

ΕΜΠ ΔΠΜΣ

Εφαρμοσμένες Μαθηματικές Επιστήμες
Αλγόριθμοι Εξόρυξης Πληροφορίας

Διάλεξη 03:
**Προεπεξεργασία & Επιλογή
Δεδομένων**



Προεπεξεργασία δεδομένων

- Ο μετασχηματισμός των δεδομένων σε μορφή κατάλληλη και αποδοτική για την επιλεγμένη (?) μέθοδο μάθησης
- Σύνολο τεχνασμάτων που εφαρμόζονται με κύριο στόχο την αύξηση του βαθμού αξιοπιστίας
- Κάποιες φορές λειτουργούν, κάποιες όχι, και, μέχρι σήμερα, είναι δύσκολο να γνωρίζει κανείς εκ των προτέρων την αποτελεσματικότητά τους
- Καθώς ο πλέον αξιόπιστος οδηγός είναι η μέθοδος 'δοκιμής & σφάλματος' ('trial & error'), η γνώση και κατανόηση των τεχνασμάτων αυτών είναι εξαιρετικά σημαντική



Τεχνοτροπίες προεπεξεργασίας δεδομένων



- Επιλογή χαρακτηριστικών (attributes)
- Διακριτοποίηση χαρακτηριστικών
- Μετασχηματισμός δεδομένων
- Καθαρισμός δεδομένων



Επιλογή χαρακτηριστικών

- Στις περισσότερες εφαρμογές, ο αριθμός των χαρακτηριστικών είναι υπερβολικά μεγάλος για τον αποτελεσματικό χειρισμό τους από τα σχήματα εκμάθησης
- Τα περισσότερα χαρακτηριστικά – συχνά η συντριπτική πλειοψηφία τους – είναι ευκρινώς μη συσχετιζόμενα ή περιττά.
- Κατά συνέπεια, πρέπει να επιλεγεί υποσύνολο των δεδομένων προς χρήση στη διαδικασία εκμάθησης
- Οι μέθοδοι εκμάθησης εκ φύσεως επιτελούν τη διαδικασία επιλογής των κατάλληλων χαρακτηριστικών και απόρριψης των υπολοίπων
- Ωστόσο η αποδοτικότητά τους μπορεί συχνά να βελτιωθεί μέσω της προεπιλογής



Διακριτοποίηση χαρακτηριστικών

- Απολύτως αναγκαία στην περίπτωση που κάποια χαρακτηριστικά είναι αριθμητικά αλλά η επιλεγμένη μέθοδος εκμάθησης μπορεί να χειριστεί μόνο ρητά (categorical) χαρακτηριστικά
- Αιόμα και οι μέθοδοι που μπορούν να χειριστούν αριθμητικά χαρακτηριστικά πολλές φορές παράγουν καλύτερα αποτελέσματα ή λειτουργούν ταχύτερα αν τα χαρακτηριστικά έχουν διακριτοποιηθεί
- Το αντίστροφο πρόβλημα, στο οποίο κατηγορικά χαρακτηριστικά μετατρέπονται σε αριθμητικά, προκύπτει επίσης, αν και λιγότερο συχνά



Μετασχηματισμός δεδομένων

- Η μορφή των δεδομένων, είτε η φύση του αλγορίθμου εκμάθησης, συχνά υποδεικνύουν διάφορους μετασχηματισμούς των δεδομένων
- Αναγκαίοι για την ανάδειξη ιδιαιτεροτήτων ή και διαφορετικών γωνιών θέασης του συνόλου των δεδομένων
- Πλήθος τεχνικών
 - Μαθηματικοί μετασχηματισμοί
 - Βασισμένοι στη γνώση πεδίου
 - Λογικοί μετασχηματισμοί
 - Αλλαγή δομής / μορφής δεδομένων



Καθαρισμός δεδομένων

- Τα ακάθαρτα δεδομένα συνιστούν σημαντικό πρόβλημα
- Η αναγκαιότητα εξοικείωσης με τα δεδομένα έχει ήδη υποδειχθεί:
 - Κατανόηση περιεχομένου χαρακτηριστικών, συμβάσεων κωδικοποίησης, σημαντικότητας άγνωστων ή και επαναλαμβανόμενων τιμών, θορύβου μέτρησης, λαθών καταγραφής, συστηματικών λαθών
- Απλές μέθοδοι οπτικοποίησης συχνά βοηθούν κατά πολύ
- Αυτοματοποιημένες μέθοδοι καθαρισμού δεδομένων, εντοπισμού τιμών προς εξαίρεση και ανωμαλιών



Επιλογή χαρακτηριστικών

- Οι περισσότεροι αλγόριθμοι εκμάθησης έχουν σχεδιαστεί ώστε να επιλέγουν τα πλέον κατάλληλα χαρακτηριστικά για τη διαμόρφωση των αποφάσεών τους
- Η ύπαρξη περισσότερων χαρακτηριστικών επομένως οδηγεί – κατά τη θεωρία – σε περισσότερο αποδοτική εκπαίδευση
- Ωστόσο:
 - "Ποια η διαφορά μεταξύ θεωρίας και πράξης;
Δεν υπάρχει διαφορά
- στη θεωρία. Αλλά στην πράξη υπάρχει."*
- Στην πράξη, η προσθήκη μη σχετικών χαρακτηριστικών συχνά 'συγχύζει' τα συστήματα μηχανικής μάθησης



Επιλογή χαρακτηριστικών

- Βέλτιστη μέθοδος επιλογής: χειροκίνητη, βασισμένη σε βαθιά κατανόηση του προβλήματος και των εκφραζόμενων από τα χαρακτηριστικά
- Αυτοματοποιημένες μέθοδοι: επίσης χρήσιμες
- Η μείωση των διαστάσεων των δεδομένων
 - Βελτιώνει την απόδοση των αλγορίθμων εκπαίδευσης
 - Παρέχει μία περισσότερο συμπαγή και κατανοητή απεικόνιση της αντίληψης – στόχου, εστιάζοντας στις συναφείς μεταβλητές



Επιλογή χαρακτηριστικών

- Δύο μέθοδοι επιλογής υποσυνόλου χαρακτηριστικών
 - Μέθοδος διήθησης (*filter*): ανεξάρτητη αποτίμηση, βασισμένη στα γενικά χαρακτηριστικά των δεδομένων
 - Μέθοδος ενσωμάτωσης (*wrapper*): προσιόληση της διαδικασίας επιλογής στη διαδικασία εκπαίδευσης και αποτίμηση του υποσυνόλου με βάση την τελική απόδοση του αλγορίθμου εκμάθησης
- Δεν υπάρχει οικουμενικά αποδεικτό μέτρο της σχετικότητας ενός χαρακτηριστικού, γεγονός που καθιστά την επιλογή υποσυνόλου δύσκολη διαδικασία



Μέθοδος διήθησης

- Εύρεση υποσυνόλου χαρακτηριστικών επαρκούς για το διαχωρισμό όλων των υποδειγμάτων
 - Επιλογή μικρότερου δυνατού συνόλου χαρακτηριστικών που ικανοποιεί την απαίτηση αυτή (μέσω εξαντλητικής (exhaustive) αναζήτησης)
- Χρήση διαφορετικού αλγορίθμου εκμάθησης (για παράδειγμα C4.5, 1R) για επιλογή χαρακτηριστικών
 - Για παράδειγμα, μπορεί να εφαρμοστεί ένα αλγόριθμος δένδρου απόφασης στο σύνολο δεδομένων και στη συνέχεια να επιλεγούν τα χαρακτηριστικά εκείνα που ενυπάρχουν στο δένδρο και μόνο αυτά.
- Μάθηση βασισμένη στα υποδείγματα για απόδοση βαρύτητας στα χαρακτηριστικά (Instance based Learning)
 - Επίσης εφαρμόσιμη, ωστόσο αδυνατεί να υποδείξει τα περιττά χαρακτηριστικά



Αναζήτηση στο χώρο των χαρακτηριστικών

- Αριθμός των υποσυνόλων χαρακτηριστικών
 - ⇒ ειθδικός ως προς τον αριθμό των χαρακτηριστικών
 - ⇒ εξαντλητική αναζήτηση: μη πρακτική & κοστοβόρα
- Συνήθειες προσεγγίσεις: άπληστη (*greedy*) αναζήτηση
 - Επιλογή προς τα εμπρός (*forward selection*)
 - Απαλοιφή προς τα πίσω (*backward elimination*)
- Περισσότερο εξελιγμένες στρατηγιές:
 - Αναζήτηση διπλής κατεύθυνσης (*bidirectional*)
 - Αναζήτηση καλύτερου πρώτου (*best-first*): δύναται να εντοπίσει τη βέλτιστη λύση
 - Αναζήτηση δέσμης (*beam*): απλοποίηση της αναζήτησης καλύτερου πρώτου
 - Γενετικοί αλγόριθμοι (*genetic algorithms*)



Μέθοδος ενσωμάτωσης

- Ενσωμάτωση της διαδικασίας επιλογής στην εκμάθηση
 - Κριτήριο αποτίμησης: απόδοση διασταυρωμένης επικύρωσης (cross-validation)
- Δαπάνη χρόνου
 - Άπληστη αναζήτηση, k χαρακτηριστικά $\Rightarrow k^2 \times$ χρόνος
 - Πρότερη κατάταξη χαρακτηριστικών \Rightarrow γραμμική στο k
 - Εξαντλητική αναζήτηση $\Rightarrow 2^k \times$ time
- *RaceSearch, Schemata*
- *RankSearch*
- Τελικό κριτή συνιστά το ύψος του σφάλματος σε νέα δεδομένα



Διακριτοποίηση χαρακτηριστικών

- Μερικοί αλγόριθμοι ταξινόμησης και ομαδοποίησης χειρίζονται μόνο ονομαστικά χαρακτηριστικά και δεν μπορούν να χειριστούν όσα βρίσκονται σε αριθμητική μορφή
- Η χρήση των αριθμητικών χαρακτηριστικών προϋποθέτει την προηγούμενη ‘διακριτοποίησή’ τους σε ξεχωριστά υποσύνολα τιμών
- Ακόμα και οι αλγόριθμοι μάθησης που μπορούν να χειριστούν αριθμητικά χαρακτηριστικά, συχνά λειτουργούν πολύ πιο γρήγορα όταν χρησιμοποιούνται μόνο με διακριτά χαρακτηριστικά
- Επομένως, ποιος ο βέλτιστος τρόπος διακριτοποίησης των αριθμητικών χαρακτηριστικών πριν τη διαδικασία εκμάθησης;
- Για παράδειγμα, εφαρμογή αλγορίθμου σε
 - Διακριτοποιημένο σε k –τιμές χαρακτηριστικό ή σε
 - $k - 1$ δυαδικά χαρακτηριστικά που κωδικοποιούν τα σημεία τομής



Διακριτοποίηση χωρίς επίβλεψη

- Διακριβωση διαστημάτων χωρίς γνώση της τάξης στην οποία ανήκει το υπόδειγμα
 - Μοναδική λύση κατά την ομαδοποίηση
- Δύο προσεγγίσεις:
 - Διαστήματα σταθερού εύρους (*equal-interval binning*)
 - Διαστήματα σταθερής συχνότητας (*equal-frequency binning*) (καλείται επίσης 'έξισορρόπηση ιστογράμματος' (*histogram equalization*))
- Υποδεέστερης αξιοπιστίας έναντι των τεχνικών με επίβλεψη σε εργασίες ταξινόμησης



Διακριτοποίηση με επίβλεψη



- Μέθοδος *εντροπίας*
- Θεωρείται ως η πλέον αποδοτική & αξιόπιστη
- Κατασκευή δένδρου απόφασης για τη διακριτοποίηση του χαρακτηριστικού
 - Χρήση του μεγέθους της εντροπίας ως κριτήριο διαχωρισμού των τιμών του αρχικού συνόλου
 - Αρχή ελαχίστου μήκους περιγραφής (*minimum description length* (MDL)) ως κριτήριο διακοπής



Παράδειγμα: Διακριτοποίηση χαρακτηριστικού θερμοκρασίας

Temperature	64	65	68	69	70	71	72	72	75	75	80	81	83	85
Play	Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

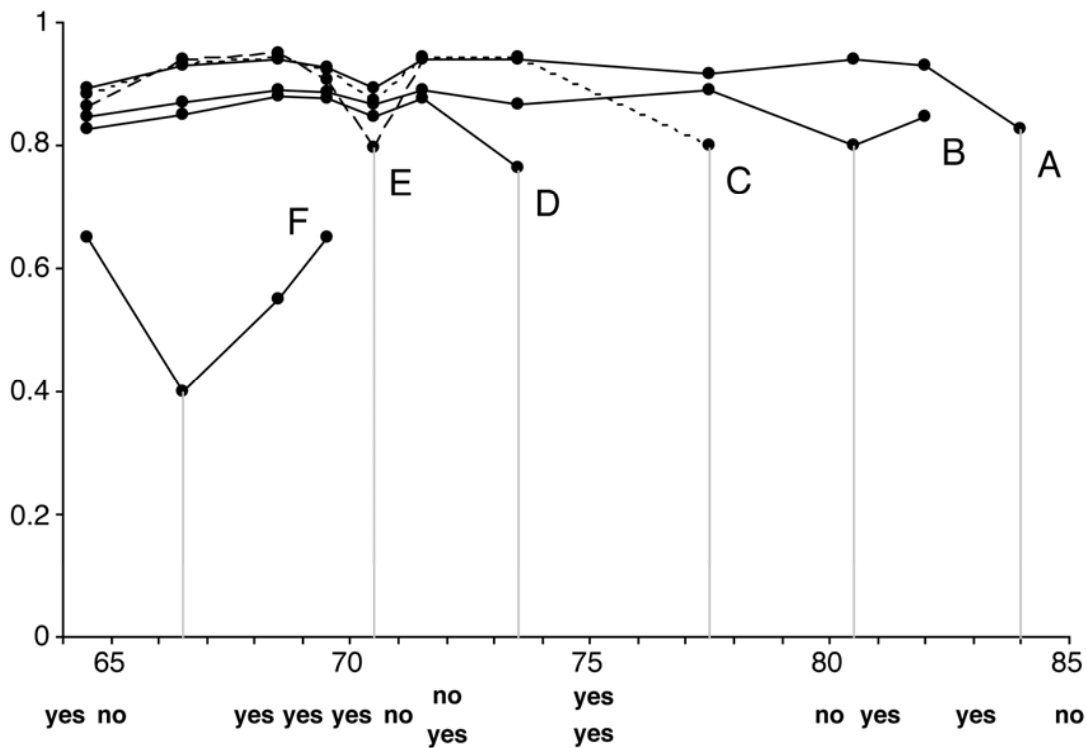
- Temperature $< 71,5 \rightarrow$ διαχωρισμός του εύρους σε
 - 4 yes & 2 no / 5 yes & 3 no
- Μέτρο κέρδους πληροφορίας βασισμένο στην εντροπία

$$\text{info} = -\sum_{i=1}^n p(i) \log_2(p(i))$$

- $\text{info}([4,2]) = -(4/(4+2)) * \log_2(4) - (2/(4+2)) * \log_2(2)$
- $\text{info}([4, 2],[5,3]) = (6/14) * \text{info}([4, 2]) + (8/14) * \text{info}([5, 3]) = 0.939 \text{ bits}$



Παράδειγμα: Διακριτοποίηση χαρακτηριστικού θερμοκρασίας



64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no yes	yes yes	no	yes	yes	no
		F			E		D	C	B		A
		66.5			70.5		73.5	77.5	80.5		84



Σχέση αρχής ελαχίστου μήκους περιγραφής

- Για την εφαρμογή της αρχής MDL :
 - Λαμβάνονται υπ' όψιν
 - Σημείο τομής ($\log_2[N - 1]$ bits)
 - Κατανομή τάξεων σε κάθε υποσύνολο
 - Σύγκριση απαιτούμενου μήκους περιγραφής πριν και μετά την προσθήκη σημείου τομής
- N υποδείγματα
 - Αρχικό σύνολο: k τάξεις, εντροπία E
 - Πρώτο υποσύνολο: k_1 τάξεις, εντροπία E_1
 - Δεύτερο υποσύνολο: k_2 τάξεις, εντροπία E_2

$$\text{info} > \frac{\log_2(N - 1)}{N} + \frac{\log_2(3^k - 2) - kE + k_1E_1 + k_2E_2}{N}$$

- Συνιστά τη μη διακριτοποίηση της θερμοκρασίας



Διακριτοποίηση με επίβλεψη: άλλες μέθοδοι

- Αντικατάσταση διαδικασίας από πάνω προς τα κάτω (top-down) με προσέγγιση από κάτω προς τα πάνω (bottom-up)
- Αντικατάσταση αρχής MDL με τεστ X^2
- Χρήση δυναμικού προγραμματισμού για την εύρεση βέλτιστου διαχωρισμού σε k -διαστήματα με βάση δοσμένο κριτήριο
 - Υπολογιστικό κόστος με κριτήριο εντροπίας $\Rightarrow N^2 \times \text{χρόνος}$
 - Υπολογιστικό κόστος με κριτήριο σφάλματος $\Rightarrow N \times \text{χρόνος}$



Σφάλμα ή εντροπία;

- Ερώτηση: Είναι δυνατόν, υποθέτοντας βέλτιστη διακριτοποίηση, δύο γειτονικά διαστήματα να περιέχουν υποδείγματα που ανήκουν κατά πλειοψηφία στην ίδια τάξη?
- Προφανής απάντηση: Όχι. Γιατί, εάν ναι,
 - Συγχώνευση των δύο
 - Απελευθέρωση ενός διαστήματος
 - Χρήση του άλλού*(προσέγγιση διακριτοποίησης με βάση κριτήριο σφάλματος)*
- Σωστή απάντηση: Ναι.
(η διακριτοποίηση με κριτήριο εντροπίας μπορεί να το υλοποιήσει)



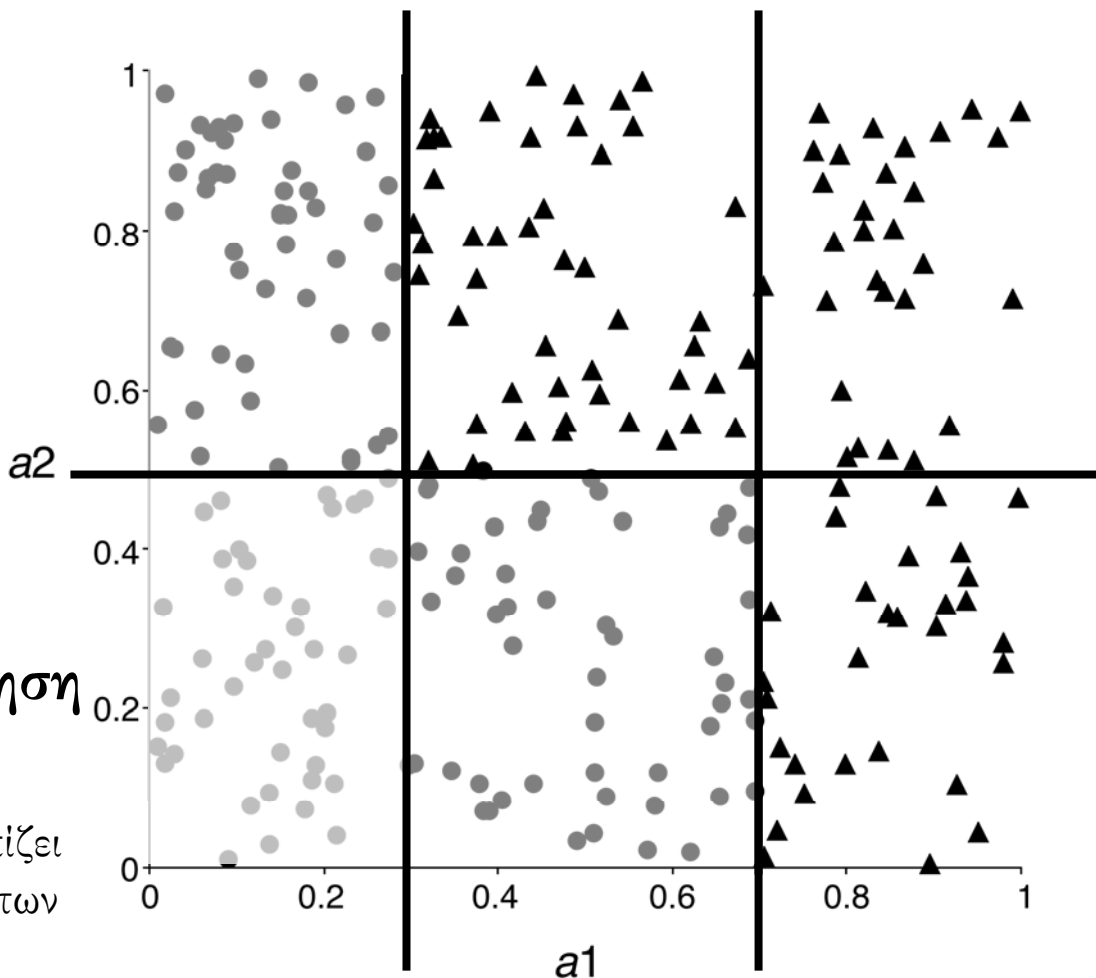
Σφάλμα ή εντροπία;

- Πρόβλημα δύο τάξεων, δύο χαρακτηριστικών

Αληθές μοντέλο

Βέλτιστη διακριτοποίηση των a_1, a_2

Η μέθοδος εντροπίας εντοπίζει μεταβολές στην κατανομή των τάξεων





Αντίστροφο πρόβλημα διακριτοποίησης

- Μετατροπή ονομαστικών τιμών σε 'αριθμητικές'
 1. Χαρακτηριστικό – δείκτης (indicator)
 - Δεν κάνει χρήση πληροφορίας περί κατάταξης
 2. Κωδικοποίηση ονομαστικού χαρακτηριστικού μέσω κατάταξης σε δυαδικά χαρακτηριστικά
 - Δύναται να χρησιμοποιηθεί από κάθε διατεταγμένο χαρακτηριστικό
 - Προτιμητέα έναντι μετατροπής σε αμέραιο (η οποία προϋποθέτει βαθμονόμηση)
- Γενικά: κωδικοποίηση υποσυνόλου των χαρακτηριστικών ως δυαδικά



Μετασχηματισμός δεδομένων

- Ένα πρόβλημα εξόρυξης γνώσης από δεδομένα σπάνια περιορίζεται στην απλή εφαρμογή ενός αλγορίθμου σε ένα σύνολο δεδομένων
- Μέθοδοι μετασχηματισμού των δεδομένων, όπως οι περι διακριτοποίησης, υπάρχουν διαθέσιμοι σε μεγάλο πλήθος και χρησιμοποιούνται σχεδόν σε κάθε περίπτωση
- Κάθε πρόβλημα είναι διαφορετικό
 - Απαιτείται κατανόηση των δεδομένων
 - Επίσης εξέτασή τους από διαφορετικές γωνίες θέασης με δημιουργικό τρόπο ώστε να φθάσει κανείς στην κατάλληλη οπτική
- Ο μετασχηματισμός σε διάφορες μορφές συμβάλλει προς αυτή την κατεύθυνση



Γενικά περί μετασχηματισμών

- Συχνά η φύση των δεδομένων υποδεικνύει μαθηματικούς μετασχηματισμούς
 - Για παράδειγμα, αφαίρεση ημερομηνιών για εύρεση ηλικίας
- Μετασχηματισμοί πιθανά να υποδεικνύονται επίσης από γνωστές ιδιότητες του αλγορίθμου εκμάθησης
 - Για παράδειγμα, κανονικοποίηση για βελτίωση απόδοσης
- Δεν είναι απαραίτητα μαθηματικοί, μπορούν για παράδειγμα να αποτυπώνουν την κοινή λογική ή τη γνώση πεδίου
 - Για παράδειγμα, ημερομηνίες αργιών, ατομικοί αριθμοί χημικών στοιχείων



Είδη μετασχηματισμού

- Συγχώνευση ονομαστικών χαρακτηριστικών
- Προσθήκη νέων χαρακτηριστικών
- Προσθήκη θορύβου στα δεδομένα (για παράδειγμα για έλεγχο αξιοπιστίας)
- Μεταλλαγή συγκεκριμένου ποσοστού
- Σύγχυση δεδομένων (*obfuscate*)
- Τυχαιοποίηση σειράς, παραγωγή τυχαίου υποσυνόλου
- Απομάκρυνση υποδειγμάτων (με ειδικά χαρακτηριστικά ή όχι)
- Απομάκρυνση τιμών προς εξαίρεση (*outliers*)
- Μετατροπή αραιών (*sparse*) δεδομένων σε μη αραιά και αντιστρόφως
- Μετασχηματισμοί κειμένου & χρονοσειρών



Ανάλυση Κύριων Συνιστωσών (Principal Components Analysis)

- Σύνολο δεδομένων με k αριθμητικά χαρακτηριστικά
 - Ένα νέφος σημείων σε ένα χώρο k -διαστάσεων – τα αστέρια στον ουρανό, ένα σμήνος μελισσών στο χώρο, ένα διάγραμμα διασποράς στο χαρτί. Τα χαρακτηριστικά αντιπροσωπεύουν τις συντεταγμένες.
- Η επιλογή αξόνων – το σύστημα συντεταγμένων – είναι αυθαίρετη
 - Η απεικόνιση του σμήνους είναι εξίσου καλή για κάθε επιλογή αξόνων
- Στην εξόρυξη γνώσης από δεδομένα, συχνά *υπάρχει* προτιμώμενο σύστημα συντεταγμένων
 - Καθορίζεται όχι από κάποια εξωτερική παραδοχή αλλά από τα ίδια τα δεδομένα



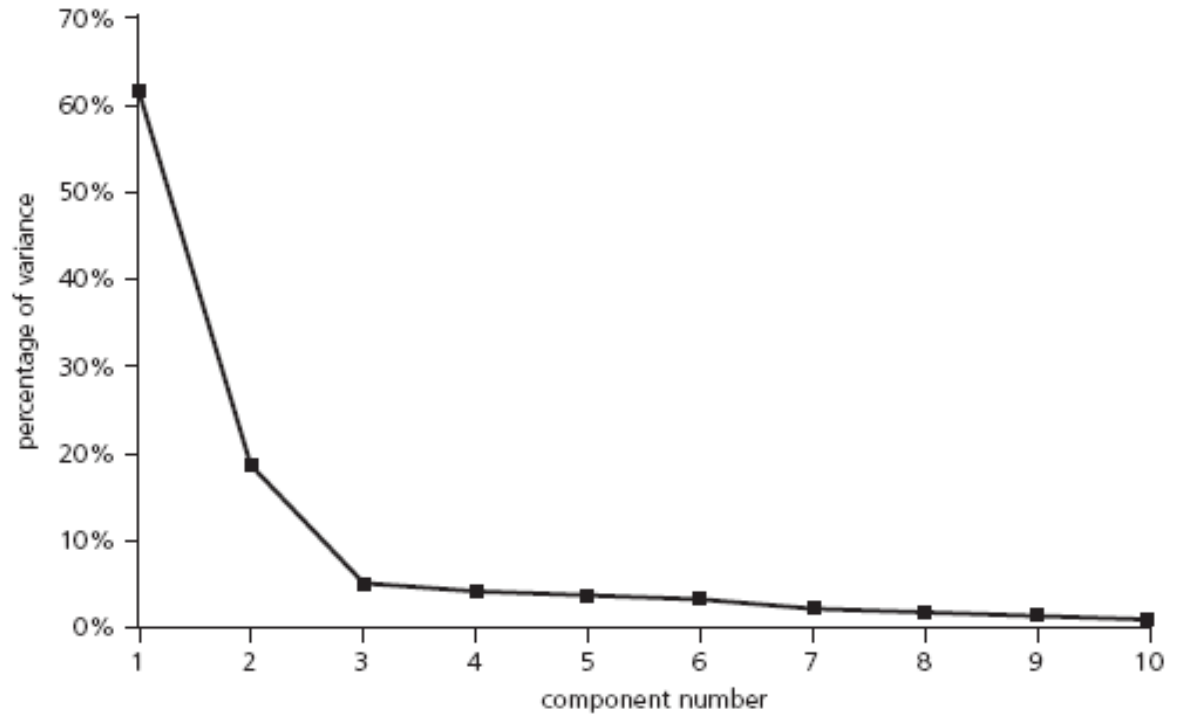
Ανάλυση Κύριων Συνιστωσών

- Για κάθε επιλογή συστήματος συντεταγμένων: το νέφος των σημείων χαρακτηρίζεται από συγκεκριμένη διακύμανση (variance) σε κάθε κατεύθυνση
 - δηλώνει τη διασπορά γύρω από τη μέση τιμή σε αυτή την κατεύθυνση
- Υπό τη συνθήκη ότι το σύστημα συντεταγμένων είναι *ορθογώνιο*, η συνολική διακύμανση είναι σταθερή.
- PCA: επέλεξε το σύστημα συντεταγμένων ως εξής
 - Τοποθέτησε τον πρώτο άξονα στην κατεύθυνση μέγιστης διακύμανσης
 - Επέλεξε τον δεύτερο άξονα κάθετα στον πρώτο, μεγιστοποιώντας πάλι τη διακύμανση
 - ...
 - Με μαθηματική διατύπωση: βρες τα *ιδιοδιανύσματα* του *διαγωνοποιημένου πίνακα συνδιακύμανσης* επί των αρχικών συντεταγμένων. αυτά αποτελούν του άξονες του μετασχηματισμένου χώρου και οι *ιδιοτιμές* αποδίδουν τη διακύμανση κατά μήκος των αξόνων.



Ανάλυση Κύριων Συνιστωσών Παράδειγμα

Axis	Variance	Cumulative
1	61.2%	61.2%
2	18.0%	79.2%
3	4.7%	83.9%
4	4.0%	87.9%
5	3.2%	91.1%
6	2.9%	94.0%
7	2.0%	96.0%
8	1.7%	97.7%
9	1.4%	99.1%
10	0.9%	100%





Ανάλυση Κύριων Συνιστωσών

- Η διακύμανση ανά άξονα (συνιστώσα / component) εκφράζεται ως % επί του (σταθερού) συνόλου
 - Για παράδειγμα, η πρώτη συνιστώσα ερμηνεύει το $x\%$ της συνολικής διασποράς
- Μπορεί κανείς να επιλέξει προς χρήση τις πιο σημαντικές (κύριες / principal) συνιστώσες –και μόνο αυτές– και να αποβάλλει τις υπόλοιπες
- Η μέθοδος χρησιμοποιείται συχνά πριν την εφαρμογή των αλγορίθμων εκπαίδευσης για τον καθαρισμό των δεδομένων και την αναπαραγωγή των χαρακτηριστικών
- Σημείωση: η κλίμακα των χαρακτηριστικών επηρεάζει το αποτέλεσμα της PCA
 - Κοινή πρακτική αποτελεί η κανονικοποίηση όλων των χαρακτηριστικών (μέση τιμή: 0, τυπική απόκλιση: 1) πριν την εφαρμογή της μεθόδου



Τυχαίες προβολές

- Υπολογιστικό κόστος PCA \Rightarrow ($\#$ διαστάσεων)³ \times χρόνος
- Απλούστερη εναλλακτική: τυχαία προβολή των δεδομένων σε υποχώρο προκαθορισμένου αριθμού διαστάσεων
- Όπως είναι αναμενόμενο, οι τυχαίες προβολές έχουν χαμηλότερη απόδοση της PCA
- Ωστόσο, πειραματικά αποτελέσματα υποδεικνύουν πως η απόκλιση δεν είναι πολύ μεγάλη και τείνει να μειωθεί καθώς ο αριθμός των διαστάσεων αυξάνεται



Μετατροπή κειμένου σε χαρακτηριστικά

- Χαρακτηριστικά που περιέχουν συμβολοσειρές (strings) κειμένου
 - Συχνά οι τιμές τους αποτελούν ολόκληρο κείμενο
 - Πώς αντιμετωπίζονται;
- Αποσύνθεση του κειμένου σε χαρακτηριστικά συμβολοσειρών
 - παραγράφους, προτάσεις, φράσεις ή – συνήθως – λέξεις
- Η μετατροπή σε λέξεις (*tokenization*) είναι πολυσύνθετη
 - Αριθμοί, σύνταξη, στίξη, διαχωριστής πεδίου (delimiter), γλωσσικές συμβάσεις, συχνές ή μη λέξεις, ...
- Οι συχνότητες f_{ij} της λέξης i στο έγγραφο j μπορούν να μετασχηματιστούν με διάφορους τρόπους
 - $\log(1 + f_{ij})$.
 - TF x IDF “term frequency times inverse document frequency.”
 - $f_{ij} \log \frac{\text{number of documents}}{\text{number of documents that include word } i}$



Μετατροπή χρονοσειρών σε χαρακτηριστικά



- Σε δεδομένα χρονοσειρών
 - κάθε υπόδειγμα αντιπροσωπεύει διαφορετική χρονική στιγμή
 - τα χαρακτηριστικά δίδουν τιμές συσχετιζόμενες με αυτό το στιγμιότυπο
 - για παράδειγμα πρόβλεψη καιρού, πρόβλεψη τιμών μετοχής
- Πολύ συνηθής είναι η αντικατάσταση των τιμών ενός χαρακτηριστικού με τη διαφορά (Δ) τους με αυτές ενός χρονικά προηγούμενου υποδείγματος
- Σε κάποιες περιπτώσεις, τα υποδείγματα δε λαμβάνονται περιοδικά, γι' αυτό και συνοδεύονται από χαρακτηριστικό ένδειξης χρόνου (*timestamp*)
 - Η διαφορά των timestamps αποτελεί το χρονικό βήμα του υποδείγματος
- Σε άλλες περιπτώσεις, κάθε χαρακτηριστικό αφορά διαφορετικό στιγμιότυπο, οπότε η χρονοσειρά ουσιαστικά προσδιορίζεται από το ένα προς το άλλο χαρακτηριστικό και όχι υπόδειγμα



Αυτοματοποιημένος καθαρισμός δεδομένων

- Η χαμηλή ποιότητα των δεδομένων αποτελεί σοβαρό πρόβλημα για την εξόρυξη γνώσης από αυτά
 - Τα λάθη σε μεγάλες βάσεις δεδομένων αποτελούν περισσότερο κανόνα παρά εξαίρεση
 - Οι τιμές των χαρακτηριστικών, πολλές φορές ακόμα και των τάξεων, είναι συχνά ανακριβείς, αναξιόπιστες και αιάθαρτες
- Συμβατική μέθοδος αντιμετώπισης: προσεκτικός έλεγχος των δεδομένων
 - Στην έκταση των προβλημάτων που πραγματεύονται, ίσως αδύνατος
- Αυτοματοποιημένες μέθοδοι:
 - Εντοπισμός διαταραχών
 - Συνδυασμός αλγορίθμων μάθησης...



Βελτίωση δένδρων απόφασης

- Για τη βελτίωση ενός δένδρου απόφασης:
 - Ιδεατή λύση: έλεγχος λανθασμένα ταξινομημένων υποδειγμάτων από εμπειρογνώμονα
 - Προσέγγιση μηχανική μάθησης στο πρόβλημα: απομάκρυνση λανθασμένα ταξινομημένων υποδειγμάτων και επανεκπαίδευση
- Θόρυβος χαρακτηριστικών και θόρυβος τάξεων
 - Ο θόρυβος στα χαρακτηριστικά πρέπει να παραμένει ανέπαφος στο σύνολο δεδομένων (*προσοχή: όχι εκπαίδευση σε 'καθαρό' σύνολο και έλεγχος σε 'μη καθαρό' σύνολο*)
 - Συστημικός θόρυβος στην τάξη (για παράδειγμα αντικατάσταση μίας τάξης με μία άλλη): πρέπει επίσης να παραμείνει στα δεδομένα εκπαίδευσης
 - Μη συστημικός θόρυβος τάξης: απομάκρυνσή του από το σύνολο εκπαίδευσης, αν αυτό είναι δυνατό

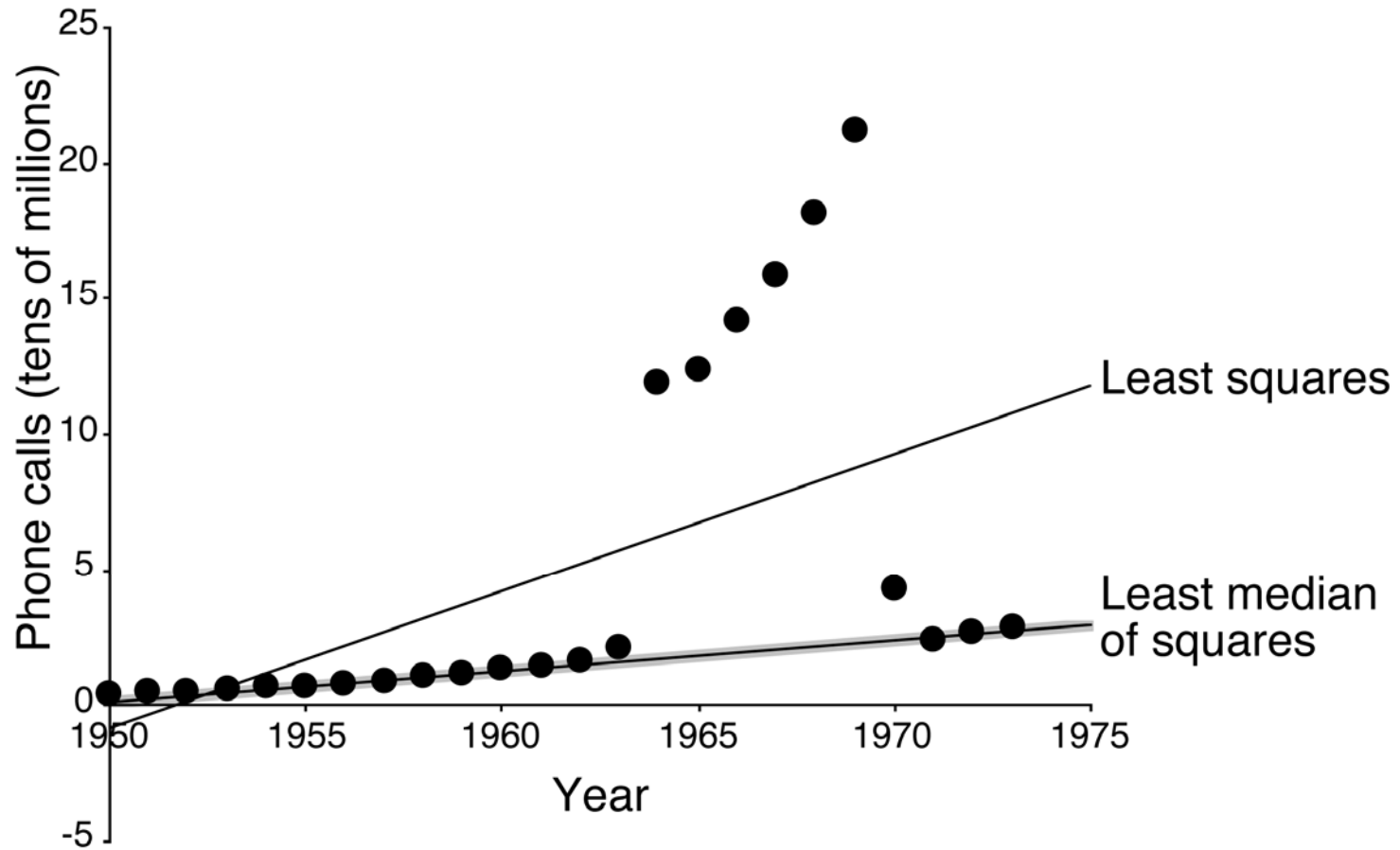


Ανθεκτική παλινδρόμηση

- Ανθεκτική (*robust*) στατιστική μέθοδος \Rightarrow επιλαμβάνεται του προβλήματος των τιμών προς εξαίρεση (*outliers*)
- Για την ανθεκτικότητα της παλινδρόμησης (*regression*):
 - Ελαχιστοποίηση του απολύτου σφάλματος, όχι του τετραγωνικού σφάλματος
 - Απομάκρυνση outliers (για παράδειγμα 10% των σημείων που απέχουν περισσότερο από την επιφάνεια παλινδρόμησης)
 - Ελαχιστοποίηση της μεσαίας τιμής (*median*) και όχι του μέσου (*mean*) των τετραγώνων (για outliers στον x και y άξονα)
 - Λαμβάνει υπ' όψιν το ήμισυ τα περισσότερα συναφή, σε όρους απόστασης, δεδομένα
 - Υψηλό υπολογιστικό κόστος



Παράδειγμα ανθεκτικής παλινδρόμησης





Εφαρμογή στο weka



Αποτίμηση & επιλογή χαρακτηριστικών



- Αναζήτηση στο χώρο των υποσυνόλων χαρακτηριστικών και αποτίμηση του καθενός
- Το περιβάλλον εργασίας παρέχει
 - 4 μεθόδους αποτίμησης υποσυνόλου χαρακτηριστικών
 - 7 μεθόδους αναζήτησης σε υποσύνολα χαρακτηριστικών
 - (άρα και 2401 συνδυασμούς τους!)
- Μία ταχύτερη αλλά λιγότερο ακριβής μέθοδος είναι η
 - αποτίμηση κάθε χαρακτηριστικού χωριστά
 - ταξινόμηση και απόρριψη όσων δεν ικανοποιούν ορισμένη τιμή του κριτηρίου
- Το περιβάλλον εργασίας παρέχει
 - 8 μεθόδους αποτίμησης μοναδιαίου χαρακτηριστικού
 - μέθοδο κατάταξης



Select attributes

Weka Explorer

Preprocess Classify Cluster Associate **Select attributes** Visualize

Attribute Evaluator
Choose **CfsSubsetEval**

Search Method
Choose **RandomSearch -F 0.0025**

Attribute Selection Mode
 Use full training set
 Cross-validation Folds: 10 Seed: 1

(Nom) class

Start Stop

Result list (right-click for options)

- 16:35:38 - RandomSearch + CfsSubsetEval
- 16:38:23 - RandomSearch + CfsSubsetEval
- 16:39:28 - RandomSearch + CfsSubsetEval

weka.gui.GenericObjectEditor

weka.attributeSelection.CfsSubsetEval

About

CfsSubsetEval : More

Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.

locallyPredictive: False

missingSeparate: False

Open... Save... OK Cancel

Attribute Subset Evaluator (supervised, Class (nominal): 5 class):
CFS Subset Evaluator

Selected attributes: 2,3,4 : 3
sepalwidth
petallength
petalwidth

Status
OK Log x 0

- Select attributes sheet
 - Choose & format *Attribute Evaluator* & *Search Method*
 - *Attribute Selection Mode* (full training set / cross-validation)
 - *Select attribute*
 - Start / stop
 - *Attribute selection output*
 - *Result list*



Αποτίμηση υποσυνόλου χαρακτηριστικών



Μέθοδοι διήθησης (filter methods)

- *CfsSubsetEval* Consider the predictive value of each subset evaluator attribute individually, along with the degree of redundancy among them
- *ConsistencySubsetEval* Project training set onto attribute set and measure consistency in class values

Μέθοδοι ενσωμάτωσης (wrapper methods)

- *ClassifierSubsetEval* Use a classifier to evaluate attribute set
- *WrapperSubsetEval* Use a classifier plus cross-validation



Μέθοδοι αναζήτησης



- *BestFirst* Greedy hill-climbing with backtracking
- *ExhaustiveSearch* Search exhaustively
- *GeneticSearch* Search using a simple genetic algorithm
- *GreedyStepwise* Greedy hill-climbing without backtracking; optionally generate ranked list of attributes
- *RaceSearch* Use race search methodology
- *RandomSearch* Search randomly
- *RankSearch* Sort the attributes and rank promising subsets using an attribute subset evaluator

Μέθοδος κατάταξης

- *Ranker* Rank individual attributes (not subsets) according to their evaluation



Αποτίμηση μοναδιαίου χαρακτηριστικού



- *ChiSquaredAttributeEval* Compute the chi-squared statistic of each attribute with respect to the class
- *GainRatioAttributeEval* Evaluate attribute based on gain ratio
- *InfoGainAttributeEval* Evaluate attribute based on information gain
- *OneRAttributeEval* Use OneR's methodology to evaluate attributes
- *PrincipalComponents* Perform principal components analysis and transformation
- *ReliefFAttributeEval* Instance-based attribute evaluator
- *SVMAttributeEval* Use a linear support vector machine to determine the value of attributes
- *SymmetricalUncertAttributeEval* Evaluate attribute based on symmetric uncertainty



Filter



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Undo Edit... Save...

Filter

weka

- filters
 - supervised
 - attribute
 - AttributeSelection
 - ClassOrder
 - Discretize
 - NominalToBinary
 - instance
 - Resample
 - SpreadSubsample
 - StratifiedRemoveFolds
 - unsupervised
 - attribute
 - instance

Selected attribute

Name: sepalength
Missing: 0 (0%) Distinct: 35 Type: Numeric
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom) Visualize All

- Preprocess sheet
 - *Filter*
 - Choose (supervised / unsupervised, attribute / instance)
 - Apply



Φίλτρα χαρακτηριστικών χωρίς επίβλεψη (Unsupervised attribute filters)



- Προσθήκη & απομάκρυνση χαρακτηριστικών
 - *Add* Add a new attribute, whose values are all marked as *missing*
 - *Copy* Copy a range of attributes in the dataset
 - *Remove* Remove attributes
 - *RemoveType* Remove attributes of a given type (nominal, numeric, string, or date)
 - *RemoveUseless* Remove constant attributes, along with nominal attributes that vary too much
 - Define attribute, eg. *first-3,5,9-last* / *except*



Φίλτρα χαρακτηριστικών χωρίς επίβλεψη (Unsupervised attribute filters)



- *AddCluster* Add a new nominal attribute representing the cluster assigned to each instance by a given clustering algorithm
- *AddExpression* Create a new attribute by applying a specified mathematical function to existing attributes
- *ClusterMembership* Use a clusterer to generate cluster membership values, which then form the new attributes
- *Normalize* Scale all numeric values in the dataset to lie within the interval $[0,1]$
- *Standardize* Standardize all numeric attributes to have zero mean and unit variance
- *NumericTransform* Transform a numeric attribute using any Java function
 - Operators $+$, $-$, $*$, $/$, $^$
 - functions *log/exp, abs/sqrt, floor/ceil/rint, sin/cos/tan*
 - Define attribute, eg. *A7*
 - Expression eg. $a1^2 * a5 / \log(a7 * 4.0)$



Φίλτρα χαρακτηριστικών χωρίς επίβλεψη (Unsupervised attribute filters)



- Αλλαγή τιμών
 - *SwapValues* Swap two values of an attribute
 - *MergeTwoValues* Merge two values of a given attribute: Specify the index of the two values to be merged
 - *ReplaceMissingValues* Replace all missing values for nominal and numeric attributes with the modes and means of the training data
- Τυχαιοποίηση
 - *AddNoise* Change a percentage of a given nominal attribute's values
 - *Obfuscate* Obfuscate the dataset by renaming the relation, all attribute names, and nominal and string attribute values
 - *RandomProjection* Project the data onto a lower-dimensional subspace using a random matrix



Φίλτρα χαρακτηριστικών χωρίς επίβλεψη (Unsupervised attribute filters)



- Μετατροπές
 - *Discretize* Convert numeric attributes to nominal: Specify which attributes, number of bins, whether to optimize the number of bins, and output binary attributes. Use equal-width (default) or equal-frequency binning
 - *PKIDiscretize* Discretize numeric attributes using equal-frequency binning, where the number of bins is equal to the square root of the number of values (excluding missing values)
 - *MakeIndicator* Replace a nominal attribute with a Boolean attribute. Assign value 1 to instances with a particular range of attribute values; otherwise, assign 0. By default, the Boolean attribute is coded as numeric
 - *NominalToBinary* Change a nominal attribute to several binary ones, one for each value
 - *NumericToBinary* Convert all numeric attributes into binary ones: Nonzero values become 1
 - *FirstOrder* Apply a first-order differencing operator to a range of numeric attributes



Φίλτρα χαρακτηριστικών χωρίς επίβλεψη (Unsupervised attribute filters)



- Μετατροπές συμβολοσειρών
 - *StringToNominal* Convert a string attribute to nominal
 - *StringToWordVector* Convert a string attribute to a vector that represents word occurrence frequencies; you can choose the delimiter(s)—and there are many more options
- Χρονοσειρές
 - *TimeSeriesTranslate* Replace attribute values in the current instance with the equivalent value in some previous (or future) instance
 - *TimeSeriesDelta* Replace attribute values in the current instance with the difference between the current value and the value in some previous (or future) instance



Φίλτρα υποδειγμάτων χωρίς επίβλεψη (Unsupervised instance filters)



- Τυχαίωση & δειγματοληψία υποσυνόλου
 - *Randomize* Randomize the order of instances in a dataset
 - *Normalize* Treat numeric attributes as a vector and normalize it to a given length
 - *Resample* Produce a random subsample of a dataset, sampling with replacement
 - *RemoveFolds* Output a specified cross-validation fold for the dataset
 - *RemovePercentage* Remove a given percentage of a dataset
 - *RemoveRange* Remove a given range of instances from a dataset
 - *RemoveWithValues* Filter out instances with certain attribute values
 - *RemoveMisclassified* Remove instances incorrectly classified according to a specified classifier—useful for removing outliers
- Αραιά Υποδείγματα
 - *NonSparseToSparse* Convert all incoming instances to sparse format
 - *SparseToNonSparse* Convert all incoming sparse instances into nonsparse format



Φίλτρα με επίβλεψη (Supervised Filters)



- Χρησιμοποιούν ιδιαίτερης προσοχής, καθώς στην πραγματικότητα δεν συνιστούν λειτουργίες προεπεξεργασίας
- Οι διαμερίσεις των δεδομένων ελέγχου απαιτείται να μη λαμβάνουν υπ' όψιν τις τιμές των τάξεων των υποδειγμάτων ελέγχου, καθώς αυτές υποτίθεται πως είναι άγνωστες.
- Φίλτρα χαρακτηριστικών
 - *Discretize* Convert numeric attributes to nominal
 - *NominalToBinary* Convert nominal attributes to binary, using a supervised method if the class is numeric
 - *ClassOrder* Randomize, or otherwise alter, the ordering of class values
 - *AttributeSelection* Provides access to the same attribute selection methods as the *Select attributes* panel
- Φίλτρα υποδειγμάτων
 - *Resample* Produce a random subsample of a dataset, sampling with replacement
 - *SpreadSubsample* Produce a random subsample with a given spread between class frequencies, sampling with replacement
 - *StratifiedRemoveFolds* Output a specified stratified cross-validation fold for the dataset



Τέλος

Επόμενη διάλεξη:
Απεικόνιση Γνώσης,
Αξιοπιστία & Αποτίμηση