Machine Learning & Knowledge Extraction

DR KONSTANTINOS KARAMPIDIS

Πληροφορίες Μαθήματος

- Ωράριο:
  - Θεωρία: **Κάθε Τρίτη 09:00-13:00 – Αίθουσα 207**
  - Εργαστήριο: **13:00-14:00 ΕΡΓ6 (σύμφωνα με το πρόγραμμα που είναι αναρτημένο στο eclass)**
- Εργασίες
  - **1 project – Ομάδες έως 2 ατόμων – 80%**
  - **Εργαστηριακές ασκήσεις – 20%**
- Προαπαιτούμενα: Κανένα

Περιεχόμενο Μαθήματος

- Εισαγωγή στη Μηχανική Μάθηση - τι είναι, γιατί μας ενδιαφέρει, παραδείγματα προβλημάτων, η μηχανική μάθηση ως αναζήτηση, υπόθεση επαγωγικής μάθησης
- **Επεξεργασία εισόδου – Μείωση διαστατικότητας- Αξιόλογηση**
- Μέθοδοι επιβλεπόμενης μάθησης
- Νευρωνικά Δίκτυα
- Εξελικτική Μάθηση – Γενετικοί Αλγόριθμοι
- Μέθοδοι μη επιβλεπόμενης μάθησης
- Βαθιά Μάθηση

Αξιολόγηση

- Πόσο καλά προβλέπει το μοντέλο που δημιουργήσαμε;
- Το λάθος (error) στα δεδομένα εκπαίδευσης δεν είναι καλός δείκτης απόδοσης για τα μελλοντικά δεδομένα
  - Διαφορετικά ο 1-NN θα ήταν ένα βέλτιστος ταξινομητής
- Μια απλή λύση αν είναι διαθέσιμο ένα μεγάλο πλήθος από δεδομένα (labeled) :
  - Διαχωρισμός δεδομένων σε σύνολα εκπαίδευσης (training set) και ελέγχου (test set)
- Όμως αυτά τα δεδομένα (labeled) είναι συνήθως περιορισμένα
  - χρειάζονται πιο προηγμένες τεχνικές

Ζητήματα στην Αξιολόγηση

- Στατιστική αξιοπιστία των εκτιμώμενων διαφορών στην απόδοση (έλεγχοι σημαντικότητας)
- Επιλογή μετρικών απόδοσης:
  - Αριθμός σωστών ταξινομήσεων
  - Ακρίβεια στην εκτίμηση πιθανοτήτων
  - Λάθος στις αριθμητικές προβλέψεις
- Τα κόστος που ανατίθεται σε διαφορετικά είδη λαθών
  - Το κόστος είναι σημαντικό για πολλές πρακτικές εφαρμογές

Έλεγχος και Εκπαίδευση

- Η αναλογία λαθών είναι η φυσική μετρική απόδοσης για προβλήματα ταξινόμησης
  - *Επιτυχία*: η κλάση μιας περίπτωσης προβλέπεται σωστά
  - *Λάθος*: η κλάση μιας περίπτωσης προβλέπεται λανθασμένα
  - Αναλογία λαθών: ποσοστό λαθών που έχουν γίνει σε όλες τις περιπτώσεις
- *Λάθος επανα-αντικατάστασης*: η αναλογία λαθών που λαμβάνεται στην Αξιολόγηση του μοντέλου σε δεδομένα εκπαίδευσης
- Αυτό το είδος λάθους είναι βέλτιστο

Έλεγχος και Εκπαίδευση

- *Σύνολο Ελέγχου (test set)*: ανεξάρτητες περιπτώσεις που δεν έχουν λάβει μέρος στην εκπαίδευση ενός ταξινομητή
  - Υπόθεση: τόσο τα δεδομένα εκπαίδευσης (training set) και ελέγχου (test set) είναι αντιπροσωπευτικά δείγματα του αντίστοιχου προβλήματος
- Τα δεδομένα ελέγχου και εκπαίδευσης μπορεί να διαφέρουν κατά φύση
  - Παράδειγμα: ταξινομητές κατασκευαζόμενοι με τη χρήση δεδομένων πελατών από δύο διαφορετικές πόλεις *A* και *B*
  - Η εκτίμηση απόδοσης του ταξινομητή της πόλης Α για μια καινούργια πόλη μπορεί να γίνει μέσω του ελέγχου του με δεδομένα απο την πόλη Β

Ρύθμιση Παραμέτρων

- Είναι σημαντικό τα δεδομένα ελέγχου να μην χρησιμοποιηθούν με οποιαδήποτε τρόπο για τη δημιουργία του ταξινομητή
- Ορισμένα σχήματα μάθησης δουλεύουν σε 2 φάσεις:
  - Φάση 1: κατασκευή της βασικής δομής
  - Φάση 2: βελτιστοποίηση της ρύθμισης παραμέτρων
- Τα δεδομένα ελέγχου δεν πρέπει να χρησιμοποιηθούν για την ρύθμιση παραμέτρων
- Μια κατάλληλη διαδικασία χρησιμοποιεί 3 σύνολα: δεδομένα εκπαίδευσης, επικύρωσης και ελέγχου
  - Τα δεδομένα επικύρωσης χρησιμοποιούνται για την βελτιστοποίηση παραμέτρων

Εκμετάλλευση Δεδομένων

- Μόλις η Αξιολόγηση ολοκληρωθεί, όλα τα δεδομένα μπορούν να χρησιμοποιηθούν για την κατασκευή του τελικού ταξινομητή
- Εν γένει, όσο περισσότερα είναι τα δεδομένα εκπαίδευσης τόσο καλύτερος γίνεται και ο ταξινομητής (αλλά τα καλά αποτελέσματα μειώνονται μετά από ένα κατώφλι)
- Όσο περισσότερα είναι τα δεδομένα ελέγχου, τόσο πιο ακριβής είναι η Αξιολόγηση λάθους
- Διαδικασία *Κράτησης* (*Holdout*): μέθοδος διαχωρισμού των αρχικών δεδομένων σε σύνολα εκπαίδευσης και ελέγχου
  - Δίλημμα: ιδανικά και τα 2 σύνολα θα πρέπει να είναι μεγάλα

Holdout estimation

- Τι συμβαίνει όταν έχουμε ένα σύνολο από δεδομένα;
- Η μέθοδος holdout διατηρεί ένα συγκεκριμένο πλήθος για έλεγχο (τεστ) και χρησιμοποιεί το υπόλοιπο για εκπαίδευση (train) αφού τα δεδομένα πρώτα ανακατευθούν
  - Συνήθως: ένα τρίτο για έλεγχο και δύο τρίτα για εκπαίδευση
- Πρόβλημα: τα δείγματα μπορεί να μην είναι αντιπροσωπευτικά
  - Παράδειγμα: μια κλάση μπορεί να μην υπάρχει σε δεδομένα ελέγχου
- Η προχωρημένη έκδοση της μεθόδου χρησιμοποιεί διαστρωμάτωση (stratification)
  - Διαβεβαιώνει ότι κάθε κλάση αναπαρίσταται με σχεδόν ισότιμη αναλογία και στα δύο υποσύνολα δεδομένων

Repeated Holdout method

- Μέθοδος επαναλαμβανόμενης κράτησης: Η εκτίμηση κράτησης μπορεί να γίνει πιο αξιόπιστη μέσω της επανάληψης της διαδικασίας με διαφορετικά υποδείγματα/υποσύνολα
  - Σε κάθε επανάληψη, μια συγκεκριμένη αναλογία επιλέγεται τυχαία για την εκπαίδευση (πιθανώς με διαστρωμάτωση)
  - Ο μέσος όρος των αναλογίων λαθών κατά μήκος όλων των επαναλήψεων υπολογίζεται για την αξιολόγηση της συνολικής αναλογίας λαθών
- Όμως, η μέθοδος αυτή δεν είναι ακόμη βέλτιστη διότι τα διαφορετικά σύνολα ελέγχου αλληλεπικαλύπτονται
  - Μπορούμε να αποφύγουμε αυτή την αλληλοεπικάλυψη;

Cross Validation

- Η *επικύρωση με Κ-αναδιπλώσεις/πτυχές (K-fold cross-validation)* αποφεύγει αλληλεπικαλυπτόμενα σύνολα ελέγχου
  - Πρώτο βήμα: διαχωρισμός δεδομένων σε *k* υποσύνολα ίσου μεγέθους
  - Δεύτερο βήμα: χρήση κάθε υποσυνόλου για έλεγχο, ενώ τα υπόλοιπα για εκπαίδευση
  - Αυτό σημαίνει ότι ο αλγόριθμος μάθησης εφαρμόζεται σε *k* διαφορετικά σύνολα εκπαίδευσης

Cross Validation

- Συχνά πραγματοποιείται stratification των υποσυνόλων πριν γίνει το cross validation, ώστε να πετύχουμε stratified *k*-fold cross validation.
- Ξανά η συνολική αξιολόγηση λάθους υπολογίζεται μέσω του μέσου όρου των εκτιμήσεων λάθους. Επίσης, συχνά υπολογίζεται και η τυπική απόκλιση
- Εναλλακτικά, οι προβλέψεις και οι πραγματικές τιμές στόχου από τις *k* πτυχές λαμβάνονται για τον υπολογισμό μιας εκτίμησης μόνο
  - Αυτό δεν οδηγεί στην εκτίμηση της τυπικής απόκλισης

Cross Validation

- Τυπική μέθοδος αξιολόγησης: stratified 10-fold cross validation
- Γιατί 10;
  - Εκτεταμένα πειράματα έχουνε δείξει ότι αυτή είναι η καλύτερη επιλογή για μια ακριβή εκτίμηση
  - Επιπλέον, υπάρχει μια θεωρητική απόδειξη για αυτό
- Η διαστρωμάτωση (stratification) μειώνει την διασπορά της εκτίμησης
- Ακόμη καλύτερα: repeated stratified cross- validation
  - Πχ., Η 10-fold cross validation επαναλαμβάνεται 10 φορές και υπολογίζεται ο μέσος όρος των αποτελεσμάτων (αυτό μειώνει τη διασπορά)

Leave-one-out cross-validation

- Leave one-out:
  συγκεκριμένη μορφή κατά της *k*-fold cross-validatoon:
  - Θέσε τον αριθμό των πτυχών ίσο με τον αριθμό των περιπτώσεων εκπαίδευσης
  - Δηλ., για *n* περιπτώσεις εκπαίδευσης, κατασκεύασε ταξινομητή *n* φορές
- Πραγματοποιεί τη βέλτιστη χρήση δεδομένων
- Δεν περιλαμβάνει τυχαία υπο-δειγματοποίηση
- Πολύ απαιτητική μέθοδος σε υπολογιστικούς πόρους (εξαίρεση: χρήση χαλαρών (lazy) ταξινομητών όπως αυτού του κοντινότερου γείτονα)

Leave-one-out cross-validation & stratification

- Στο προηγούμενο είδος επικύρωσης δεν είναι δυνατή η διαστρωμάτωση
  - Εξασφαλίζει ένα μη διαστρωματωμένο δείγμα διότι υπάρχει μόνο μια περίπτωση στο σύνολο ελέγχου
- Οριακό παράδειγμα: τυχαίο σύνολο δεδομένων που διαχωρίζεται εξίσου σε δύο κλάσεις
  - 50% ακρίβεια σε «φρέσκα» δεδομένα
  - Η εκτίμηση μέσω του προηγούμενου είδους επικύρωσης παρέχει λάθη 100%

Αυτοδύναμη Εκκίνηση (Bootstrap)

- Η μέθοδος cross-validation χρησιμοποιεί δειγματοληψία χωρίς αντικατάσταση
  - Όταν επιλεχθεί μια περίπτωση, δεν μπορεί να επιλεχθεί ξανά για ένα συγκεκριμένο σύνολο εκπαίδευσης/ελέγχου
- Αντιθέτως, η μέθοδος bootstrap χρησιμοποιεί δειγματοληψία με αντικατάσταση για το σχηματισμό του συνόλου εκπαίδευσης
  - Δειγματοληψία συνόλου δεδομένων από *n* περιπτώσεις *n* φορές με αντικατάσταση για τον σχηματισμό ενός νέου συνόλου δεδομένων με *n* περιπτώσεις
  - Χρήση αυτών των δεδομένων ως το σύνολο εκπαίδευσης
  - Χρήση των περιπτώσεων του αρχικού συνόλου δεδομένων, που δεν συμπεριλαμβάνονται στο νέο σύνολο εκπαίδευσης, για έλεγχο

Αυτοδύναμη Εκκίνηση (Bootstrap)

Ας υποθέσουμε ότι έχουμε ένα μικρό δείγμα δεδομένων που αντιπροσωπεύει το ύψος (σε ίντσες) 10 ατόμων:

Ύψος = [65.2, 67.1, 68.5, 69.3, 70.0, 71.2, 72.4, 73.1, 74.5, 75.8]

Θέλουμε να εκτιμήσουμε το 95% διάστημα εμπιστοσύνης για το μέσο ύψος στον πληθυσμό χρησιμοποιώντας bootstrapping.

Ακολουθούν τα βήματα που θα ακολουθούσαμε:

- Υπολογίζουμε τον μέσο όρο από τα αρχικά δεδομένα:

Μέσος όρος δείγματος = (65,2 + 67,1 + 68,5 + 69,3 + 70,0 + 71,2 + 72,4 + 73,1 + 74,5 + 75,8) / 10 = 70,71 ίντσες

Αυτοδύναμη Εκκίνηση (Bootstrap)

- Δημιουργήστε έναν μεγάλο αριθμό δειγμάτων bootstrap από τα αρχικά δεδομένα με επαναδειγματοληψία με αντικατάσταση. Για παράδειγμα, ας δημιουργήσουμε 10.000 δείγματα bootstrap, το καθένα μεγέθους 10.

- Για κάθε δείγμα bootstrap, υπολογίστε το μέσο ύψος.

Αφού υπολογίσουμε τους μέσους όρους για όλα τα 10.000 δείγματα bootstrap, έχουμε τώρα μια εμπειρική κατανομή δειγματοληψίας bootstrap του μέσου όρου.

Αυτοδύναμη Εκκίνηση (Bootstrap)

Από αυτή την εμπειρική κατανομή δειγματοληψίας bootstrap, μπορούμε να προσδιορίσουμε το 95% διάστημα εμπιστοσύνης βρίσκοντας το 2,5ο και το 97,5ο εκατοστημόριο (percentile) της κατανομής.

Ας υποθέσουμε ότι το 2,5ο εκατοστημόριο είναι 69,8 ίντσες και το 97,5ο εκατοστημόριο είναι 71,6 ίντσες.

Τότε, το 95% διάστημα εμπιστοσύνης (confidence interval) για το μέσο ύψος είναι [69,8, 71,6] ίντσες.

Αυτό το διάστημα εμπιστοσύνης σημαίνει ότι αν επαναλαμβάναμε τη διαδικασία λήψης ενός δείγματος μεγέθους 10 και κατασκευής ενός διαστήματος εμπιστοσύνης bootstrap πολλές φορές, το 95% αυτών των διαστημάτων θα περιείχε το πραγματικό μέσο ύψος του πληθυσμού.

Αυτοδύναμη Εκκίνηση (Bootstrap) – Πλεονεκτήματα

- **Μη παραμετρική φύση:** Η μέθοδος Bootstrap δεν βασίζεται σε υποθέσεις σχετικά με την υποκείμενη κατανομή των δεδομένων. Αυτό την καθιστά ιδιαίτερα χρήσιμη όταν έχει να κάνει με σύνθετες ή άγνωστες κατανομές, επιτρέποντας πιο ευέλικτη και ισχυρή στατιστική ανάλυση.
- **Ευελιξία:** Μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα στατιστικών μέτρων, συμπεριλαμβανομένων των μέσων όρων, των διαμέσων, των αποκλίσεων και των συντελεστών παλινδρόμησης. Αυτή η ευελιξία επεκτείνεται σε διάφορους τύπους δεδομένων, είτε πρόκειται για συνεχή, διακριτά ή κατηγορικά δεδομένα.
- **Ακρίβεια σε μικρά δείγματα:** Σε περιπτώσεις όπου τα μεγέθη των δειγμάτων είναι μικρά, οι παραδοσιακές μέθοδοι ενδέχεται να μην παρέχουν αξιόπιστες εκτιμήσεις. Η μέθοδος Bootstrap μπορεί να βελτιώσει την ακρίβεια αυτών των εκτιμήσεων αυξάνοντας αποτελεσματικά το μέγεθος του δείγματος μέσω επαναδειγματοληψίας.

Αυτοδύναμη Εκκίνηση (Bootstrap) – Πλεονεκτήματα

- **Απλή εφαρμογή:** Η μέθοδος Bootstrap είναι απλό να εφαρμοστεί με τη χρήση σύγχρονων υπολογιστικών εργαλείων. Περιλαμβάνει επαναλαμβανόμενη επαναληπτική δειγματοληψία και μπορεί να προγραμματιστεί εύκολα, καθιστώντας την προσιτή σε ερευνητές και αναλυτές.
- **Εσωτερική επικύρωση:** Η μέθοδος Bootstrap επιτρέπει την εσωτερική επικύρωση των στατιστικών μοντέλων, δημιουργώντας πολλαπλά σύνολα δεδομένων με νέα δειγματοληψία. Αυτό βοηθά στην αξιολόγηση της σταθερότητας και της αξιοπιστίας των μοντέλων χωρίς την ανάγκη για πρόσθετα εξωτερικά δεδομένα.
- **Χειρισμός πολύπλοκων δομών δεδομένων:** Η μέθοδος Bootstrap είναι ικανή να χειρίζεται πολύπλοκες δομές δεδομένων. Αυτή η προσαρμοστικότητα την καθιστά κατάλληλη για ένα ευρύ φάσμα εφαρμογών σε διάφορους τομείς.

Αυτοδύναμη Εκκίνηση (Bootstrap) – Μειονεκτήματα

- **Χρονοβόρα:** Απαιτεί χιλιάδες προσομοιωμένα δείγματα.
- **Υπολογιστικό κόστος:** Επειδή απαιτεί χιλιάδες δείγματα και είναι χρονοβόρο, απαιτεί επίσης μεγαλύτερη υπολογιστική ισχύ.
- **Μερικές φορές ασύμβατη:** Το bootstrapping δεν είναι πάντα η καλύτερη λύση για την περίπτωσή σας, ιδίως όταν πρόκειται για χωρικά δεδομένα ή χρονοσειρές.

Υπολογισμός Υπερ-παραμέτρων

- *Υπερ-παράμετρος*: παράμετρος που μπορεί να ρυθμιστεί για την βελτιστοποίηση της απόδοσης ενός αλγορίθμου μάθησης
  - Διαφορετική από μια βασική παράμετρος που μπορεί να είναι μέρος του μοντέλου, όπως είναι ο συντελεστής σε ένα μοντέλο γραμμικής παρεμβολής
  - Παράδειγμα υπερ-παραμέτρου: *k* σε ένα ταξινομητή *k*-κοντινότερων γειτόνων (*k*-nearest neighbors)
- Δεν επιτρέπεται να χρησιμοποιήσουμε τα δεδομένα ελέγχου για την επιλογή της τιμής αυτής της παραμέτρου
  - Η ρύθμιση της υπερπαραμέτρου μέσω των δεδομένων ελέγχου θα οδηγούσε σε βέλτιστες εκτιμήσεις απόδοσης για τα δεδομένα ελέγχου
  - Η ρύθμιση παραμέτρων θα πρέπει να θεωρηθεί ως μέρος του αλγορίθμου μηχανικής μάθησης και πρέπει να πραγματοποιηθεί μόνο στα δεδομένα εκπαίδευσης

Υπολογισμός Υπερ-παραμέτρων

- Πως όμως να εκτιμήσουμε την απόδοσης για διαφορετικές τιμές παραμέτρων ώστε να επιλέξουμε τη καλύτερη δυνατή τιμή;
  - Απάντηση: διαχώρισε τα δεδομένα σε ένα μικρότερο σύνολο εκπαίδευσης και σε ένα σύνολο επικύρωσης (validation set - αφού ανακατευτούν τα δεδομένα αρχικά)
  - Κατασκεύασε μοντέλα με τη χρήση διαφορετικών τιμών του *k* στο νέο, μικρότερο σύνολο εκπαίδευσης και αποτίμησέ τα στο σύνολο επικύρωσης.
  - Επέλεξε την καλύτερη τιμή για το *k* και ανασκεύασε το μοντέλο σε ολόκληρο το αρχικό σύνολο εκπαίδευσης

Υπερπαράμετροι και cross validation

- Να σημειώσουμε ότι η μέθοδος cross validation με *k-folds* εκτελεί *k* διαφορετικές αποτιμήσεις των συνόλων εκπαίδευσης και ελέγχου.
  - Η προηγούμενη διαδικασία ρύθμισης παραμέτρων με τη χρήση συνόλων επικύρωσης πρέπει να εφαρμοστεί ξεχωριστά για κάθε ένα από τα *k* σύνολα εκπαίδευσης!
- Αυτό σημαίνει ότι όταν η ρύθμιση υπερπαραμέτρων εφαρμοστεί τότε *k* διαφορετικές τιμές υπερπαραμέτρων μπορούν να επιλεγούν
  - Αυτό δεν αποτελεί πρόβλημα διότι η ρύθμιση υπερπαραμέτρων είναι μέρος της διαδικασίας μάθησης
  - Η μέθοδος cross validation αποτιμά την ποιότητα της διαδικασίας μάθησης και όχι την ποιότητα ενός συγκεκριμένου μοντέλου

Υπερπαράμετροι και cross validation

- Τι συμβαίνει όταν τα δεδομένα εκπαίδευσης είναι λίγα έτσι ώστε οι εκτιμήσεις απόδοσης σε ένα σύνολο επικύρωσης να μην είναι αξιόπιστες;
- Χρήση εμφωλευμένης cross validation (υψηλό υπολογιστικό κόστος!)
  - Για κάθε σύνολο εκπαίδευσης για την «εξωτερική» (outer) cross validation *k*-folds εκτέλεσε "εσωτερικές" (inner) cross validations *p-folds* για την επιλογή της καλύτερης τιμής υπερ-παραμέτρου
  - Η εξωτερική cross validation χρησιμοποιείται για την εκτίμηση της ποιότητας της διαδικασίας μάθησης
  - Οι εσωτερικές cross validations χρησιμοποιούνται για την επιλογή των τιμών υπερπαραμέτρου και είναι μέρος της διαδικασίας μάθησης

Σύγκριση σχημάτων μηχανικής μάθησης

- Συχνή ερώτηση: ποιο από δύο σχήματα μάθησης έχει καλύτερη απόδοση;
- Αυτό εξαρτάται από το πεδίο εφαρμογής
- Προφανής τρόπος: σύγκρινε εκτιμήσεις 10-folds cross validation
- Εν γένει επαρκής για τις εφαρμογές (δεν χάνουμε πολύ αν η επιλεγμένη μέθοδος δεν είναι πραγματικά η καλύτερη)
- Όμως, η έρευνα στη μηχανική μάθηση πρέπει να επιδείξει πειστικά ότι μια συγκεκριμένη μέθοδος μπορεί να λειτουργήσει καλύτερα σε ένα πεδίο εφαρμογής από το οποίο λαμβάνονται δεδομένα

Σύγκριση σχημάτων μηχανικής μάθησης

- Θέλουμε να επιδείξουμε ότι το σχήμα Α είναι καλύτερο από το σχήμα Β σε ένα συγκεκριμένο πεδίο εφαρμογής
  - Για ένα δεδομένο αριθμό από δεδομένα εκπαίδευσης
  - Κατά μέσο όρο κατά μήκος όλων των πιθανών συνόλων εκπαίδευσης για το πεδίο αυτό
- Ας υποθέσουμε ότι έχουμε ένα άπειρο αριθμό από δεδομένα στο πεδίο
- Τότε απλά μπορούμε να
  - δειγματοληπτήσουμε απεριόριστα πολλά σύνολα δεδομένων ενός συγκεκριμένου μεγέθους
  - υπολογίσουμε μια εκτίμηση cross validation για κάθε σύνολο δεδομένων για κάθε σχήμα
  - ελέγξουμε αν η μέση ακρίβεια για το σχήμα Α είναι καλύτερη από αυτή του σχήματος Β

Πρόβλεψη Πιθανοτήτων

- Το μέτρο απόδοσης μέχρι τώρα ο βαθμός επιτυχίας (success rate)
- Επίσης ονομάζεται *0-1 συνάρτηση απωλειών (loss function)*:

$$\sum_i error_i \text{ όπου } error_i = \begin{cases} 1, & \text{αν πρόβλεψη για περίπτωση } i \text{ είναι λάθος} \\ 0, & \text{αν πρόβλεψη για περίπτωση } i \text{ είναι σωστή} \end{cases}$$

- Οι πιο πολλοί ταξινομητές παράγουν πιθανότητες κλάσεων
- Ανάλογα με την εφαρμογή, μπορεί να επιθυμούμαι να ελέγξουμε την ακρίβεια των εκτιμήσεων πιθανοτήτων
- Η παραπάνω συνάρτηση επομένως δεν είναι κατάλληλη προς χρήση σε αυτές τις περιπτώσεις

Quadratic Loss Function

- *Έστω* $p_1 \ldots p_k$ είναι εκτιμήσεις πιθανότητας για μια περίπτωση και c ο δείκτης θέσης για την πραγματική κλάση της περίπτωσης
- $a_1 \ldots a_k = 0$, εκτός από την a_c που είναι 1
- Η τετραγωνική ζημία εκφράζεται ως εξής: $\sum_j (p_j - a_j)^2$

- Η αναμενόμενη τιμή για αυτήν ελαχιστοποιείται όταν $p_j = p_j^*$, όπου οι τελευταίες είναι αληθινές πιθανότητες

$$E\left[\sum_j (p_j - a_j)^2\right] = \sum_j (E[p_j^2] - 2E[p_j a_j] + E[a_j^2])$$
$$= \sum_j (p_j^2 - 2p_j p_j^* + p_j^*) = \sum_j ((p_j - p_j^*)^2 + p_j^*(1 - p_j^*))$$

Informational Loss Function

- Η έκφραση αυτής είναι –log(p_c), όπου c είναι ο δείκτης θέσης για την πραγματική κλάση της περίπτωσης
- Έστω ότι p_1^* … p_k^* είναι οι αληθινές πιθανότητες κλάσης
- Τότε η αναμενόμενη τιμή για τη loss function είναι:

$$-p_1^* \log_2 p_1 - p_2^* \log_2 p_2 - \ldots - p_k^* \log_2 p_k$$

- Η informational loss function ελαχιστοποιείται όταν $p_j = p_j^*$:

$$-p_1^* \log_2 p_1^* - p_2^* \log_2 p_2^* - \ldots - p_k^* \log_2 p_k^*$$

Loss Functions

- Ποια loss function να επιλέξουμε;
  - Η quadratic loss function λαμβάνει υπόψη όλες τις εκτιμήσεις πιθανότητας κλάσης για μια περίπτωση
  - Η informational loss function εστιάζει μόνο στην εκτίμηση πιθανότητας για την πραγματική κλάση

Υπολογισμός Κόστους

- Στην πράξη διαφορετικοί τύποι λαθών ταξινόμησης συχνά αντιστοιχούν σε διαφορετικά κόστη
- Παραδείγματα:
  - Προφίλ τρομοκρατών: "Δεν είναι τρομοκράτης" είναι σωστό στο 99.99…% των περιπτώσεων
  - Αποφάσεις έγκρισης δανείων
  - Ανίχνευση πετρελαιοκηλίδας
  - Διάγνωση λαθών
  - Καμπάνιες Διαφημιστικών Mail

Υπολογισμός Κόστους

Έστω ένα πρόβλημα με 2 κατηγορίες

- Παραδείγματα από την κατηγορία 0 = 9990
- Παραδείγματα από την κατηγορία 1 = 10

- Εάν ένα μοντέλο προβλέπει ότι κάθε τι ανήκει στην κατηγορία 0, η ακρίβεια είναι 9990/10000 = 99.9 %

- Η ακρίβεια σε αυτή την περίπτωση είναι παραπλανητική γιατί το μοντέλο αποτυγχάνει να ανιχνεύσει οποιοδήποτε από τα παραδείγματα που ανήκουν στην κατηγορία 1

Υπολογισμός κόστους

Confusion matrix :

		Predicted class	
		Yes	No
Actual class	Yes	True positive	False negative
	No	False positive	True negative

True Positives - Σωστές προβλέψεις αληθινών γεγονότων.

False Positives - Λανθασμένες προβλέψεις αληθινών γεγονότων

True Negatives- Σωστές προβλέψεις ψευδών γεγονότων

False Negatives- Λανθασμένες προβλέψεις ψευδών γεγονότων.

Υπολογισμός κόστους

Confusion matrix :

		Predicted class	
		Yes	No
Actual class	Yes	True positive	False negative
	No	False positive	True negative

Διαφορετικά κόστη λανθασμένης ταξινόμησης μπορούν να αντιστοιχιστούν στα λανθασμένα θετικά και λανθασμένα αρνητικά

- Sensitivity : είναι ένα μέτρο του πόσο καλά ένα τεστ μπορεί να εντοπίσει τα πραγματικά θετικά αποτελέσματα

$TPR = TP / (TP + FN)$

- Specificity: είναι ένα μέτρο του πόσο καλά ένα τεστ μπορεί να εντοπίσει τα πραγματικά αρνητικά αποτελέσματα.

$TNR = TN / (FP + TN)$

Υπολογισμός κόστους

Confusion matrix :

		Predicted class	
		Yes	No
Actual class	Yes	True positive	False negative
	No	False positive	True negative

False positive rate: Το ποσοστό των αρνητικών παραδειγμάτων που ταξινομούνται λάθος (δηλαδή, ως θετικά)

$$\mathrm{FPR} = \frac{FP}{TN + FP}$$

False negative rate: Το ποσοστό των θετικών παραδειγμάτων που ταξινομούνται λάθος (δηλαδή, ως αρνητικά)

$$\mathrm{FNR} = \frac{FN}{TP + FN}$$

Kappa statistic

Τρέχουσα μέθοδος

τυχαία μέθοδος (δεξιά)

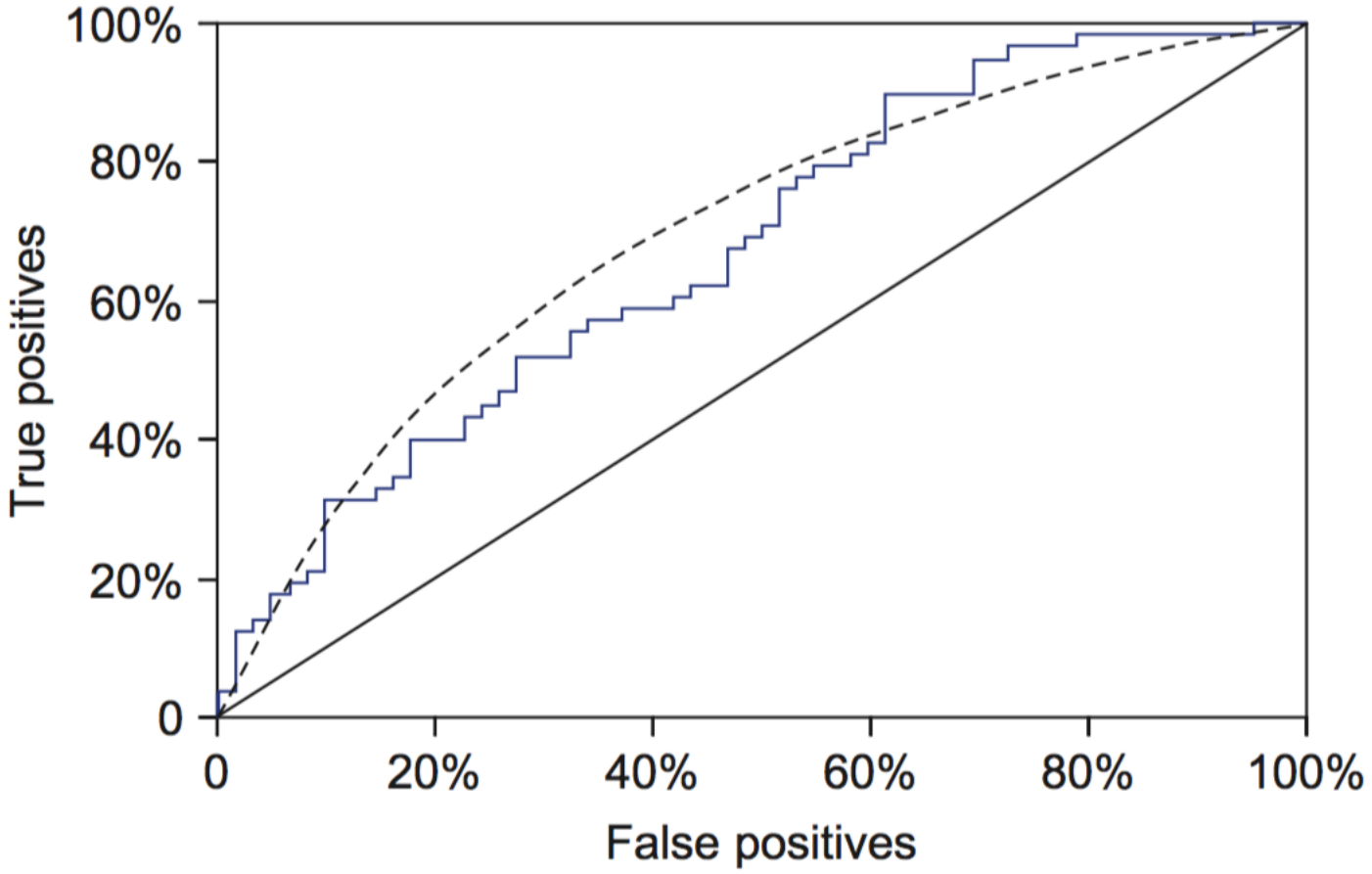
	Εκτιμώμενη Κλάση					Εκτιμώμενη Κλάση				
(A) Πραγματική Κλάση		a	b	c	Συν.	(B) Πραγματική κλάση		a	b	c	Συν.
	a	88	10	2	100		a	60	30	10	100
	b	14	40	6	60		b	36	18	6	60
	c	18	10	12	40		c	24	12	4	40
	Συν.	120	60	20			Συν.	120	60	20	

- Αριθμός επιτυχιών: αριθμός καταχωρήσεων στη διαγώνιο (*D*)
- Στατιστική *Kappa*: (αναλογία λαθών για την τρέχουσα μέθοδο – αναλογία λαθών για την τυχαία) / (1 – αναλογία λαθών για την τυχαία)
- Μετράει τη σχετική βελτίωση σε σχέση με την τυχαία μέθοδο: 1 σημαίνει ιδανική ακρίβεια και 0 ότι δεν υπάρχει καμία βελτίωση

Καμπύλες ROC

- Το ROC σημαίνει “receiver operating characteristic” (λειτουργικό χαρακτηριστικό λήπτη)
- Χρησιμοποιείται στην ανίχνευση σήματος για να επιδείξει τον συμβιβασμό (tradeoff) μεταξύ της αναλογίας των hits και της αναλογίας λανθασμένων συναγερμών πάνω από ένα θορυβώδες κανάλι
- Ο άξονας *y* προσδιορίζει το ποσοστό των αληθινών θετικών (true positives) στο δείγμα
- Ο άξονας x προσδιορίζει το ποσοστό των λανθασμένων θετικών (false positives) στο δείγμα

Παράδειγμα καμπύλης ROC

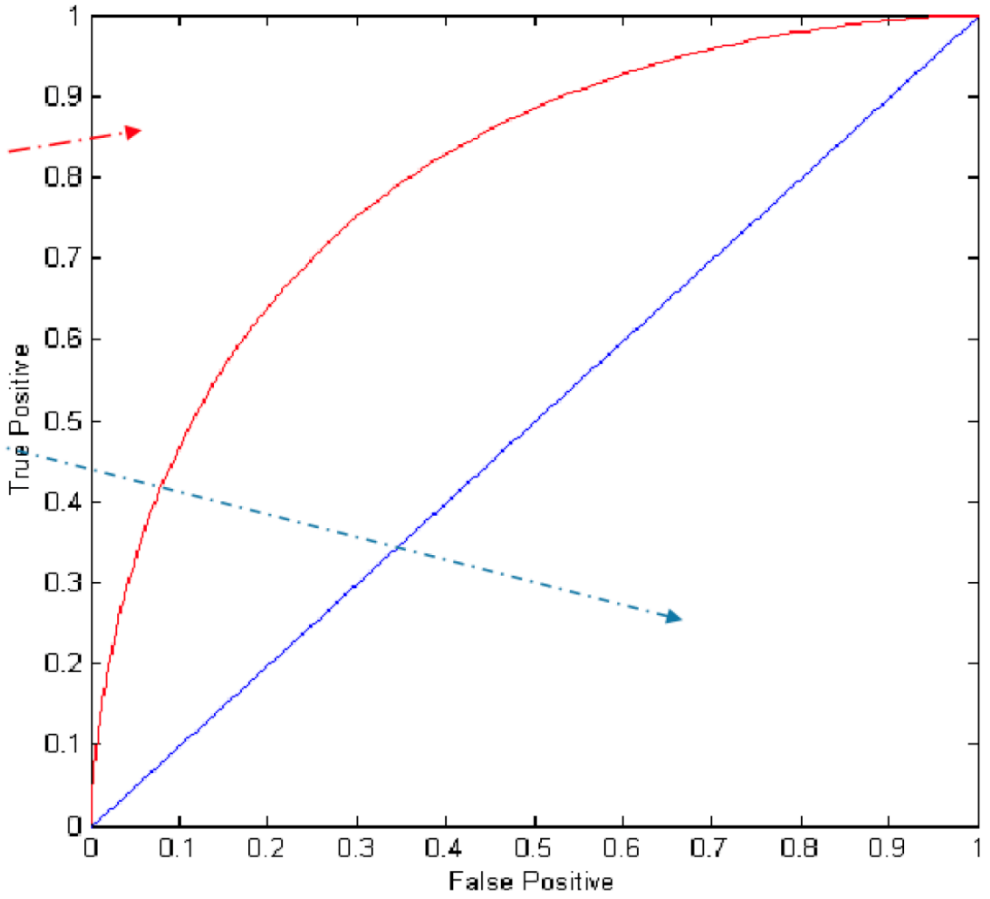


- Μπλέ καμπύλη Jagged— ένα test set
- Μαύρη καμπύλη — cross validation

Παράδειγμα καμπύλης ROC

Καλοί ταξινομητές κοντά στην αριστερή πάνω γωνία του διαγράμματος

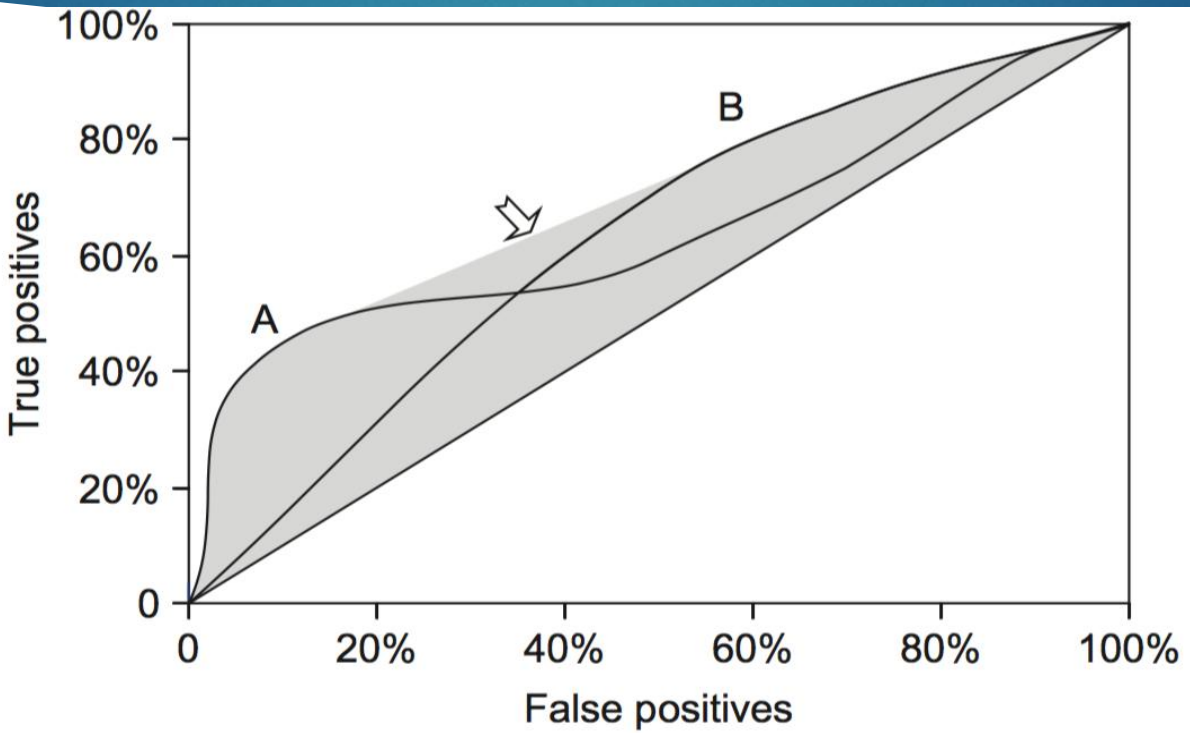
Κάτω από τη διαγώνιο Πρόβλεψη είναι το αντίθετο της πραγματικής κλάσης



Cross Validation & ROC curves

- Απλή μέθοδος για την λήψη για καμπύλης ROC μέσω της κατά μήκους επικύρωσης:
  - Συλλογή πιθανοτήτων για περιπτώσεις σε folds ελέγχου
  - Ταξινόμηση περιπτώσεων με βάση τις πιθανότητες
- Όμως αποτελεί απλώς μια περίπτωση
  - Μια άλλη δυνατότητα είναι η δημιουργία μιας καμπύλης ROC για κάθε fold και έπειτα να υπολογιστεί ο μέσος όρος

Καμπύλες ROC για δύο σχήματα μάθησης



- Για ένα μικρό εστιασμένο δείγμα, χρησιμοποίησε τη μέθοδο Α
- Για ένα μεγαλύτερο τη Β
- Ανάμεσα, διάλεξε μεταξύ Α και Β ανάλογα με τις αντίστοιχες πιθανότητες

Επιπλέον Μετρικές

- *Ακρίβεια - Precision*=TP/(TP+FP)

Πόσα από τα παραδείγματα που ο ταξινομητής έχει ταξινομήσει ως θετικά είναι πραγματικά θετικά. Όσο πιο μεγάλη η ακρίβεια, τόσο μικρότερος ο αριθμός των FP

- Ανάκληση - *Recall* =TP/(TP+FN)

Πόσα από τα θετικά παραδείγματα κατάφερε ο ταξινομητής να βρει Όσο πιο μεγάλη η ανάκληση, τόσο λιγότερα θετικά παραδείγματα έχουν ταξινομεί λάθος (=TPR - Sensitivity)

- Οι καμπύλες Precision/Recall έχουν σχήμα υπερβολής (hyperbolic shape)

Επιπλέον Μετρικές

- *F-measure*=(2 × recall × precision)/(recall+precision)

Τείνει να είναι πιο κοντά στο μικρότερο από τα δύο. Υψηλή τιμή σημαίνει ότι και τα δύο είναι ικανοποιητικά μεγάλα

- Περιοχή κάτω από τη καμπύλη ROC (*AUC*): πιθανότητα μια τυχαίως επιλεγμένη θετική περίπτωση να ταξινομηθεί πάνω από μια επιλεγμένη αρνητική

Αξιολόγηση Αριθμητικής Πρόβλεψης

- Ίδιες στρατηγικές: ανεξάρτητο σύνολο ελέγχου, cross validation, έλεγχοι σημαντικότητας
- Διαφορά: μετρικές λάθους
- Πραγματικές τιμές στόχου: $a_1\ a_2\ \ldots a_n$
- Εκτιμώμενες τιμές στόχου: $p_1\ p_2\ \ldots\ p_n$
- Πιο δημοφιλής μετρική: μέσο τετραγωνικό λάθος

$$\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n}$$

  - Εύκολα διαχειρίσιμο με μαθηματικό τρόπο

Άλλες Μετρικές

- Η ρίζα του μέσου τετραγωνικού λάθους (Root mean-squared error) :

$$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$$

- Το μέσο απόλυτο λάθος (mean absolute error) που είναι λιγότερο ευαίσθητο σε outliers σχέση με το τετραγωνικό λάθος:

$$\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n}$$

Η αρχή MDL

- MDL σημαίνει *minimum description length* (ελάχιστο μέγεθος περιγραφής)
- Το μέγεθος περιγραφής ορίζεται ως:

  απαιτούμενος χώρος περιγραφής μιας θεωρίας
  +
  απαιτούμενος χώρος περιγραφής των σφαλμάτων της θεωρίας
- Στη δική μας περίπτωση η θεωρία είναι ο ταξινομητής και τα σφάλματα είναι τα λάθη στα δεδομένα εκπαίδευσης
- Στόχος: εύρεση ταξινομητή με ελάχιστο DL (μέγεθος περιγραφής)
- Η αρχή MDL είναι ένα κριτήριο επιλογής μοντέλων
  - Επιτρέπει την επιλογή ενός ταξινομητή με μια κατάλληλη πολυπλοκότητα για την αντιμετώπιση του overfitting

Κριτήρια επιλογής μοντέλων

- Τα κριτήρια επιλογής μοντέλων προσπαθούν να βρουν ένα κατάλληλο σημείο συμβιβασμού μεταξύ:
  - Της πολυπλοκότητας του μοντέλου
  - Της ακρίβειας πρόβλεψής του στα δεδομένα εκπαίδευσης
- Λογική: ένα καλό μοντέλο είναι απλό και επιτυγχάνει υψηλή ακρίβεια στα τρέχοντα δεδομένα
- Αυτό ονομάζεται και ως το ξυράφι του *Occam*: η καλύτερη θεωρία είναι η μικρότερη που περιγράφει όλα τα γεγονότα

Κομψότητα εναντίων λαθών

- Θεωρία 1: πολύ απλή και κομψή θεωρία που περιγράφει σχεδόν ιδανικά τα δεδομένα
- Θεωρία 2: αρκετά πιο πολύπλοκη θεωρία που αναπαράγει τα δεδομένα χωρίς σφάλματα
- Η θεωρία 1 είναι πιθανώς προτιμότερη
- Κλασικό παράδειγμα: Οι τρεις νόμοι του Kepler για την κίνηση πλανητών
  - Λιγότερη ακριβής σε σχέση με τη τελευταία τροποποίηση του Κοπέρνικου στη Πτολεμαϊκή θεωρία των επικύκλων στα τρέχοντα διαθέσιμα δεδομένα

MDL και Συμπίεση

- Η αρχή MDL συσχετίζεται με τη συμπίεση δεδομένων:
  - Η καλύτερη θεωρία είναι αυτή που συμπιέζει τα δεδομένα το περισσότερο
  - Στην επαγωγική μάθηση, για τη συμπίεση των labels σε ένα σύνολο δεδομένων, παράγουμε ένα μοντέλο και έπειτα το αποθηκεύουμε μαζί με τα σφάλματά του
- Πρέπει να υπολογίσουμε:
  (α) το μέγεθος του μοντέλου και
  (β) το χώρο που χρειάζεται για την κωδικοποίηση των λαθών
- Το (β) είναι εύκολο: χρήση της informational loss function
- Για το (α) χρειαζόμαστε μια μέθοδο για την κωδικοποίηση του μοντέλου

Χρήση validation set για επιλογή μοντέλου

- Η αρχή MDL είναι ένα παράδειγμα κριτηρίου επιλογής μοντέλων
  - Επιλογή μοντέλων: εύρεση της κατάλληλης πολυπλοκότητας μοντέλου
- Κλασικό πρόβλημα επιλογής μοντέλων στη στατιστική:
  - Εύρεση του υποσυνόλου των ιδιοτήτων για χρήση στη γραμμική παρεμβολή
- Άλλα προβλήματα επιλογής μοντέλων: επιλογή μεγέθους δένδρου αποφάσεων ή ενός δικτύου ιδεατών νευρώνων
- Υπάρχουν πολλά κριτήρια επιλογής μοντέλων με βάση μια ποικιλία από θεωρητικές υποθέσεις
- Απλή προσέγγιση επιλογής μοντέλου: χρήση validation set
  - Χρήση της πολυπλοκότητας μοντέλου που επιτυγχάνει την καλύτερη απόδοση για το validation set
- Εναλλακτική προσέγγιση αν τα δεδομένα είναι λίγα: εσωτερική cross validation

Μηχανική Μάθηση & Εξόρυξη Γνώσης

Ερωτήσεις

?

Βιβλιογραφία

- Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, Η. Σακελλαρίου, Τεχνητή Νοημοσύνη - Γ' Έκδοση, ISBN: 978-960-8396-64-7, Έκδοση/Διάθεση: Εκδόσεις Πανεπιστημίου Μακεδονίας, 2011
- Ian H. Witten and Eibe Frank. 2005. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Κ. Διαμαντάρας, Ι. Μπότσης, Μηχανική Μάθηση – Α' Έκδοση, ISBN: 978-960-461-955-5, Εκδόσεις Κλειδάριθμος, 2019
- P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining», Addison Wesley, 2006