώφλι στη βελτίωση που επιτυγχάνεται, αποτελεί δε μια σημαντική σχεδιαστική επιλογή, που οδηγεί εύκολα σε σφάλμα, καθώς για ακατάλληλα μικρή τιμή υπάρχει ο κίνδυνος εγκλωβισμού σε τοπικό μέγιστο, ενώ για μεγάλη τιμή η διαδικασία θα γίνει πολύ αργή.

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:

$$x_{t+1} = argminE(|f_{t+1}(x) - f(x^*)|D_{1:t})$$

Ο Mockus όρισε με πολύ πρακτικό τη συνάρτηση βελτίωσης ως εξής:

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

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:

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

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

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

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

- *Χρήση άνω ορίων εμπιστοσύνης.* Στην τεχνική αυτή είναι ξεκάθαρη η προσπάθεια συμβιβασμού 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 προσέγιση της L που εμφανίζεται στην 2.13 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) [36], Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [46], Ακολουθιακή Βελτιστοποίηση χωρίς Μοντέλο (Sequential Model-free Optimization - ΣΜΦΟ) [81].

Ακολουθιακή βελτιστοποίηση βασισμένη σε μοντέλο

Οι αλγόριθμοι βελτιστοποίησης που χρησιμοποιούν μοντέλα εκπαίδευσης εκκολάφτηκαν από τη διαπίστωση πως η συνάρτηση, $f : \Theta \rightarrow Y$ που καθορίζει πως επιδρούν οι 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 υπερπαραμέτρων και παρέκτασης (extrapolation) σε άγνωστες περιοχές του χώρου παραμέτρων. Επίσης, ποσοτικοποιούν τη σημασία κάθε υπερπαραμέτρου και των 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) Ο αλγόριθμος αυτός αποτελεί χαρακτηριστικό παράδειγμα αυτής της οικογένειας. Παρουσιάστηκε το 2011 από τους Hutter, Hoos, and Leyton-Brown [36]. Χρησιμοποιεί instance-based 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 ορίζεται ως:

$$EI(\theta)_{y^*}(\theta) = \int_{-\infty}^{y^*} (y^* - y) p(y \mid \theta) dy$$

Δέντρο εκτιμητών Parzen (TPE) Παρουσιάστηκε από τους Bergstra et al. [8]. Ενώ ο αλγόριθμος SMAC υπολογίζει απευθείας την ποσότητα $p(y \mid \theta)$ κατά την εύρεση της προσδοκώμενης βελτίωσης, εδώ θα την προσεγγίσουμε 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^*:

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

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 υπερπαραμέτρους 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 υπερπαραμέτρων που έχουν καλή απόδοση, έναντι αυτών που παρουσιάζουν φτωχή απόδοση.

Οι κατανομές $l(\theta)$ και $g(\theta)$ παρουσιάζουν μία ιεραρχική δομή, καθώς αντιπροσωπεύουν τις υπερπαραμέτρους και τη συσχέτιση μεταξύ τους. Όσο αφορά τις υπερπαραμέτρους με συνεχείς τιμές, μπορούμε να φανταστούμε πως έχουμε υπολογίσει τον Parzen εκτιμητή για κάθε μια από αυτές. Για να υπολογίσουμε την πιθανότητα ενός παραδείγματος υπερπαραμέτρων θ, ξεκινούμε από την κορυφή του δέντρου και κατευθυνόμαστε προς τα φύλλα ακολουθώντας τις υπερπαραμέτρους 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:

$$EI(I_{y_{min}}(\theta)) \propto (\gamma + \frac{g(\theta)}{l(\theta)} \cdot (1 - \gamma))^{-1}$$

δοκιμάζοντας διάφορους υποψήφιους συνδυασμούς των υπερπαραμέτρων 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.
- 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 μεταγνώσης. 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 [24] 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 (landmarking). 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 προ-επεξεργασία, η επιλογή αλγορίθμου και η ρύθμιση ενός μοντέλου. Σχετικές προσπάθειες στον τομέα της αυτοματοποιημένης μηχανικής μάθησης αποτελούν οι Feurer, Springenberg, and Hutter [26], οι οποίοι χρησιμοποιούν μετα-μάθηση για να αρχικοποιήσουν την αναζήτηση υπερ-παραμέτρων και οι Soares, Brazdil, and Kuba [69], 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 υπερπαραμέτρων, που παρέχει μία κοινή διεπαφή προς τρία σύγχρονα, αναγνωρισμένα πακέτα:

- *SMAC.*
- *Spearmint.* Γραμμένο σε python, χρησιμοποιείται για την εφαρμογή bayesian βελτιστοποίησης.
- *Hyperopt.* Γραμμένο σε python, αναλαμβάνει τη βελτιστοποίηση σε ιδιαίτερους χώρους αναζήτησης με τη χρήση τυχαία αναζήτησης και του αλγορίθμου TPE.

[5] https://azure.microsoft.com/en-us/
[6] https://github.com/automl/HPOlib

auto-sklearn [7] Μία εργαλειοθήκη αυτοματοποιημένης μηχανικής μάθησης,η οποία με βάση την python βιβλιοθήκη skiki-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 Autoweka, που εισήχθη το 2013 και αναλαμβάνει την επίλυση του προβλήματος CASH (εξίσωση 2.14) . Συνεχίζοντας την παράδοση του εργαλείου αυτού στην απλότητα χρήσης, το Autoweka 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 (SMOC, TPE).

Caret [10] Το πακέτο αυτό είναι γραμμένο σε R, που αποτελεί μία γλώσσα προγραμματισμού και ένα περιβάλλον λογισμικού εξειδικευμένο στη στατιστική. Η R είναι το κατεξοχήν εργαλείο για 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 προ επεξεργασίας δεδομένων και ρύθμισης μοντέλου, οι διεπαφές των οποίων είναι αναμενόμενα ανομοιόμορφες. Το πακέτο caret (classification and regression training) αποτελεί προσπάθεια τυποποίησης της διαδικασίας της εκπαίδευσης παρέχοντας ομοιόμορφες διεπαφές και εργαλεία για διαχωρισμό των δεδομένων, προ επεξεργασία, επιλογή χαρακτηριστικών, ρύθμιση των παραμέτρων του μοντέλου, εκτίμηση της σημασίας των χαρακτηριστικών και άλλων λειτουργιών χρήσιμων στην προσπάθεια αυτοματοποίησης της εκπαίδευσης ενός μοντέλου.

[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/

Κεφάλαιο 3

Περιγραφή Συστήματος

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

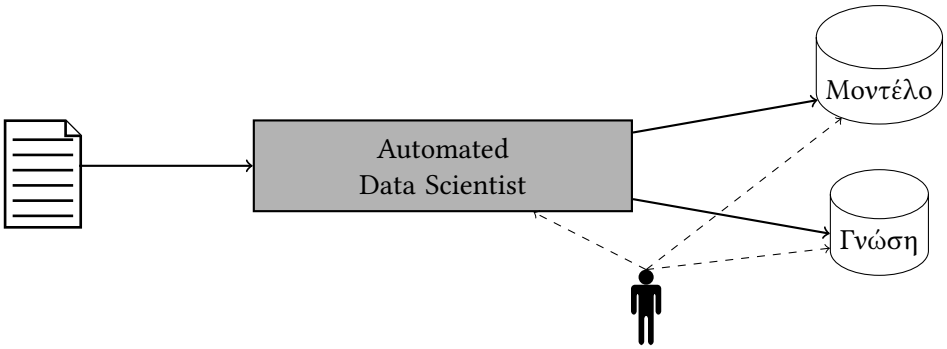
3.1 Σκοπός

Η εργασία μας 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 του state of the art στη μετα-μάθηση εφαρμόζοντας μία τεχνική πρόβλεψης υπερ-παραμέτρων για τη ρύθμιση ενός μοντέλου.
- 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.
- Ενσωματώνει την εμπειρία της κοινότητας της μηχανικής μάθησης μέσω ευριστικών και 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.
- 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 Τεχνικές

Προκειμένου να ικανοποιήσει το στόχο του το σύστημα χρησιμοποιεί διάφορες τεχνικές εφαρμογής μηχανικής μάθησης, εμπνευσμένες από τη βιβλιογραφία και προσαρμοσμένες στις ανά-



Σχήμα 3.1: Το σύστημα Automated Data Scientist ως μαύρο κουτί: Δέχεται ως είσοδο ένα σετ δεδομένων δυαδικής ταξινόμησης και ως έξοδο παράγει το βέλτιστο μοντέλο και γνώση σχετικά με το πείραμα. Ο χρήστης μπορεί να επέμβει στο πείραμα μέσω μιας σαφώς καθορισμένης διεπαφής,.

γκες του. Οι τεχνικές αυτές εξασφαλίζουν στο σύστημα αποδοτικότητα και το καθιστούν εκπαιδευόμενο, έμπειρο και επεκτάσιμο. Ένα εργαλείο AutoML θα μπορούσε να αποτελείται εξολοκλήρου από έναν άκριτο συγκερασμό μεθόδων, αποτελέσματα των οποίων θα συγκεντρώνει και παρουσιάζει περιληπτικά στον αναλυτή, αλλά ποιό το κέρδος (μαθησιακά, τελεολογικά, πρακτικά) σε μια τέτοια προσέγγιση;

3.2.1 Βελτιστοποίηση 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 εξής οφέλη:

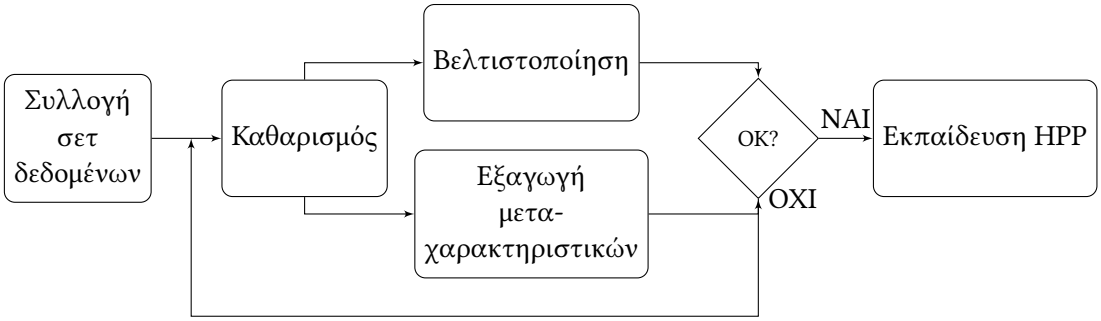
- Υπολογιστική και χρονική βελτίωση. Το πρόβλημα ης βελτιστοποίησης, όπως έχουμε δει στην ενότητα 2.3.2, απαιτεί μία επαναληπτική διαδικασία αξιολόγησης μίας κοστοβόαρς συνάρτησης ενέχοντας το κίνδυνο προσκόλλησης σε τοπικά μέγιστα, αποτελώντας το κατεξοχήν χρονοβόρο στάδιο ενός πειράματος. Αντιθέτως η πρόβλεψη των τιμών με χρήση μοντέλων μηχανικής μάθησης είναι χρονικά σύντομη και υπολογιστικά απλή.
- Ακόμη σημαντικότερο είναι το όφελος ης εκπαιδευσιμότητας της μεθόδου μας. Η χρήση του συστήματός μας σε νέα προβλήματα το καθιστά πιο έμπειρο, γεγονός που αποτελεί μελλοντική επένδυση στην προσπάθειά του. Σε αντίθεση 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 στη δουλειά της ομάδας auto-sklearn (Ενότητα 2.3.4). Στο σχήμα 3.3 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 TNN. 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 TNN. 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 3.2: 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 Καθώς η απαίτηση ακριβούς πρόβλεψης της βέλτιστης τιμής μιας υπερπαραμέτρου κρίνεται, τουλάχιστον με τα τρέχοντα χαρακτηριστικά, υπερβολικά απαιτητική, όπως σχολίασαν και οι Feurer, Springenberg, and Hutter [25], οι οποίοι αρκέστηκαν στη χρήση των προβλέψεων τους για warmstart αλγορίθμων βελτιστοποίησης, θα χρειαστεί περαιτέρω επεξεργασία του μοντέλου. Προς αυτό το σκοπό εκμεταλλεύτηκα τα διαστήματα πρόβλεψης που παράγονται από ένα μοντέλο παλινδρόμησης ώστε να ορίσω ένα σύνολο βέλτιστων υπερ-παραμέτρων για κάθε σετ δεδομένων και τελικά να δημιουργήσω έναν ensemble με αυτά. Το σύνολο αυτό ορίζεται ως οι υπερ-παράμετροι που βρίσκονται στο 90% διάστημα εμπιστοσύνης της πρόβλεψης. 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′

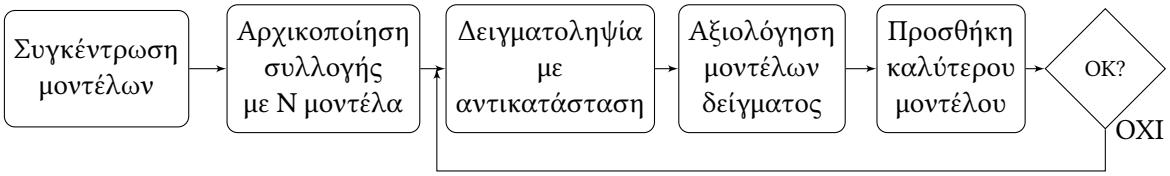
3.2.2 Ensemble με προς τα εμπρός επιλογή μοντέλων

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 [22]

> 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. [13], 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 προβλέ-

ψεων της συλλογής μετά την προσθήκη του. Στο σχήμα 3.3 μπορούμε να παρατηρήσουμε τον τρόπο με τον οποίο λειτουργεί η προτεινόμενη μέθοδος.



Σχήμα 3.3: 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.

3.2.3 Ευριστικές

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 μιας 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 TNN.
- το πλήθος των διανυσμάτων στήριξης, για ένα SVM.

Κανόνας για επιλογή μεγέθους τεστ ελέγχου Ο ίδιος καθηγητής εισάγει την ακόλουθη ευριστική. Κατά το διαχωρισμό του σετ δεδομένων σε υποσέτ για εκπαίδευση και έλεγχο πρέπει να επέλθει συμβιβασμός, καθώς είναι σημαντική και η καλή εκτίμηση της απόδοσης 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.

όπου k το πλήθος των παραδειγμάτων στο σετ ελέγχου και N το συνολικό πλήθος παραδειγμάτων.

Κανόνας Tukey για αναγνώριση ακραίων τιμών Η αναγνώριση των ακραίων τιμών σε ένα δείγμα γίνεται συνήθως οπτικά, καθώς όπως δήλωσε ο Grubbs [27] ως ακραία τιμή ορίζεται αυτή που απέχει πολύ από τις υπόλοιπες. Ο Tukey [76] 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 πληθυσμού.

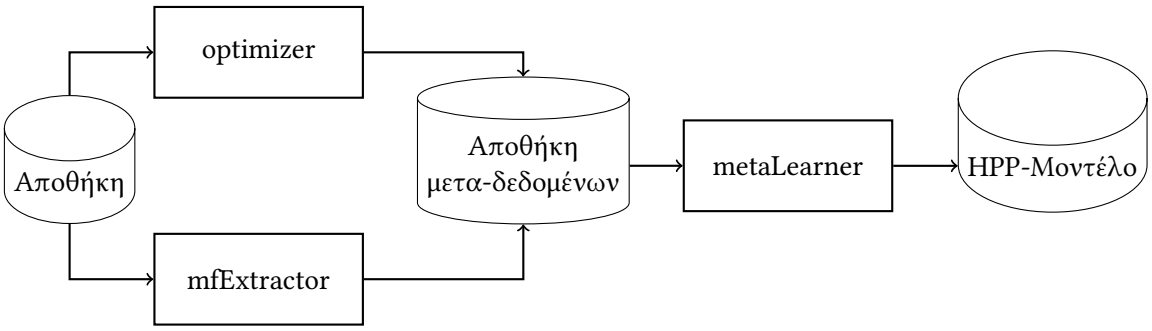
[2] http://work.caltech.edu/telecourse.html
[3] https://www.coursera.org/learn/machine-learning

3.3 Αρχιτεκτονική

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.

3.3.1 Υποσύστημα εκπαίδευσης



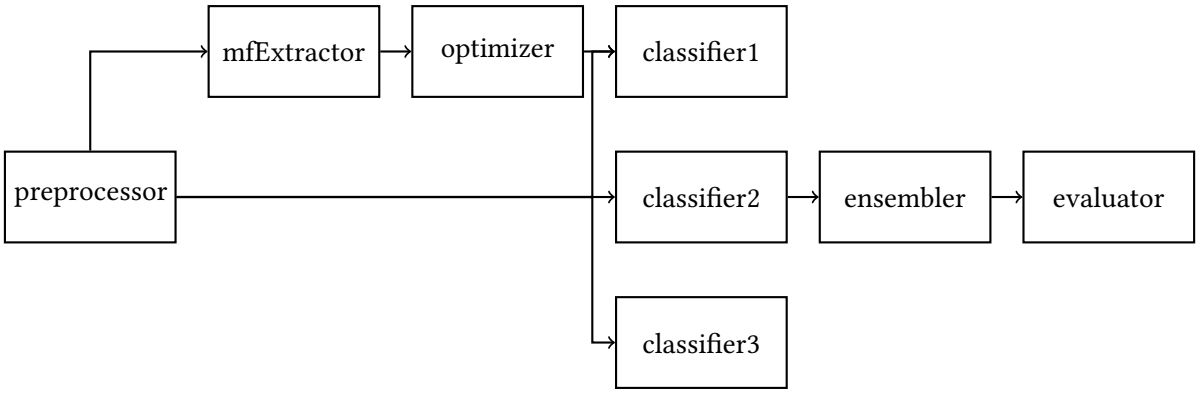
Σχήμα 3.4: Το υποσύστημα εκπαίδευσης: Για κάθε σετ δεδομένων δυαδικής ταξινόμησης της Αποθήκης γίνεται εξαγωγή των μετα-χαρακτηριστικών και εύρεση των βέλτιστων υπερ-παραμέτρων με αποτέλεσμα τη δημιουργία ενός σετ-μεταδεδομένων για κάθε υπερ-παράμετρο ενός αλγορίθμου μάθησης. Το πακέτο metaLearner αναλαμβάνει 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 ένα σετ δεδομένων στο οποίο 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.

3.3.2 Υποσύστημα πειράματος



Σχήμα 3.5: 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 του model selection.
- evaluator. Υπεύθυνο για την αξιολόγηση του τελικού ensemble με χρήση τεχνικών που είδαμε στην ενότητα **??** και τη σύγκριση της μεθόδου μας με άλλες μεθόδους αναφοράς, οι οποίες αναλύονται στην ενότητα **??** με χρήση στατιστικών τεστ.

Κεφάλαιο 4

Πειραματικά αποτελέσματα

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 3.2. Προς αυτό το σκοπό σχεδιάσαμε τα ακόλουθα πειράματα, τα οποία θα αναλύσουμε στη συνέχεια:

- αξιολόγηση των HPP μοντέλων
- αξιολόγηση του ensemble με 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.

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

Όπως είδαμε στην ενότητα 3.2.1 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 μετα-χαρακτηριστικά που χρησιμοποιήσαμε περιγράφονται στον Πίνακα 4.2, υπολογίστηκαν από το πακέτο mf-Extractor του συστήματος μας και βασίστηκαν στη δουλειά της ομάδας που υλοποίησε το λογισμικό autosklearn [1]. Καθώς τα μετα-χαρακτηριστικά που επιλέξαμε αφορούν μόνο μη-κατηγορικά χαρακτηριστικά, πριν τον υπολογισμό τους έγινε μετατροπή των κατηγορικών χαρακτηριστικών σε μεταβλητές-δείκτες.

Μετα-χαρακτηριστικά
Κλάσμα χαρακτηριστικών για 95% διακύμανση των PCA
Κυρτότητα πρώτης PCA συνιστώσας
Ασσυμετρία πρώτης PCA συνιστώσας
Ελάχιστη ασυμμετρία
Μέγιστη ασυμμετρία
Μέση τιμή ασυμμετρίας
Τυπική απόκλιση ασυμμετρίας
Ελάχιστη κυρτότητα
Μέγιστη κυρτότητα
Μέση τιμή κυρτότητας
Τυπική απόκλιση κυρτότητας

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

- Εύρεση των βέλτιστων υπερ-παραμέτρων για κάθε αλγόριθμο. Προς αυτό ο 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 **??** μπορούμε να δούμε τους αλγορίθμους μάθησης με τους οποίους ασχοληθήκαμε, καθώς και τις υπερπαραμέτρους τους.

knn	rpart	nnet	svmRadial	nb
k	cp	size decay	C sigma	fL usekernel adjust

Πίνακας 4.2: Οι αλγόριθμοι που χρησιμοποιεί το σύστημα Automated Data Scientist και οι υπερπαράμετροί τους, όπως τις ορίζει το πακέτο caret. knn: κ-κοντινότερος γείτονας, rpart: δέντρο ταξινόμησης και παλινδρόμησης (CART), nnet: TNN, svmRadial: svm με χρήση γκαουσιανού πυρήνα, nb: Naive Bayes.

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).

[1] https://github.com/automl/auto-sklearn/tree/master/autosklearn

	lm	lm+BoxCox	svmRadial
rmse			
Rsquared			
p-value			

Πίνακας 4.3

	lm	lm+BoxCox	svmRadial
rmse			
Rsquared			
p-value			

Πίνακας 4.4

Συμπεράσματα Τα μοντέλα HPP που εκπαιδεύσαμε απέχουν κατά πολύ από το να προβλέπουν επακριβώς τις υπερ-παραμέτρους. Το γεγονός αυτό μάλλον οφείλεται στα μετα-χαρακτηριστικά και συγκεκριμένα την αδυναμία τους να περιγράψουν τις συναρτήσεις-στόχους που θέσαμε. Είναι γεγονός πως δεν είμαστε βέβαιοι για την επιτευξιμότητα της πρόβλεψης υπερ-παραμέτρων, τα πειράματά μας ωστόσο δεν απορρίπτουν 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 σχεδόν πάντα μέσα στο διάστημα πρόβλεψης, προσδίδοντας βαρύτητα στην αξιολόγηση του ensemble, η οποία ακολουθεί.

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

Η αξιολόγηση της τεχνικής ensemble που χρησιμοποιήσαμε επιχειρεί να επιβεβαιώσει δύο προσδοκίες:

- Ο ensemble παρουσιάζει τουλάχιστον το ίδιο καλή απόδοση με το καλύτερο μοντέλο, το οποίο βρίσκεται στην αποθήκη βελτιστοποιημένων μοντέλων. Προς αυτό το σκοπό θα συγκρίνουμε την απόδοση του ensemble με αυτήν του εκάστοτε βέλτιστου μοντέλου με δύο τεχνικές: στατιστικά τεστ υπόθεσης και διαγράμματα προφίλ απόδοσης.
- Ο ensemble προσθέτει μοντέλα με το βέλτιστο τρόπο. Ουσιαστικά θέλουμε να επιβεβαιώσουμε τη σωστή λειτουργία του ensemble, δηλαδή ότι σε κάθε επανάληψη έχουμε είτε σταθερή είτε βελτιωμένη απόδοση.

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 βέλτιστοι 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.

Εφαρμόζωντας το Wilcoxon-rank sum τεστ με επίπεδο εμπιστοσύνης 95% διαπιστώνουμε πως ..., καθώς το p-value ισούται με ... .

Διαγράμματα προφίλ απόδοσης Τα διαγράμματα προφίλ απόδοσης (performance profile plots) [23] αποτελούν ένα εργαλείο αξιολόγησης και σύγκρισης της απόδοσης εργαλείων βελτιστοποίησης. Χρησιμοποιούνται σε περιπτώσεις εφαρμογής διαφορετικών τεχνικών βελτιστο-

ποίησης 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 ακρίβεια, p το σετ δεδομένων και s η τεχνική.

Το διάγραμμα απεικονίζει τη τιμή

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

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.

Σχήμα 4.1: Διάγραμμα προφίλ απόδοσης για τη σύγκριση του ensemble με το καλύτερο μοντέλο: Παρατηρούμε πως

Σχήμα 4.2: Διάγραμμα εξέλιξης ensemble για το σετ δεδομένων (όνομα): Παρατηρούμε πως σε κάθε επανάληψη η απόδοση του ensemble είτε μειώνεται είτε παραμένει σταθερή.

Συμπεράσματα

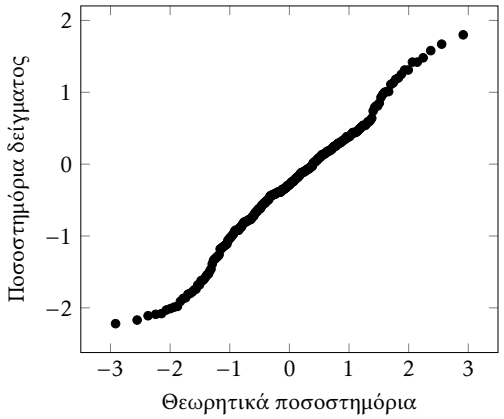
4.4 Αξιολόγηση συστήματος Automated Data Scientist

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:

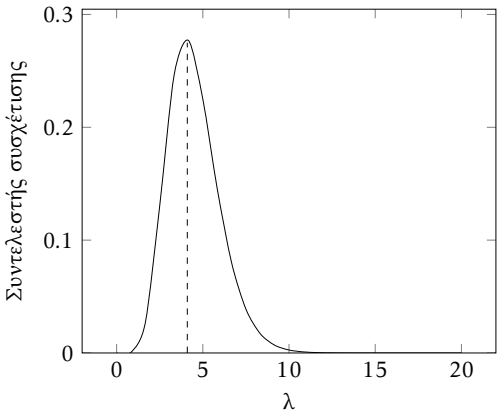
- πλεγματική αναζήτηση. Πρόκειται για τη συνηθέστερη τεχνική αναζήτησης υπερ-παραμέτρων μέχρι και σήμερα.
- Tree Parzen Estimator. Η τεχνική αυτή, που έχει περιγραφεί στην ενότητα **??** αποτελεί state of the art στο χώρο του AutoML.

Η διεξαγωγή των πειραμάτων ακολουθεί τη λογική της ενότητας 4.3. Έχουμε πραγματοποιήσει 6 διαφορετικά πειράματα: ένα για το συνολικό σύστημα και ένα χρησιμοποιώντας στον ensemble μοντέλα μόνο ενός αλγορίθμου μάθησης. Οι υποπεριπτώσεις αυτές λαμβάνονται υπόψη για τη σύγκριση των τριών τεχνικών επί ίσοις όροις. Φυσικά εμείς ενδιαφερόμαστε περισσότερο για την απόδειξη υπεροχής του συνολικού συστήματος.

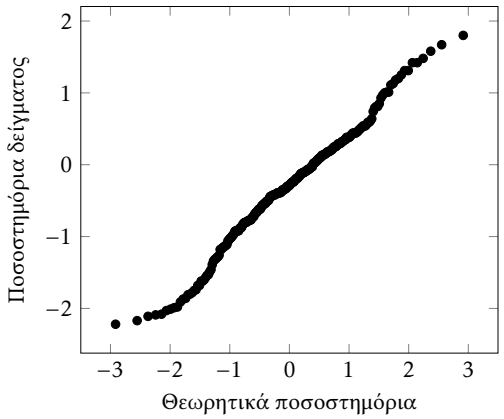
Συμπεράσματα



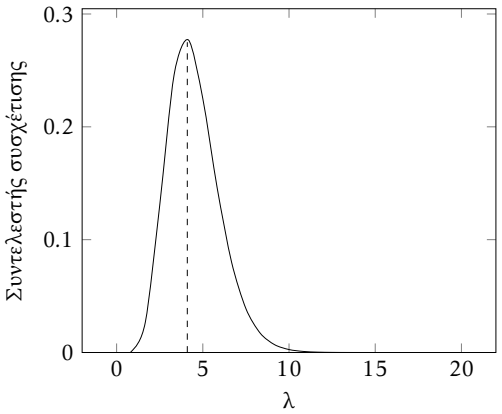
Σχήμα 4.3: Οι άξονες του διαγράμματος αντιστοιχούν στα χαρακτηριστικά του προβλήματος και τα σημεία στα παραδείγματα, για τα οποία η κλάση απεικονίζεται με το χρώμα. Η υπόθεση h αντιστοιχεί στη γραμμή, η οποία διαχωρίζει το πρόβλημα σε δύο υποχώρους.



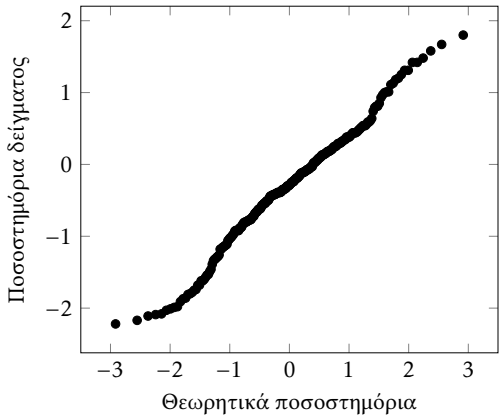
Σχήμα 4.4: Οι άξονες του διαγράμματος αντιστοιχούν στα χαρακτηριστικά του προβλήματος και τα σημεία στα παραδείγματα, για τα οποία η κλάση απεικονίζεται με το χρώμα. Η υπόθεση h αντιστοιχεί στη γραμμή, η οποία διαχωρίζει το πρόβλημα σε δύο υποχώρους.



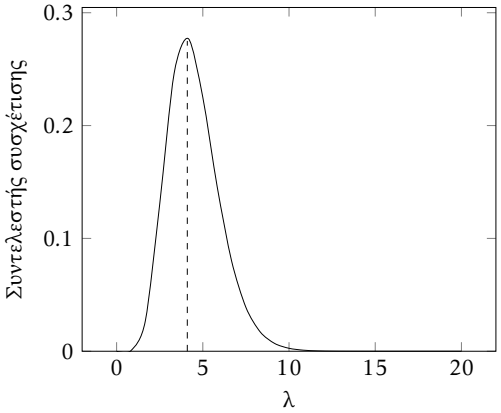
Σχήμα 4.5: Οι άξονες του διαγράμματος αντιστοιχούν στα χαρακτηριστικά του προβλήματος και τα σημεία στα παραδείγματα, για τα οποία η κλάση απεικονίζεται με το χρώμα. Η υπόθεση h αντιστοιχεί στη γραμμή, η οποία διαχωρίζει το πρόβλημα σε δύο υποχώρους.



Σχήμα 4.6: 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.



Σχήμα 4.7: 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.



Σχήμα 4.8: 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.

Κεφάλαιο 5

Σχετική Δουλειά

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 το state-of-the-art. 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 απέδειξαν πως η τεχνική της πλεγματικής αναζήτησης επιφέρει αποτέλεσμα χειρότερο από την τυχαία αναζήτηση. Έκτοτε έχουν δοκιμαστεί γενετικοί αλγόριθμοι [66], αναζήτηση κλίσης [59] και η bayesian βελτιστοποίηση, η οποία φαίνεται να έχει επικρατήσει με τη μορφή της τεχνικής SMBO [**DBLP:journals/corr/abs-1208-3719**]. Οι **HutHooLeyMur10** εισάγουν την έννοια των χρονικών ορίων στη διαδικασία της βελτιστοποίησης λαμβάνοντας υπόψιν ως κόστος τόσο την ποιότητα όσο και το χρόνο.

Προσπάθειες εισαγωγής μετα-μάθησης κατέβαλαν οι Feurer, Springenberg, and Hutter [25], 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. Η τεχνική πρόβλεψης που χρησιμοποιούν είναι

Οι **kuba2002exploiting** 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 [69] 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 κ-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 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 "Extending Metalearning to Data Mining and KDD" [24] 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:2014:MCA:3015544.3015557**. 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 υπερ-παραμέτρων και τη δυνατότητά τους να ενημερώνονται χωρίς εκπαίδευση.

Βιβλιογραφία

[1] David W. Aha. *Tic-Tac-Toe Endgame Data Set.*

[2] *AP_Endometrium_Breast.*

[3] *AP_Prostate_Lung.*

[4] Mathieu Bally. *DCG.*

[5] Mathieu Bally. *Utube.*

[6] Arnaud Barragao. *Musk.*

[7] Hans Jesus Bauer and Deter Bergman. *PieChart2.*

[8] 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.

[9] Rajen Bhatt. " Planning-Relax Dataset for Automatic Classification of EEG Signals." In: *UCI Machine Learning Repository* ().

[10] M. Bohanec. *Car Evaluation Data Set.* 2013.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] William G. Cochran. "Some Methods for Strengthening the Common χ² Tests". In: *Biometrics* 10.4 (1954), pp. 417–451. ISSN: 0006341X, 15410420.

[16] 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. 4, 1993, pp. 5–24. DOI: 10.1017/CBO9780511569944.002.

[17] *Credit Card Fraud Detection.*

[18] *Crowdedness at the campus gym.*

[19] 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.

[20] *Default of Credit Card Clients Dataset.*

[21] National Institute of Diabetes, Digestive, and Kidney Diseases. *Pima Indians Diabetes Data Set.*

[22] 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.

[23] 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.

[24] "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.

[25] 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.

[26] Matthias Feurer, Jost Springenberg, and Frank Hutter. *Initializing Bayesian Hyperparameter Optimization via Meta-Learning*. 2015.

[27] F. E Grubbs. "Procedures for detecting outlying observations in samples". In: *Technometrics, Volume 11 (1969) - Volume 1201*. 1969.

[28] 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.

[29] *Hepatitis Data Set*.

[30] *Historical Sales and Active Inventory*.

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

[32] G. HOMMEL. "A stagewise rejective multiple test procedure based on a modified Bonferroni test". In: *Biometrika* 75.2 (1988), p. 383. DOI: 10.1093/biomet/75.2.383.

[33] Mark Hopkins et al. *Spambase Data Set*.

[34] 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.

[35] *Human Resources Analytics*.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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).

[41] M. Kohavi and B. Becker. *Census Income Data Set*. 2013.

[42] Janez Kranjc. *IMDb*.

[43] 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.

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

[45] MA Little et al. "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection". In: *BioMedical Engineering OnLine* (2007), pp. 6–23.

[46] Ilya Loshchilov and Frank Hutter. "CMA-ES for Hyperparameter Optimization of Deep Neural Networks". In: *CoRR* abs/1604.07269 (2016).

[47] *lupus*.

[48] 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.

[49] 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.

[50] 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.

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

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

[53] 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.

[54] Jan Motl. *FTP*.

[55] *Mushroom Classification*.

[56] 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.

[57] I. Nilsel and H. Guvenir Altay. *Dermatology Data Set*.

[58] M. Pazzani. *Balloons Data Set*. 1996.

[59] Fabian Pedregosa. *Hyperparameter optimization with approximate gradient*. Version 1.

[60] 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.

[61] Brett Presnell. "An introduction to Categorical Data Analysis Using R". In: ().

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

[63] Quinlan. *Credit Approval Data Set*. 1987.

[64] Ross Quinlan. *Thyroid Disease Data Set*.

[65] V. Rajkovic. *Nursery Data Set*.

[66] 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.

[67] Oliver Schulte. *University*.

[68] V. Sigillito. *Ionosphere Data Set*.

[69] 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.

[70] *SP1 factor binding sites on Chromosome1*.

[71] "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.

[72] *Student Alcohol Consumption*.

[73] S. Thrun. *MONK's Problems Data Set*.

[74] L Tjen-Sien. *Haberman's Survival Data Set*.

[75] *Top 500 Indian Cities*.

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

[77] 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.

[78] *Video Game Sales*.

[79] *Video Game Sales*.

[80] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6 (Dec. 1945), pp. 80–83. ISSN: 00994987. DOI: 10.2307/3001968.

[81] 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.

Παραρτήματα

Παράρτημα Α΄

Κανονικοποίηση

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: 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. 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 υπερπροσαρμογή.

Ο θόρυβος, που μας παρασύρει σε υπερπροσαρμογή, αποτελείται από δύο συνιστώσες: το στοχαστικό θόρυβο, ο οποίος κάθεται τυχαία στα δεδομένα μας και χαρακτηρίζεται από μία κατανομή $\epsilon(x)$ και τον ντετερμινιστικό. Ο τελευταίος οφείλεται στην πολυπλοκότητα της συνάρτησης-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:

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

Η ιδέα της κανονικοποίησης Η τεχνική της κανονικοποίησης στοχεύει στη μείωση της διακύμανσης με ταυτόχρονη διατήρηση χαμηλής πόλωσης. Με απλά λόγια, 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.

Στη συνέχεια θα δούμε πώς εφαρμόζεται η κανονικοποίηση σε ένα πρόβλημα λογιστικής παλινδρόμησης με πολυωνυμικό μοντέλο και θα κατανοήσουμε την επίλυση γραφικά.

Όπως έχουμε δει στο κεφάλαιο της λογιστικής παλινδρόμησης, η συνάρτηση λάθους που προσπαθούμε να ελαχιστοποιήσουμε ώστε να βρούμε τη βέλτιστη υπόθεση είναι η εξής:

$$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}(w^T z_n - y_n)^2$$

όπου w είναι οι συντελεστές του μοντέλου, y το διάνυσμα κλάσης και x τα διανύσματα εισόδου των N παραδειγμάτων. Η λύση της παραπάνω εξίσωσης είναι:

$$w_{in} = (Z^T Z^{-1} y)$$

Αν ορίσουμε την προσπάθειά μας να περιορίσουμε τις παραμέτρους ως τοποθέτηση ενός άνω φράγματος στο μέτρο των συντελεστών, τότε το αναδιατυπωμένο, κανονικοποιημένο πλέον, πρόβλημα έχει ως εξής:

$$\text{Ελαχιστοποίηση} \qquad E_{in}(w) = \frac{1}{N}(Zw - y)^T(Zw - y)$$

$$\text{Υπό τον περιορισμό} \qquad w^T w \leq C$$

Το παραπάνω πρόβλημα καθιστά πρόβλημα βελτιστοποίησης με περιορισμό που περιέχει ανισότητα, επομένως μπορεί να επιλυθεί με πολλαπλασιαστές Lagrane και τη βοήθεια της θεωρίας Karush-Kuhn- Tucker. Ωστόσο η γραφική αναπαράσταση του προβλήματος ενδεχομένως να είναι ενδιαφέρουσα, καθώς θα δώσει καλύτερη κατανόηση της φύσης του.

Όπως έχουμε δει στο κεφάλαιο της λογιστικής παλινδρόμησης, η εξίσωση του σφάλματος έχει τη μορφή ελλειψοειδούς καμπύλης. Ο περιορισμός που θέσαμε ορίζει έναν κύκλο στον οποίο κινούνται τα επιτρεπτά w, με ακτίνα C. Η πληροφορία της κατεύθυνσης προς την οποία πρέπει να κινηθούμε, ώστε να μειώσουμε το E_{in} δίνεται ως γνωστόν από την κλίση, ∇E_{in} στο σημείο που βρισκόμαστε και φαίνεται στο παραπάνω σχήμα. Όσο αφορά τα w, είναι διανύσματα που κινούνται πιο κοντά γίνεται στη βέλτιστη λύση w_{lin}, παραμένοντας βέβαια στον κύκλο. Στο σχήμα φαίνεται πως στην καλύτερη περίπτωση βρίσκονται πάνω στην περιφέρειά του. Παρατηρώ πως, επειδή μεταξύ των δύο διανυσμάτων υπάρχει γωνία, καθώς κινούμαι στον κύκλο, το σφάλμα θα αυξομειώνεται και επομένως δε θα το βελτιστοποιώ. Αντιθέτως, αν το w πάρει την αντίθετη κατεύθυνση από το ∇E_{in}, θα κινούμαι προς τη βέλτιστη λύση. Επομένως ορίζω τη συνθήκη:

$$\nabla E_{in} = -2\frac{\lambda}{N} w_{reg}$$

όπου ως w_{reg} συμβολίζονται οι κανονικοποιημένοι πλέον συντελεστές, λ μια σταθερή παράμετρος που επηρεάζει την κανονικοποίηση και N, το γνωστό μας πλήθος παραδειγμάτων.

Σκεπτόμενοι αντιστρόφως από ότι συνήθως, παρατηρούμε πως η παραπάνω εξίσωση μοιάζει με μηδενισμό της παραγώγου μιας συνάρτησης και επομένως θα μπορούσε να αποτελέσει την προσπάθεια βελτιστοποίησης του παρακάτω προβλήματος:

$$E_{in} + \frac{\lambda}{N} w_{reg}^T w_{reg} = 0$$

Η παραπάνω διατύπωση είναι αποκαλυπτική: ένα πρόβλημα βελτιστοποίησης υπό συνθήκη αναδιατυπώθηκε ως ένα απλό πρόβλημα βελτιστοποίησης μιας συνάρτησης, που αντικαθιστώντας με τη γνωστή μας συνάρτηση του E_{in} εκφράζεται ως:

$$E_{in}(w) = \frac{1}{N}(Zw - y)^T(Zw - y) + \frac{\lambda}{N}w_{reg}^T w_{reg}$$

Με τη ίδια λογική, που είχαμε επιλύσει το πρόβλημα χωρίς κανονικοποίηση, το ψευδοαντίστροφο, η λύση προκύπτει ως:

$$w_{lin} = (Z^T Z)^{-1} Z^T y$$

Αν και είδαμε την εφαρμογή της κανονικοποίησης σε ένα συγκεκριμένο μοντέλο, μπορούμε να κατανοήσουμε τη συνολική διαδικασία ως εξής:

$$E_{aug}(h) = E_{in}(h) + \frac{\lambda}{N}\Omega(h) = 0$$

όπου ως $E_{in(h)}$ συμβολίζουμε το σφάλμα μιας υπόθεσης πριν την κανονικοποίηση και $E_{aug}(h)$ το επαυξημένο σφάλμα, δηλαδή έπειτα από την κανονικοποίηση. Η συνάρτηση $\Omega(h)$ εκφράζει τον τρόπο που θα γίνει η κανονικοποίηση και, ανεξάρτητα από το μοντέλο, έχει ως στόχο την επιβολή περιορισμών που εξασφαλίζουν ομαλότητα ή/και χαμηλότερη πολυπλοκότητα της υπόθεσης.

Η επίδραση της παραμέτρου λ Αν η συνάρτηση $\Omega(h)$ εκφράζει τον τρόπο, τότε το λ εκφράζει το μέγεθος της κανονικοποίησης.

Όπως παρατηρούμε στο παραπάνω διάγραμμα υποπεριπτώσεων, επιβολή μικρής τιμής λ, αντιστοιχεί σε μεγάλο C, δηλαδή ασθενή περιορισμό του μοντέλου, που συνεχίζει να κινδυνεύει από υπερπροσαρμογή. Καθώς αυξάνουμε την τιμή του λ γινόμαστε αυστηρότεροι, με κίνδυνο να είμαστε τόσο περιοριστικοί με το μοντέλο που θα οδηγηθούμε σε υπόθεση υψηλής πόλωσης.

Σχήμα Α΄.1: Μοντέλο υψηλής πόλωσης

Σχήμα Α΄.2: Μοντέλο υψηλής διακύμανσης

Σχήμα Α΄.3: Παράδειγμα κανονικοποίησης: Η άγνωστη συνάρτηση που προσπαθούμε να προβλέψουμε είναι ένα ημίτονο και το μοντέλο που επιλέγουμε είναι το γραμμικό, για χάρη απλότητας. Αριστερά, πριν την κανονικοποίηση, το μοντέλο μας είναι ελεύθερο να επιλέξει οποιαδήποτε ευθεία επιθυμεί. Δεξιά, έχουμε περιορίσει την κλίση και το σταθερό συντελεστή της ευθείας, καταφέρνοντας μικρότερη διακύμανση στις υποθέσεις μας και μικρότερο σφάλμα, καθώς έχουν αποκλειστεί κάποιες πολύ κακές υποθέσεις.

Σχήμα Α΄.4: Γραφική αναπαράσταση κανονικοποίησης.

Σχήμα Α΄.5: Επίδραση λ στην κανονικοποίηση.

Παράρτημα Β΄

Μηχανές Διανυσματικής Στήριξης

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. 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 είναι κάπως έτσι: Θα θέλαμε η υπόθεσή μας να διαχωρίσει τα παραπάνω δεδομένα με βάση την κλάση τους, πράγμα που διαπιστώνουμε πως μπορεί να επιτευχθεί με μερικές διαφορετικές υποθέσεις: 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. Ένας πιο επίσημος ορισμός, που επεκτείνεται σε περισσότερες διαστάσεις, είναι πως ορίζεται ένα υπερεπίπεδο που διαχωρίζει τις κατηγορίες.

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 $\left\|w^T x\right\| = 1$
- 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:

$$distance = |\bar{w}(x_n - x)| = \frac{1}{\|w\|}|w^T x_n - w^T x| = \frac{1}{\|w\|}|w^T x_n + b - w^T x - b| = \frac{1}{\|w\|}$$

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:

Μεγιστοποίηση $\frac{1}{||w||}$

υπό τον περιορισμό ότι $\min_{n=1,2,...,N} |w^T x + b| = 1$

Η παραπάνω διατύπωση δεν είναι φιλική προς επίλυση, κυρίως λόγω της μορφής του περιορισμού, για αυτό θα την αναδιατυπώσουμε ως εξής:

Ελαχιστοποίηση $\frac{1}{2} w^T w$

υπό τον περιορισμό ότι $y_n(w^T x_n + b) \geq 1, n = 1, ..., N$

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, επεκτείνουν την εφαρμογή των πολλαπλασιαστών Lagrange, επιτρέποντας τη βελτιστοποίηση προβλήματα υπό περιορισμούς σε μορφή ανισοτήτων.

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

Ελαχιστοποίηση $L(w,b,a) = \frac{1}{2} w^T w - \sum_{n=1}^{N} a_n (y_n(w^T x_n + b) - 1)$

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:

Άρα $\nabla_w L = w - \sum_{n=1}^{N} a_n y_n x_n = 0$

και $\frac{\partial L}{\partial b} = \sum_{n=1}^{N} a_n y_n = 0$

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

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

$$\text{υπό τη συνθήκη} \qquad \sum_{n=1}^{N} a_n y_n = 0$$

$$\text{και} \qquad a_n \geq 0$$

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:

$$\begin{bmatrix} y_1 y_1 x_1^T x_1 & y_1 y_2 x_1^T x_2 & \dots & y_1 y_N x_1^T x_N \\ y_2 y_1 x_2^T x_1 & y_2 y_2 x_2^T x_2 & \dots & y_2 y_N x_2^T x_N \\ \vdots & \vdots & \ddots & \vdots \\ y_N y_1 x_N^T x_1 & y_N y_2 x_N^T x_2 & \dots & y_N y_N x_N^T x_N \end{bmatrix}$$

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, επομένως 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:

$$X \to Z$$

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:

$$K(x, x') = z \cdot z'$$

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, που εξασφαλίζει πως οποιαδήποτε συνάρτηση πυρήνα

$$K(x, x')$$

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

$$\begin{bmatrix} (x_1, x_1) & (x_1, x_2) & \dots & (x_1, x_N) \\ (x_2, x_1) & (x_2 x_2) & \dots & (x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ (x_N, x_1) & (x_N, x_2) & \dots & (x_N, x_N) \end{bmatrix}$$

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

- *Πολυωνυμικός.* Δίνεται από τον τύπο:

  $$K(x, x') = (x^T x' + c)^d$$

  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:

  $$K(x, x') = e^{(-\frac{\|x-x'\|^2}{2\sigma^2})}$$

  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 Κάθε φορά λοιπόν που παρατηρούμε μη γραμμικότητα στα δεδομένα μας θα εφαρμόζουμε τη συνάρτηση πυρήνα; Ας δούμε το παρακάτω παράδειγμα: Και τα δύο σχήματα αντιστοιχούν σε μη γραμμικά διαχωρίσιμα προβλήματα, ωστόσο διαφοροποιούνται ποιοτικά: στα αριστερά, ένας γραμμικός διαχωρισμός θα προκαλούσε πολύ μικρό σφάλμα, καθώς μόνο δύο παραδείγματα του σετ εκπαίδευσης θα κατηγοριοποιηθούν λανθασμένα. Αν προσπαθήσω να τα κατατάξω και αυτά σωστά, τότε η υπόθεσή μου θα γίνει πολύ περίπλοκη, καθώς θα χρειαστούν πολλά διανύσματα στήριξης και υποπτεύομαι πως το μοντέλο μου δεν θα γενικεύει. Αντιθέτως, η δεξιά εικόνα αντιστοιχεί σε εμφανώς μη γραμμικά διαχωρίσιμο πρόβλημα, που επιλύεται μόνο με τη χρήση συνάρτησης πυρήνα.

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 :

$$y_n(w^T x_n + b) \geq 1 - \xi_n, n = 1, ..., N$$

και η εξίσωση που βελτιστοποιεί πλέον ο αλγόριθμος είναι:

$$\begin{aligned} \text{Ελαχιστοποίηση} \quad & \frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n \\ \text{υπό τον περιορισμό ότι} \quad & y_n(w^T x_n + b) \geq 1 - \xi_n, n = 1, \ldots, N \\ \text{και} \quad & \xi_n \geq 0 \end{aligned}$$

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.

Σχήμα Β΄.1: Χώρος ταξινόμησης

Σχήμα Β΄.2: Υπόθεση Α

Σχήμα Β΄.3: Υπόθεση Β

Σχήμα Β΄.4: Υπόθεση Γ

Σχήμα Β΄.5: Υπερεπίπεδο μηχανών διανυσματικής στήριξης

Σχήμα Β΄.6: Χρήση πυρήνα για επίλυση μη γραμμικού διαχωρισμού

Σχήμα Β΄.7: Ελάχιστα μη διαχωρίσιμα δεδομένα

Σχήμα Β΄.8: Εμφανώς μη διαχωρίσιμα δεδομένα

Σχήμα Β΄.9: Μηχανές διανυσματικής στήριξης χαλαρού περιθωρίου

Παράρτημα Γ΄

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:

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)}$$

όπου h είναι η υπόθεση και d τα παραδείγματα.

Ως γνωστόν, σε ένα πρόβλημα ταξινόμησης η υπόθεση ισοδυναμεί με την κλάση που αναθέτουμε σε ένα παράδειγμα. Ας δούμε λίγο πιο αναλυτικά τις πιθανότητες, με τις οποίες ασχολούμαστε:

- $P(h \mid d)$ 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:

$$P(cloudy =' yes' \mid rain =' yes') = \frac{count(cloudy =' yes', rain =' yes')}{count(rain =' yes')}$$

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:

$$MAP(h) = P(d \mid h)P(h)$$

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 στο λογισμικό που τις εκτελεί. Για αυτό το λόγο συνηθίζουμε να δουλεύουμε με τους λογαρίθμους των πιθανοτήτων και όχι απευθείας με τις πιθανότητες.
- μόλις έχουμε ολοκληρώσει μια πρόβλεψη και είμαστε σίγουροι για αυτήν, μπορούμε να επαναυπολογίσουμε το μοντέλο για να το εμπλουτίσουμε με τη νέα γνώση.

Παράρτημα Δ΄

Λογιστική Παλινδρόμηση

Σκοπός αυτού του αλγορίθμου δεν είναι ακριβώς να ταξινομήσει τα δεδομένα, αλλά να δώσει πιθανότητες σε κάθε κλάση δεδομένων των χαρακτηριστικών.

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:

$$\theta(s) = \frac{e^s}{1+e^s}$$

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 πως 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:

$$s = w^T x \rightarrow h(x) = \theta(s) = \frac{e^{w^T x}}{1+e^{w^T x}}$$

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

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

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:

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

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

$$P(y \mid x) = \theta(y w^T x)$$

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

$$\prod_{n=1}^{N} P(y \mid x) = \prod_{n=1}^{N} \theta(y_n w^t x_n)$$

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:

$$Error(w) = \frac{1}{N} \sum_{n=1}^{N} \ln(1 + e^{-y_n w^T x_n})$$

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 απόδοση 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.

Σχήμα Δ΄.1: Η λογιστική συνάρτηση

Σχήμα Δ΄.2: Steepest descent

Σχήμα Δ΄.3: Gradient descent με πολύ μικρό βήμα

Σχήμα Δ΄.4: Gradient descent με πολύ μεγάλο βήμα

Παράρτημα Ε΄

Κ-κοντινότερος γείτονας

Ο αλγόριθμος αυτός ανήκει στην κατηγορία των αλγορίθμων αλγορίθμων βασισμένων σε παραδείγματα, 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.

Αν φανταστούμε πως τα δεδομένα ζουν σε ένα χώρο ίσων διαστάσεων με το πλήθος των χαρακτηριστικών και διαφοροποιούνται ως προς την κλάση τους, τότε 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:
$$\sqrt{\sum_{i=1}(x_i - y_i)^2}$$
- *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:
$$\sum_{i=1}|x_i - y_i|$$

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

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

Η επιλογή της βέλτιστης υπόθεσής 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:

$$E_{in} = 0 \rightarrow \sum_{n=1}^{N} w_n e^{-\gamma \|x_n - x_m\|^2} = y_n \forall n \in D_N$$

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

Ο παραπάνω τύπος δίνει ένα σύστημα Ν γραμμικών εξισώσεων με Ν 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:

$$w = \Phi^{-1} y$$

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 Κ 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 Κ 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 Κ και όχι Ν παραμέτρων.

Πλέον η υπόθεση δίνεται από τον τύπο:

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

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:

$$w = \Phi^T \Phi^{-1} \Phi^T y$$

Επίλυση υπερορισμένων συστημάτων με χρήση ψευδοαντίστροφου

Ένα γραμμικό σύστημα $y = Ax$ χαρακτηρίζεται ως υπερορισμένο 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, που ορίζεται ως $A^{\dagger} = (A^T A)^{-1} A^T$, ώστε να ισχύει $A^{\dagger} =$, 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 $x \approx A^{\dagger} y$

Το νέο πρόβλημα που αναδύεται είναι αυτό της επιλογής των βέλτιστων μ_k. Το πρόβλημα αυτό ονομάζεται k-means ομαδοποίηση 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:

$$\sum_{k=1}^{K} \sum_{x_n \in S_k} \left\| x_n - \mu_k \right\|^2$$

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

Η λύση του δίνεται από τον αλγόριθμο του Lloyd, 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.

Σχήμα Ε΄.1: K-NN ταξινομητής: Το άγνωστο σημείο θα ταξινομηθεί ως θετικό.

Σχήμα Ε΄.2: Radial Basis συνάρτηση με γκαουσιανή βάση.

Σχήμα Ε΄.3: RBF με μικρό γ

Σχήμα Ε΄.4: RBF με μεγάλο γ

Παράρτημα ΣΤ΄

Στατιστικά τεστ Υπόθεσης

Γενικά χαρακτηριστικά των στατιστικών τεστ υπόθεσης έχουν περιγραφεί στην Ενότητα 2.2.3. Σε αυτό το σημείο θα αναφέρουμε περιληπτικά μερικά είδη τέτοιων τεστ, τα οποία διαφοροποιούνται κυρίως ως προς:

- τις υποθέσεις που κάνουν για τους πληθυσμούς.
-

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 από τον "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" [71] 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 δεδομένων.

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

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 και των ίσων διακυμάνσεων των residuals και την απώλεια ισχύος.

Σημαντική προσθήκη αποτελεί η εναλλακτική test statistic που εισήγαγαν οι Iman and Davenport [37], 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 [38, 31, 32].

Fisher's exact τεστ Εισήχθη από τον Fisher μέσω ενός παραδείγματος [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 [80] και αναλύθηκε διεξοδικά από τους Mann and Whitney [48]. 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 [50] 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 [15] and Mantel and Haenszel [49] και αποτελεί γενίκευση του McNemar, καθώς υποστηρίζει διαστρωμάτωση των δεδομένων σε αυθαίρετο πλήθος ομάδων.

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

Παράρτημα Ζ΄

Αλγόριθμοι

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

Αλγόριθμος 1: Γενικός Ψευδοκώδικας SMBO

Παράρτημα Η΄

Εξαγωγή διαστημάτων πρόβλεψης από μοντέλα παλινδρόμησης

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.

[ΒΑΛΕ ΚΑΙ ΑΛΛΑ ΣΤΑΤΙΣΤΙΚΣ ΚΑΙ ΟΝΟΜΑΣΕ ΑΥΤΟ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΣΕΣ]

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 [60] για μοντέλα ταξινόμησης και η μέθοδος των Jiang, Zhang, and Cai [39] και Lin and Weng [44] 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 [14] ,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 [44] 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

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

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

Παράρτημα Θ΄

Σετ δεδομένων

	Όνομα	Πηγή	Πεδίο	Τύπος	Πλήθος παρα-δειγ-μάτων	Πλήθος χαρα-κτη-ριστι-κών	Κλάση	NAs
1	accident							
2	ad							
3	adni_demographic_master_kaggle							
4	adult [41]	UCI	social	data	48842	14	binary	Ναι
5	adult-stretch [58]	UCI	social	data	16	4	binary	Όχι
6	adult+stretch [58]	UCI	social	data	16	4	binary	Όχι
7	AP_Endometrium_Breast [2]	openml	N/A	arff	405	10937	binary	Όχι
8	AP_Endometrium_Lung [3]	openml	N/A	arff	195	10937	binary	Όχι
9	AP_Endometrium_Omentum							
10	arsenic-male-bladder							
11	Attribute_DataSet	UCI	Computers	xlsx	501	13	binary	Ναι
12	australian							
13	baboon_mating							
14	bands							
15	bank-additional	UCI	csv	business	2002	20	binary	Ναι
16	bank-additional-full	UCI	csv	business	45211	20	binary	Ναι
17	biodeg							
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 [10]	UCI	N/A	data	1728	6	binary	Όχι
28	chess	relational	sport	mysql	296	19	binary	Όχι
29	chscase_health							
30	cities_r2 [75]	kaggle	N/A	csv	494	21	continuous	Όχι
31	confidence							

32	creditcard [17]	kaggle	N/A	csv	284808	30	binary	Όχι
33	crx [63]	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 [4]	relational	Synthetic	mysql	7129	3		
43	default_of_credit_card_clients [20]	UCI	Business	xls	30000	24	binary	Όχι
44	dermatology [57]	UCI	Life	data	366	33	multiclass	Ναι
45	diagnosis							
46	fertility_Diagnosis	UCI	Life	txt	100	10	binary	Όχι
47	ftp [54]	relational	Synthetic	mysql	29555	2	binary	Ναι
48	gym [18]	kaggle	N/A	csv	26067	6	continuous	Όχι
49	haberman [74]	UCI	Life	data	306	3	binary	Όχι
50	heart							
51	hepatitis [29]	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 [35]	kaggle	N/A	csv	15000	9	multiclass	Όχι
56	imdb [42]	relational	Real	mysql	986583	5	continuous	Όχι
57	indian_ilpd	UCI	Life	csv	583	10	binary	Όχι
58	ionosphere [68]	UCI	Physical	data	351	34	binary	Όχι
59	kohkiloyeh	UCI	computer	xlsx	100	6	binary	Όχι
60	krk [**krk**]	relational	Synthetic	mysql	1000	6	binary	Όχι
61	lupus [47]	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 [73]	UCI	N/A	txt	432	7	binary	Όχι
67	monks-2.train [73]	UCI	N/A	txt	432	7	binary	Όχι
68	monks-3.train [73]	UCI	N/A	txt	432	7	binary	Όχι
69	mushrooms [55]	kaggle	N/A	csv	8125	22	binary	Ναι
70	musk [6]	relational	Real	mysql	6599	6598	binary	Όχι
71	mutagenesis [19]	relational	Real	mysql	5244	16	binary	Όχι
72	numeric sequence [70]	kaggle	N/A	csv	2401	28	binary	Όχι
73	nursery [65]	UCI	social	data	12960	8	multiclass	Όχι
74	parkinsons [45]	UCI	Life	data	197	23	binary	Όχι
75	php3BOEY5 [7]	openml	N/A	arff	745	37	binary	Όχι
76	php4ylQmK [64]	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 [9]	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 [21]	UCI	Life	data	768	8	binary	Ναι
90	Pokemon							
91	Political-media-DFE							
92	prostate_TumorVSNormal	openml	N/A	arff	136	12601	binary	Όχι
93	ptc [28]	relational	Real	mysql	18313	6	binary	Όχι
94	Qualitative_Bankruptcy							
95	rabe_97							
96	reviews							
97	SalesKaggle3 [30]	kaggle	N/A	csv	198918	14	continuous	Όχι
98	seismic-bumps							
99	shuttle-landing-control							
100	sonar	UCI	Physical	data	208	60	binary	Όχι
101	spambase [33]	UCI	Computer	data	4601	57	binary	Ναι
102	SPECTF							
103	student-mat [72]	kaggle	N/A	csv	396	32	multiclass	Όχι
104	testing [40]	UCI	Life	csv	500	5	binary	Όχι
105	ThoracicSurgery							
106	tic-tac-toe [1]	UCI	Game	data	958	9	binary	Όχι
107	trains [51]	relational	Synthetic	mysql	64	7	binary	Όχι
108	training [40]	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 [67]	relational	Synthetic	mysql	93	5	continuous	Όχι
112	university_salary [67]	relational	Synthetic	mysql	26	5	continuous	Όχι
113	utube [5]	relational	Real	mysql	100001	4	continuous	Όχι
114	vgsales_EU [78]	kaggle	N/A	csv	16598	10	continuous	Όχι
115	vgsales_global [78]	kaggle	N/A	csv	16598	10	continuous	Όχι
116	vgsales_JP [78]	kaggle	N/A	csv	16598	10	continuous	Όχι
117	vgsales_NA [78]	kaggle	N/A	csv	16598	10	continuous	Όχι
118	vgsales_other [78]	kaggle	N/A	csv	16598	10	continuous	Όχι
119	Video_Games_Sales_as_at_30_Nov_2016 [79]	kaggle	N/A	csv	15850	15	continuous	Ναι
120	visualizing_soil							
121	voice [**Gender Recognition by Voice**]	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	relational	Real	mysql	30671	15	continuous	Όχι
125	yellow-small							
126	yellow-small+adult-stretch							

Πίνακας Θ΄.1: Τελική λίστα μετα-χαρακτηριστικών