Προεπεξεργασία δεδομένων

Ιωάννης Μαδεμλής

Δεδομένα

Σύνολα δεδομένων

- Τα **σύνολα δεδομένων** συνήθως δεν είναι κατάλληλα για άμεση επεξεργασία από έναν μηχανισμό/αλγόριθμο εξόρυξης γνώσης.
  - Π.χ., δεν ταιριάζουν με τις απαιτήσεις του, έχουν κακή ποιότητα, κλπ.
- Ως εκ τούτου, είναι απαραίτητο ένα βήμα μετασχηματισμού των δεδομένων πριν από την τροφοδότησή τους στον αλγόριθμο εξόρυξης (**προεπεξεργασία**).
- Μπορεί να είναι θεμιτό και ένα βήμα **μετεπεξεργασίας** των αποτελεσμάτων, προκειμένου να παρουσιαστεί ορθώς στο χρήστη η εξηγμένη νέα γνώση.

Δεδομένα

Σύνολα δεδομένων

- Κάθε σύνολο δεδομένων είναι ένα σύνολο **προτύπων/στοιχείων**, όπου κάθε πρότυπο είναι ένα διατεταγμένο σύνολο **γνωρισμάτων**.
- Τα διαφορετικά πρότυπα του ίδιου συνόλου μπορούν να έχουν διαφορετικές **τιμές** στο αντίστοιχο γνώρισμα.
- Άρα, γνώρισμα είναι μία ιδιότητα ενός στοιχείου η οποία μπορεί να αλλάζει από πρότυπο σε πρότυπο, ή από στιγμή σε στιγμή.
- Προφανώς, δύο πρότυπα τα οποία έχουν μεταξύ τους ίδιες τιμές σε κάθε τους γνώρισμα είναι πανομοιότυπα.

Δεδομένα

Τύποι γνωρισμάτων

- Πάντοτε, για να κατασκευάσουμε ένα πρότυπο το οποίο εκφράζει κάποιο φυσικό αντικείμενο επιλέγουμε κάποιες συμβάσεις μέτρησης ανά γνώρισμα.
- Αυτές ορίζουν τις δυνατές **συμβολικές** ή **αριθμητικές τιμές** του εκάστοτε γνωρίσματος.
- Η ίδια ιδιότητα του ίδιου φυσικού αντικειμένου την οποία το γνώρισμα αναπαριστά, μπορεί να αναπαρασταθεί με τελείως διαφορετικές συμβάσεις μέτρησης.

Τύποι γνωρισμάτων

- Πρόκειται για μία πολύ σημαντική επιλογή, διότι η κλίμακα μέτρησης μπορεί:
  - είτε να έχει κάποια χαρακτηριστικά τα οποία δεν διαθέτει στην πραγματικότητα η φυσική ιδιότητα την οποία προσπαθούμε να μετρήσουμε,
  - είτε το αντίστροφο.

Δεδομένα

- Παράδειγμα: σε ένα σύνολο εγγραφών με ιδιότητες των υπαλλήλων ενός οργανισμού, το πρώτο γνώρισμα είναι ένας αύξων αριθμός μετρούμενος ως ακέραιος.
  - Δεν έχει φυσικό νόημα η αναμενόμενη τιμή όλων των αυξόντων αριθμών, αν και είναι έγκυρη πράξη ακεραίων.
  - Η επιλεγμένη κλίμακα μέτρησης μας δίνει τη δυνατότητα να κάνουμε πράξεις χωρίς φυσικό νόημα για αυτό το γνώρισμα.

Δεδομένα

Τύποι γνωρισμάτων

- Τα ζητήματα αυτά περιγράφονται συστηματικά με τον **τύπο** κάθε γνωρίσματος.
- Τα γνωρίσματα μπορούν να είναι είτε **κατηγορικά**, είτε **αριθμητικά**.
  - Στα κατηγορικά γνωρίσματα δεν έχει νόημα η εφαρμογή αριθμητικών πράξεων.
  - Στα αριθμητικά γνωρίσματα έχει νόημα.
- Τα κατηγορικά δεδομένα είναι είτε **ονομαστικά** (π.χ., ένας αύξων αριθμός), είτε **διατακτικά** (π.χ., αριθμοί μίας οδού).
- Μπορούν να αναπαρασταθούν είτε με **συμβολικές** είτε με **αριθμητικές** τιμές.
  - Π.χ.: «ανοιχτό/κλειστό» ή «0/1».

Τύποι γνωρισμάτων

- Στα ονομαστικά δεδομένα έχει νόημα μόνον η εφαρμογή *τελεστών ισότητας/ανισότητας*.
  - Π.χ., ο υπάλληλος Α είναι διαφορετικός από τον υπάλληλο Β, γιατί έχουν *διαφορετικό* αύξοντα αριθμό.
- Στα διατακτικά δεδομένα εφαρμόζονται επιπροσθέτως και *τελεστές σύγκρισης*.
  - Π.χ., το σπίτι Α ακολουθεί το σπίτι Β επί του ίδιου δρόμου, διότι έχει *μεγαλύτερο* αριθμό στη διεύθυνσή του.
- Τα αριθμητικά δεδομένα είναι είτε δεδομένα **διαστήματος** (π.χ., ημερομηνίες), είτε δεδομένα **αναλογίας** (π.χ., χρήματα, μήκη).

Τύποι γνωρισμάτων

- Στα γνωρίσματα διαστήματος εφαρμόζονται επιπροσθέτως και *τελεστές πρόσθεσης /αφαίρεσης*.
  - Η αφαίρεση της ημερομηνίας έναρξης μίας διοργάνωσης από την ημερομηνία λήξης της, μας δίνει τη διάρκεια της διοργάνωσης.

Δεδομένα

- Στα γνωρίσματα αναλογίας εφαρμόζονται επιπροσθέτως και *τελεστές πολλαπλασιασμού /διαίρεσης*.
  - Ο μισθός του υπαλλήλου Α είναι διπλάσιος από τον μισθό του υπαλλήλου Β.

Δεδομένα

Τύποι γνωρισμάτων

- Για κάθε τύπο γνωρισμάτων, υπάρχουν διαφορετικοί **μετασχηματισμοί** οι οποίοι, αν εφαρμοστούν με συνέπεια σε τιμές τέτοιου τύπου για όλα τα πρότυπα ενός συνόλου δεδομένων, διατηρούν απαράλλακτο το φυσικό τους νόημα.
  - Απλώς αλλάζει η σύμβαση μέτρησης.
- Ένας *αφινικός μετασχηματισμός* μετρήσεων *διαστήματος* διατηρεί ανέπαφη την ουσία τους.
  - Μετατροπή από βαθμούς Φάρεναϊτ σε Κελσίου.
  - Αφινικός μετασχηματισμός: $Y = aX + b$.
- Όμως μόνον ένας *γραμμικός μετασχηματισμός* αφήνει ανέπαφη την ουσία μετρήσεων αναλογίας.
  - Μετατροπή από μέτρα σε πόδια.
  - Γραμμικός μετασχηματισμός: $Y = aX$.

Τύποι γνωρισμάτων

Δεδομένα

Τύπος γνωρίσματος	Μετασχηματισμός	Τελεστές	Παράδειγμα
Ονομαστικό	Μεταθέσεις τιμών 1-1 συνάρτηση	Ισότητας	Αύξων αριθμός
Διατακτικό	Μονότονη συνάρτηση	Ισότητας Σύγκρισης	Αριθμός σε διεύθυνση οδού
Διαστήματος	Αφινικός μετασχηματισμός	Ισότητας Σύγκρισης Πρόσθεσης	Ημερομηνία
Αναλογίας	Γραμμικός μετασχηματισμός	Ισότητας Σύγκρισης Πρόσθεσης Πολλαπλασιασμού	Μισθός

Τύποι γνωρισμάτων

- Μία ορθογώνια, ανεξάρτητη διάκριση είναι μεταξύ **διακριτών** και **συνεχών** γνωρισμάτων.
- Τα διακριτά γνωρίσματα λαμβάνουν μία από ένα *πεπερασμένο* ή *μετρήσιμα άπειρο* πλήθος δυνατών τιμών.
  - Το πεδίο ορισμού είναι ένα υποσύνολο των ακεραίων.
- Τα συνεχή γνωρίσματα μπορούν να λάβουν απειρία τιμών.
  - Το πεδίο ορισμού είναι οι πραγματικοί αριθμοί.
- Τα διακριτά γνωρίσματα με δύο μόνο δυνατές τιμές ονομάζονται **δυαδικά**.
  - Π.χ., boolean μεταβλητές.

Τύποι γνωρισμάτων

- Τα γνωρίσματα όπου μας ενδιαφέρει μόνο το αν έχουν *μη μηδενική τιμή* ονομάζονται **ασύμμετρα**.
  - Π.χ., αν μία μη μηδενική τιμή δηλώνει παρουσία κάποιας ιδιότητας.
- Στην περίπτωση αυτή, είναι συνήθως απαραίτητο να ληφθούν μέτρα ώστε οι μηδενικές τιμές να μην επηρεάζουν σημαντικά τους χειρισμούς μας επί των δεδομένων.
  - Π.χ., αν σε όλα τα πρότυπα οι μηδενικές τιμές είναι πολύ περισσότερες από τις μη μηδενικές.
- Τα ασύμμετρα γνωρίσματα μπορούν να είναι είτε δυαδικά, είτε διακριτά είτε συνεχή.

Τύποι γνωρισμάτων

- Παράδειγμα συνόλου δεδομένων με *ασύμμετρα δυαδικά* γνωρίσματα:
  - Κάθε γνώρισμα αντιστοιχεί σε ένα διαθέσιμο μάθημα ενός Πανεπιστημίου και έχει τιμή 1 αν κάποιος συγκεκριμένος φοιτητής το έχει διαλέξει, διαφορετικά έχει τιμή 0.
  - Τόσα πρότυπα όσοι οι φοιτητές, τόσα γνωρίσματα ανά πρότυπο όσα τα διαθέσιμα μαθήματα.
  - Προφανώς, η μεγάλη πλειονότητα των τιμών του κάθε προτύπου θα είναι μηδενικά.
    - Αν λάβουμε υπόψη τα μηδενικά κατά τη σύγκριση μεταξύ προτύπων, όλοι οι φοιτητές μοιάζουν *πολύ* μεταξύ τους.

- *Ασύμμετρα διακριτά* γνωρίσματα:
  - Στο ίδιο παράδειγμα, κάθε τιμή 1 μπορεί να αντικατασταθεί με τις πιστωτικές μονάδες του αντίστοιχου μαθήματος.

Τύποι συνόλων δεδομένων

- Εκτός από τους διαφορετικούς τύπους γνωρισμάτων, έχουμε και διαφορετικούς **τύπους συνόλων δεδομένων**.
- Τρία βασικά χαρακτηριστικά τα οποία προσδιορίζουν κάθε τύπο συνόλου δεδομένων είναι:
  - Η **διαστατικότητά** του (dimensionality).
  - Η **αραιότητά** του (sparsity).
  - Η **ανάλυσή** του (resolution).
- Διαστατικότητα είναι το πλήθος γνωρισμάτων ανά πρότυπο.
  - Συνήθως, υψηλή διαστατικότητα προξενεί προβλήματα στους αλγορίθμους εξόρυξης («**κατάρα της διαστατικότητας**»).

Τύποι συνόλων δεδομένων

- Αραιότητα σημαίνει πως τα περισσότερα γνωρίσματα των περισσότερων, ή όλων των προτύπων έχουν μηδενική τιμή.
  - Συνήθως εμφανίζεται όταν έχουμε ασύμμετρα γνωρίσματα.

- Η αραιότητα κατά κανόνα είναι πλεονέκτημα, διότι μας επιτρέπει:
  - Να αποθηκεύσουμε μόνο τις μη μηδενικές τιμές, ώστε να εξοικονομήσουμε αποθηκευτικό χώρο.
  - Να εφαρμόσουμε ειδικούς (π.χ., ταχύτερους) αλγορίθμους εξόρυξης για αραιά δεδομένα.

Δεδομένα

Τύποι συνόλων δεδομένων

- Η χωρική ή χρονική ανάλυση απορρέει από το πώς προέκυψαν τα δεδομένα.
  - Μελέτη της Γης σε κλίμακα: α) δεκάδων χιλιομέτρων, ή β) σε κλίμακα μέτρων.

- Η ανάλυση επηρεάζει:
  - Το πλήθος των προτύπων, άρα το μέγεθος του συνόλου δεδομένων.
  - Το είδος της γνώσης την οποία μπορούμε να εξορύξουμε από αυτά.
    - Π.χ., οι κινήσεις των καταιγίδων μπορούν να συσχετιστούν με αλλαγές στην ατμοσφαιρική πίεση μόνο αν έχουμε ατμοσφαιρικά δεδομένα σε κλίμακα ωρών, όχι σε κλίμακα μηνών.

Τύποι συνόλων δεδομένων

- Ο συνηθέστερος τύπος συνόλων δεδομένων είναι τα δεδομένα **εγγραφών**.
  - Κάθε εγγραφή έχει τα ίδια πεδία/γνωρίσματα και το σύνολο δεδομένων είναι το σύνολο των εγγραφών.

Δεδομένα

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Πηγή: Tan, “Introduction to Data Mining”, 2006.

Τύποι συνόλων δεδομένων

Δεδομένα

- Όταν όλα τα γνωρίσματα των εγγραφών έχουν αριθμητική αναπαράσταση, τότε μιλάμε για την ειδική περίπτωση του **πίνακα δεδομένων**.
  - Σε αυτόν μπορούν να εφαρμοστούν πράξεις γραμμικής άλγεβρας.
  - Κάθε εγγραφή/πρότυπο είναι ένα διάνυσμα σε έναν διανυσματικό χώρο διάστασης ίσης με το πλήθος των πεδίων/γνωρισμάτων.
  - Οι περισσότεροι συνηθισμένοι αλγόριθμοι εξόρυξης επεξεργάζονται πίνακες δεδομένων.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

Πηγή: Tan, “Introduction to Data Mining”, 2006.

Τύποι συνόλων δεδομένων

Δεδομένα

- Μία ακόμα συνηθισμένη περίπτωση είναι τα **συναλλακτικά δεδομένα**.
  - Π.χ., καλάθι αγορών.
  - Κάθε εγγραφή/πρότυπο μπορεί να αναπαρασταθεί με *ασύμμετρα δυαδικά γνωρίσματα* και, επομένως, κατά κανόνα το σύνολο δεδομένων είναι αραιό.

Αρχικό καλάθι αγορών, πριν από την επεξεργασία του.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

Πηγή: Tan, "Introduction to Data Mining", 2006.

- Ενδεχομένως να πρέπει να επεξεργαστούμε χειροκίνητα τα αρχικά δεδομένα για να τα φέρουμε σε αραιή αριθμητική μορφή.

Τύποι συνόλων δεδομένων

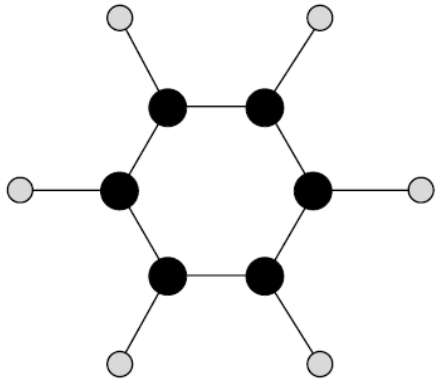
- *Αραιά*, αλλά *μη δυαδικά* σύνολα αριθμητικών δεδομένων (είτε διακριτά, είτε συνεχή) θεωρούνται *ειδική περίπτωση πίνακα δεδομένων*.
  - Π.χ., αναπαραστάσεις εγγράφων (τόσα γνωρίσματα όσες οι δυνατές λέξεις της γλώσσας).

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Πηγή: Tan, “Introduction to Data Mining”, 2006.

Τύποι συνόλων δεδομένων

- Μία πιο σπάνια αλλά σημαντική περίπτωση τύπου συνόλου δεδομένων είναι όταν κάθε πρότυπο είναι ένας **γράφος**.
  - Σε αυτή την περίπτωση έχει νόημα η εξόρυξη *συσχετίσεων* μεταξύ *επιμέρους δομών* (υπογράφων) και ολιστικών ιδιοτήτων του γράφου.
  - Π.χ., εύρεση συσχετίσεων της μορφής «η εμφάνιση του υπογράφου Α σε ένα νέο πρότυπο συνεπάγεται με μεγάλη πιθανότητα την παρατήρηση της ιδιότητας Β στο πρότυπο».



Πηγή: Tan, “Introduction to Data Mining”, 2006

Τύποι συνόλων δεδομένων

- Τέλος, έχουμε τα **διατεταγμένα σύνολα δεδομένων** όπου, συνήθως, η διάταξη των προτύπων είναι *χρονική* (π.χ., χρονοσειρές) ή χωρική (π.χ., *γεωχωρικά* δεδομένα).

- Σε αυτές τις περιπτώσεις υπάρχουν περισσότερο ή λιγότερο ισχυρές *χρονικές ή χωρικές αυτοσυσχετίσεις*.
  - Τα χρονικώς ή χωρικώς γειτονικά πρότυπα έχουν παρόμοιες ιδιότητες.
  - Αυτό πρέπει να ληφθεί υπόψη κατά την εξόρυξη.
    - Π.χ., στη συνόψιση βίντεο τα χρονικώς γειτονικά καρέ έχουν συνήθως πολύ παρόμοιο οπτικό περιεχόμενο.

Τύποι συνόλων δεδομένων

- Σε πολλές περιπτώσεις μετασχηματίζουμε όλους τους τύπους συνόλων δεδομένων σε πίνακες δεδομένων.
  - Ευκολότερη ανάλυση από συνηθισμένους αλγορίθμους εξόρυξης.
- Αυτή η πρακτική όμως απαιτεί προσοχή, καθώς μπορεί να προκαλέσει απώλεια πληροφορίας.
  - Π.χ., πώς θα αναπαρασταθούν χωροχρονικώς διατεταγμένα δεδομένα ως ένας πίνακας;
  - Μπορεί να χαθούν χωρικές ή χρονικές συσχετίσεις.

Δεδομένα

Ποιότητα δεδομένων

- Σε αντίθεση με τη στατιστική, στην εξόρυξη γνώσης τα αναλυόμενα δεδομένα είναι καιροσκοπικά.
  - Η ποιότητα τέτοιων δεδομένων δεν είναι πάντοτε βέλτιστη για ποικίλους λόγους.
- Μπορεί να υπάρχουν σφάλματα στις μετρήσεις κάποιων τιμών γνωρισμάτων, ή σε ολόκληρα πρότυπα:
  - Λείπουν στοιχεία ή τιμές γνωρισμάτων.
  - Υπάρχει *θόρυβος* στις μετρήσεις: → τυχαία συνιστώσα ενός σφάλματος μέτρησης, το οποίο εκδηλώνεται ως παραμόρφωση της τιμής.
  - Υπάρχουν *τεχνουργήματα* στις τιμές των γνωρισμάτων: → ντετερμινιστικές παραμορφώσεις των πραγματικών τιμών.

Ποιότητα δεδομένων

Δεδομένα

- Μπορεί να:
  - Υπάρχουν ασυνεπείς ή παράλογες τιμές σε κάποια γνωρίσματα.
  - Υπάρχουν διπλότυπα πρότυπα/στοιχεία.
  - Μην υπάρχουν έγκυρα πρότυπα.
  - Κλπ.

- Συνολικά, πρόκειται για υποπεριπτώσεις **σφαλμάτων μέτρησης** ή **σφαλμάτων συλλογής των δεδομένων**.
  - Π.χ., παραμορφώσεις μετρούμενων τιμών.
  - Π.χ., προσθήκη ψευδών προτύπων ή παράλειψη έγκυρων προτύπων.

Δεδομένα

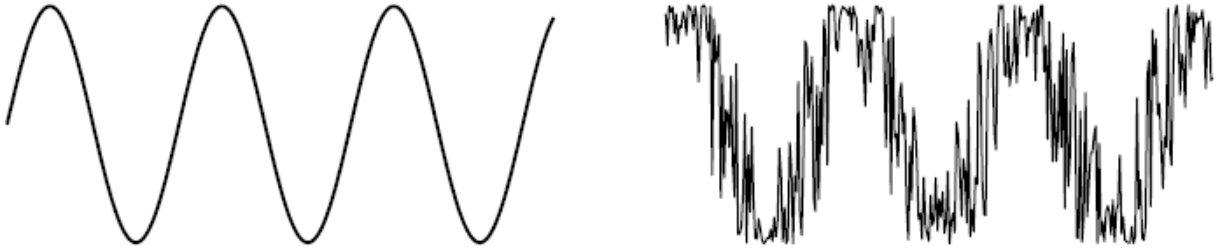
Ποιότητα δεδομένων

- Ενδείξεις της ποιότητας των δεδομένων μας δίνουν τα μεγέθη της **διακύμανσης** και της **μεροληψίας**.
  - Μετρούνται μία φορά για γνωστές ποσότητες και χαρακτηρίζουν ενδεικτικά όλο το σύνολο δεδομένων.
- Η διακύμανση είναι η τυπική απόκλιση πολλαπλών ανεξάρτητων μετρήσεων της ίδιας φυσικής ιδιότητας (γνωρίσματος) του ίδιου φυσικού αντικειμένου (προτύπου).
- Η μεροληψία είναι η διαφορά της μέσης τιμών των πολλαπλών ανεξάρτητων μετρήσεων από την πραγματική τιμή της μετρούμενης ιδιότητας.

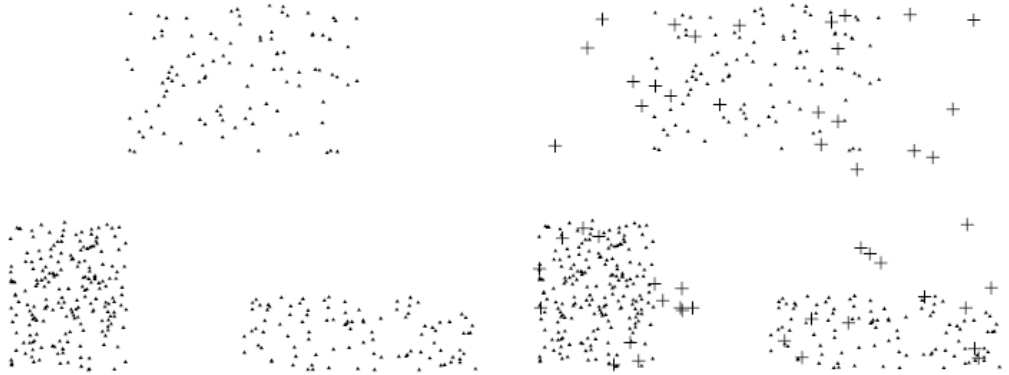
Ποιότητα δεδομένων

- Η λεγόμενη **ακρίβεια** εξαρτάται από τη μεροληψία και τη διακύμανση.
  - Δηλώνει την εγγύτητα των μετρήσεών μας στις αντίστοιχες πραγματικές τιμές.
  - Γενικός όρος ο οποίος εξειδικεύεται αναλόγως με την περίπτωση.

Θόρυβος μέτρησης.



Σφάλματα συλλογής δεδομένων.



Πηγή: Tan, “Introduction to Data Mining”, 2006.

Δεδομένα

Καθαρισμός δεδομένων

- Η εξόρυξη γνώσης είναι προσανατολισμένη στη σχεδίαση **ευσταθών** αλγορίθμων ικανών να λειτουργούν επαρκώς *καλά* με δεδομένα *κακής* ποιότητας.
- Ωστόσο, το στάδιο της προεπεξεργασίας, όπου γίνεται ανίχνευση ή/και διόρθωση των σφαλμάτων, εξακολουθεί να είναι απαραίτητο.
- Απούσες τιμές: Έστω ότι λείπει από ένα πρότυπο η τιμή ενός γνωρίσματος:
  - είτε επειδή δεν μετρήθηκε,
  - είτε επειδή εγγενώς το εν λόγω γνώρισμα δεν υπάρχει σε όλα τα πρότυπα.
    - Π.χ., πεδία σε υπεύθυνες δηλώσεις τα οποία δεν συμπληρώνονται από όλους.

Καθαρισμός δεδομένων

- Στις περιπτώσεις απουσίας τιμών, μπορεί:
  - είτε να αφαιρεθεί το ελαττωματικό στοιχείο,
  - είτε να μη ληφθεί υπ' όψη το γνώρισμα σε κανένα στοιχείο,
  - είτε να γίνει μία εκτίμηση της απούσας τιμής.
    - Π.χ., ο μέσος όρος των τιμών που έχουν στο γνώρισμα αυτό τα *n* πιο όμοια με το τρέχον πρότυπα, για αριθμητικά γνωρίσματα.
    - Π.χ., η συχνότερα εμφανιζόμενη τιμή του γνωρίσματος, για κατηγορικά γνωρίσματα.

- <u>Ασυνεπείς τιμές γνωρισμάτων</u>: Π.χ., αρνητικό ύψος.
  - Σε αυτές τις περιπτώσεις πρέπει κατ' αρχάς το πρόβλημα ποιότητας να ανιχνεύεται μέσω κατάλληλων ελέγχων και, στη συνέχεια, να αντιμετωπίζεται με βάση κάποια πολιτική.

Καθαρισμός δεδομένων

- Στις περιπτώσεις ασυνεπών τιμών, μπορεί:
  - είτε να αντικατασταθεί η εσφαλμένη τιμή με μία προεπιλεγμένη (π.χ., 1),
  - είτε το πρόβλημα να θεωρηθεί ζήτημα απούσας τιμής,
  - είτε το πρόβλημα να διορθωθεί.
    - Φυσικά, για την αυτόματη διόρθωση απαιτείται είτε καταφυγή σε κάποια *εξωτερική πηγή δεδομένων*, είτε ανίχνευση και διόρθωση αριθμητικών σφαλμάτων μέσω *πλεοναζόντων* ψηφίων (CRC, ECC).
- <u>Διπλότυπα πρότυπα</u>: Σε αυτές τις περιπτώσεις πρέπει να εντοπίσουμε τα διπλότυπα και να τα ενοποιήσουμε.
  - Μπορεί να διαφέρουν σε ορισμένες τιμές γνωρισμάτων τους, λόγω θορύβου ή τεχνουργημάτων.
  - Δεν πρέπει να ενοποιήσουμε παρόμοια πρότυπα τα οποία όμως αναφέρονται σε διαφορετικά φυσικά αντικείμενα (π.χ., δύο φορολογούμενοι με το ίδιο όνομα).

Καθαρισμός δεδομένων

- Αναλόγως με τον στόχο της εξόρυξης, η προεπεξεργασία μπορεί να θεωρεί και τις **ανωμαλίες** των δεδομένων ως ζήτημα προς αντιμετώπιση.
  - Πρόκειται είτε για ιδιότυπες τιμές *γνωρισμάτων*, σε σχέση με τις τιμές των γνωρισμάτων αυτών στα υπόλοιπα στοιχεία, είτε για ιδιότυπα *πρότυπα*, με χαρακτηριστικά σημαντικά διαφορετικά από την πλειονότητα των στοιχείων του συνόλου δεδομένων μας.
- Κατ' αντιδιαστολή με τον θόρυβο ή τα τεχνουργήματα, οι ανωμαλίες είναι *έγκυρα* πρότυπα ή τιμές γνωρισμάτων.
- Έτσι, σε ορισμένα προβλήματα, το ζητούμενο είναι ακριβώς η ανίχνευση ανωμαλιών.

Προδιαγραφές ιδανικών δεδομένων

- Ένα σύνολο δεδομένων πρέπει ιδανικά να ανταποκρίνεται σε ορισμένες **προδιαγραφές**, κατά τη στιγμή της εξόρυξης γνώσης από αυτό.
  - Αν τα δεδομένα στο εν λόγω αντικείμενο/πρόβλημα απαρχαιώνονται γρήγορα, θα πρέπει να εξακριβωθεί πριν την εξόρυξη πως το σύνολο δεδομένων είναι πράγματι **επίκαιρα**.
  - Θα πρέπει να εξακριβωθεί πως τα δεδομένα είναι **σχετικά** με το πρόβλημα και, επομένως, ενσωματώνουν όλη την απαιτούμενη γνώση (π.χ., να μην υπάρχει μεροληψία δειγματοληψίας, λόγω του καιροσκοπικού χαρακτήρα τους).
  - Πρέπει να είναι γνωστή (π.χ., μέσω σχετικής **τεκμηρίωσης**) η φύση και το πεδίο τιμών κάθε γνωρίσματος.

Προεπεξεργασία

Τύποι προεπεξεργασίας

- Αφού εξακριβωθεί πως το σύνολο δεδομένων ικανοποιεί τις προδιαγραφές μας και έχει καθαριστεί, συνήθως εφαρμόζεται σε αυτό κάποια προεπεξεργασία.
  - Στόχος είναι η αύξηση της καταλληλότητας των δεδομένων για τροφοδότηση σε κάποιον αλγόριθμο εξόρυξης.
- Οι βασικότεροι τύποι προεπεξεργασίας είναι οι μέθοδοι **επιλογής** και οι μέθοδοι **μετασχηματισμού**:
  - Οι πρώτες επιλέγουν ένα υποσύνολο των στοιχείων/προτύπων ή των γνωρισμάτων για ανάλυση, απορρίπτοντας τα υπόλοιπα.
  - Οι δεύτερες δημιουργούν νέα διανύσματα χαρακτηριστικών, ή τροποποιούν τα υπάρχοντα, πριν τα τροφοδοτήσουν στον αλγόριθμο.

Συνάθροιση

- Μία συνήθης μέθοδος προεπεξεργασίας είναι η **συνάθροιση** (aggregation), όπου συνθέτονται πολλαπλά στοιχεία/πρότυπα/διανύσματα χαρακτηριστικών σε ένα.
- Οι τιμές των *αριθμητικών* γνωρισμάτων των συνδυαζόμενων στοιχείων συνήθως:
  - προστίθενται, ή
  - εξάγεται ο μέσος όρος τους.
- Τα *κατηγορικά* γνωρίσματα:
  - παραλείπονται, ή
  - συνοψίζονται οι τιμές τους σε ένα σύνολο τιμών.

Προεπεξεργασία

Συνάθροιση

- Με τη συνάθροιση:
  - μειώνεται το **κόστος επεξεργασίας και αποθήκευσης** του συνόλου δεδομένων,
  - υλοποιείται μία **αλλαγή κλίμακας ή σκοπιάς** αντιμετώπισης των δεδομένων,
    - Π.χ., συνάθροιση προτύπων στον χρόνο, ώστε να αλλάξει η μακροσκοπική ανάλυση των δεδομένων.
  - μειώνεται ο **θόρυβος** των δεδομένων.
    - Π.χ., η αναμενόμενη τιμή ενός συνόλου μετρήσεων έχει μικρότερη διακύμανση από τις επιμέρους μετρήσεις.
- Ωστόσο, με τη συνάθροιση ελλοχεύει πάντοτε ο κίνδυνος απώλειας ενδιαφερόντων λεπτομερειών.
  - Π.χ., πληροφορία max ή min μεταξύ των αθροιζόμενων προτύπων.

Προεπεξεργασία

Δειγματοληψία

- Εναλλακτική μέθοδος: **δειγματοληψία**.
  - Εφαρμόζεται σε μεγάλα σύνολα δεδομένων, για να μειωθεί ο χρόνος επεξεργασίας τους από τους αλγορίθμους εξόρυξης.
  - Επιλέγεται αντιπροσωπευτικό υποσύνολο των δεδομένων για ανάλυση.
- Υπάρχουν πολλαπλές στρατηγικές δειγματοληψίας:
  - *Δειγματοληψία χωρίς επανατοποθέτηση*.
    - Κάθε επιλεγμένο στοιχείο αφαιρείται από τη δεξαμενή των προτύπων πριν από την επιλογή του επόμενου.
  - *Δειγματοληψία με επανατοποθέτηση*.
    - Είναι πιθανή η επιλογή του ίδιου στοιχείου πολλαπλές φορές για συμμετοχή στο τελικό δείγμα.
  - *Διαστρωματική δειγματοληψία*.
    - Ανατίθενται διαφορετικά βάρη σε διαφορετικά υποσύνολα/ομάδες των δεδομένων και, άρα, διαφορετική πιθανότητα επιλογής.

Προεπεξεργασία

Δειγματοληψία

- Στρατηγικές δειγματοληψίας:
  - *Προοδευτική δειγματοληψία*.
    - Η μέθοδος ξεκινά με ένα μικρό δείγμα και σταδιακά αυξάνει το μέγεθός του μέχρι να βρεθεί ένα κατάλληλο δείγμα.
- **Διαστρωματική δειγματοληψία**:
  - Εφαρμόζεται συνήθως όταν το σύνολο δεδομένων είναι εκ φύσεως διαμερισμένο σε ομάδες διαφορετικού μεγέθους.
  - Θέλουμε να εξασφαλίσουμε πως στο τελικό δείγμα θα αντιπροσωπεύονται αναλογικά ή εξίσου όλες οι εν λόγω ομάδες.
    - Ξεχωριστή δειγματοληψία ανά ομάδα ενός ποσοστού του ολικού δείγματος ανάλογου με το σχετικό μέγεθος της ομάδας, ή ίσου για όλες τις ομάδες.
  - <u>Προσοχή</u>: δεν αναθέτουμε διαφορετικό βάρος δειγματοληψίας ανά πρότυπο, απλώς επιλέγουμε ξεχωριστό υποδείγμα ανά αρχική ομάδα.

Δειγματοληψία

Προεπεξεργασία

- **Προοδευτική δειγματοληψία**:
  - Εφαρμόζεται όταν το μέγεθος του δείγματος είναι παράγοντας *κρίσιμος,* ώστε να καταλήξουμε με δείγμα επαρκώς αντιπροσωπευτικό του ολικού συνόλου δεδομένων, αλλά *δύσκολο να προσδιοριστεί* με ακρίβεια.
  - Απαιτείται κάποιος τρόπος να εξακριβώσουμε την ποιότητα κάθε δυνατού δείγματος, ώστε να αποφασίσουμε αν θα συνεχίσουμε να αυξάνουμε το μέγεθός του.

Προεπεξεργασία

Μείωση διαστατικότητας

- Κεντρικό ζήτημα κατά την προεπεξεργασία των δεδομένων είναι η μείωση της διαστατικότητάς τους, δηλαδή *του πλήθους των γνωρισμάτων τους*.
- Αυτή η πρακτική:
  - επιτρέπει να εξοικονομήσουμε χρόνο επεξεργασίας, υπολογιστική ισχύ και μνήμη κατά την εξόρυξη,
  - μειώνει τον εγγενή θόρυβο του συνόλου δεδομένων,
  - επιτρέπει την ευκολότερη κατανόηση των γνωρισμάτων και την οπτικοποίηση των δεδομένων,
  - αντιμετωπίζει αποτελεσματικά την **κατάρα της διαστατικότητας** (curse of dimensionality).
    - Όσο αυξάνει η διαστατικότητα, τόσο πιο αραιά μεταξύ τους είναι πλέον τα δεδομένα στον διανυσματικό τους χώρο.
    - Δυσκολεύονται περισσότερο οι αλγόριθμοι εξόρυξης.

Προεπεξεργασία

Μείωση διαστατικότητας

- Για τη μείωση της διαστατικότητας χρησιμοποιούνται:
  - Αλγόριθμοι μείωσης διάστασης (π.χ., Ανάλυση Κυρίων Συνιστωσών, PCA).
  - Μέθοδοι **επιλογής** ή **εξαγωγής χαρακτηριστικών**.
- Οι μέθοδοι επιλογής επιχειρούν να αφαιρέσουν *πλεονάζοντα* ή *μη ενδιαφέροντα* γνωρίσματα από τα πρότυπα.
  - Στόχος τους είναι να διατηρήσουν μόνον ένα υποσύνολο των αρχικών γνωρισμάτων στο σύνολο δεδομένων.
- Παράδειγμα πλεονάζοντος γνωρίσματος: τελική λιανική τιμή (με ΦΠΑ), αν υπάρχει ήδη γνώρισμα με την τιμή προ φόρων.
- Παράδειγμα μη ενδιαφέροντος γνωρίσματος: αύξων αριθμός ενός φοιτητή στο μητρώο του πανεπιστημίου, αν στόχος μας είναι η πρόβλεψη των ακαδημαϊκών του επιδόσεων.

Προεπεξεργασία

Μείωση διαστατικότητας

- Η διαδικασία επιλογής χαρακτηριστικών ίσως είναι **ενσωματωμένη** εγγενώς στον ίδιο τον αλγόριθμο εξόρυξης.
  - Π.χ., δένδρα απόφασης.
  - Τότε δεν είναι απαραίτητη η προεπεξεργασία.
- Διαφορετικά, κατά την προεπεξεργασία εκτελείται μία **μη εξαντλητική, επαναληπτική αναζήτηση** στον χώρο των δυνατών υποσυνόλων γνωρισμάτων.
  - Με κατάλληλη *αξιολόγηση* του τρέχοντος υποσυνόλου ανά επανάληψη, η αναζήτηση τερματίζεται πρόωρα με βάση ένα *κριτήριο τερματισμού*. Π.χ.:
    - Έχει βρεθεί ήδη υποσύνολο με τιμή αξιολόγησης μεγαλύτερη από ένα χειροκίνητο κατώφλι.
    - Η τιμή αξιολόγησης έχει πάψει να αυξάνεται από επανάληψη σε επανάληψη, κλπ.

Προεπεξεργασία

Μείωση διαστατικότητας

- Πώς γίνεται η αξιολόγηση κάθε υποσυνόλου γνωρισμάτων;
- Δύο τρόποι:
  - **Προσέγγιση περιβλήματος** (wrapper): χρησιμοποιείται η ακρίβεια του ίδιου του αλγορίθμου εξόρυξης στον οποίον θα τροφοδοτηθούν τα τελικά δεδομένα.
  - **Προσέγγιση φίλτρου** (filter): χρησιμοποιείται κάποιο εξωτερικό κριτήριο αξιολόγησης υποσυνόλων.
    - Π.χ., στατιστική συσχέτιση μεταξύ των γνωρισμάτων του εκάστοτε υποσυνόλου → ζητούμε υποσύνολα ασυσχέτιστων μεταξύ τους γνωρισμάτων.
- Αντί να αποτιμούμε δυαδικά τα γνωρίσματα (διατηρούνται ή απορρίπτονται), μπορούμε εναλλακτικά να τα σταθμίζουμε με διαφορετικά πραγματικά βάρη (συντελεστές στο [0, 1]).
  - Έτσι οι διαφορετικές συνιστώσες των προτύπων θα συνεισφέρουν σε διαφορετικό βαθμό στην εξόρυξη γνώσης.
  - Υπολογισμός βαρών με προσέγγιση περιβλήματος ή φίλτρου.

Προεπεξεργασία

Μείωση διαστατικότητας

- **Μέθοδοι εξαγωγής χαρακτηριστικών**: η μείωση της διαστατικότητας των δεδομένων γίνεται διά της κατασκευής ενός νέου, μικρότερου συνόλου γνωρισμάτων το οποίο συλλαμβάνει πιο αποτελεσματικά τη σημαντική πληροφορία.

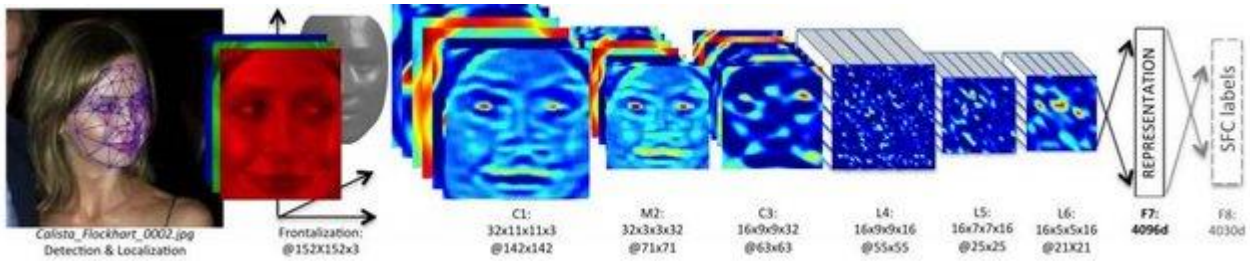
- Τα αρχικά γνωρίσματα μπορεί:
  - είτε να περιέχουν αυτήν την πληροφορία, απλώς όχι σε μορφή κατάλληλη για ανάλυση,
  - είτε να μην την περιέχουν καθόλου και η πληροφορία να κατασκευάζεται από μία ερμηνεία των αρχικών δεδομένων.

- Οι μέθοδοι εξαγωγής χαρακτηριστικών εξαρτώνται *σχεδόν απολύτως* από το εκάστοτε πρόβλημα και από τον τύπο των δεδομένων.

Μείωση διαστατικότητας

- Τι γίνεται όταν τα διαθέσιμα δεδομένα είναι επαρκώς μεγάλα σε πλήθος;
  - Σύνολα δεδομένων πολύ μεγάλου μεγέθους.
- Τα **βαθιά νευρωνικά δίκτυα** είναι σε θέση να μάθουν να εξάγουν τα κατάλληλα χαρακτηριστικά για το ζητούμενο πρόβλημα.
  - Έτσι δεν χρειάζεται προεπεξεργασία με κάποιον χειροποίητο αλγόριθμο εξαγωγής χαρακτηριστικών.
  - Μάθηση χαρακτηριστικών.

Προεπεξεργασία



Πηγή: Alashkar, “Real-time lane and car detection on highway using vision system”, DOI: 10.13140/RG.2.2.18416.05128 2016.

Προεπεξεργασία

Διακριτοποίηση

- Δεν είναι ασυνήθιστο κατά την προεπεξεργασία κάποια συνεχή γνωρίσματα των προτύπων να **διακριτοποιούνται**.
  - Π.χ., για να μπορέσει το σύνολο δεδομένων να τροφοδοτηθεί σε κάποιον αλγόριθμο εξόρυξης ο οποίος απαιτεί διακριτά γνωρίσματα.
- Εναλλακτικά, ίσως απαιτείται **δυαδικοποίηση** κάποιων κατηγορικών γνωρισμάτων.
  - Ένα επίμαχο κατηγορικό γνώρισμα m δυνατών τιμών αντικαθίσταται από m ασύμμετρα δυαδικά γνωρίσματα, καθένα από τα οποία κωδικοποιεί ανεξάρτητα την αντίστοιχη τιμή.
  - Π.χ.: το γνώρισμα «καιρός» με δυνατές τιμές «{αίθριος}, {νεφελώδης}, {βροχερός}» αντικαθίσταται από τρία δυαδικά γνωρίσματα. Άρα:
    - Πρότυπο με {αίθριο} καιρό θα έχει τώρα τιμή: [1,0,0].
    - Πρότυπο με {νεφελώδη} καιρό θα έχει τώρα τιμή [0,1,0].
    - Πρότυπο με {βροχερό} καιρό θα έχει τώρα τιμή [0,0,1].

Διακριτοποίηση

- Η κατάσταση είναι περισσότερο περίπλοκη με τη διακριτοποίηση *συνεχών* γνωρισμάτων.
  - Υπάρχουν πολλαπλοί εναλλακτικοί τρόποι.
  - Ο καλύτερος εξαρτάται από το ποιος αλγόριθμος εξόρυξης πρόκειται να εφαρμοστεί κάθε φορά.
  - Κεντρική ιδέα είναι η ταξινόμηση των δυνατών τιμών κάποιου γνωρίσματος και η εύρεση n-1 σημείων διαμέρισης που διαχωρίζουν n υποδιαστήματα.
    - Το n είναι υπερπαράμετρος κβαντισμού.
    - Σε κάθε υποδιάστημα A_i αντιστοιχίζεται μία αντιπροσωπευτική τιμή α.
    - Η τιμή κάθε στοιχείου/προτύπου που σε αυτό το γνώρισμα έχει τιμή εντός του A_i, αλλάζει κατά την προεπεξεργασία σε α.

Προεπεξεργασία

Προεπεξεργασία

Διακριτοποίηση

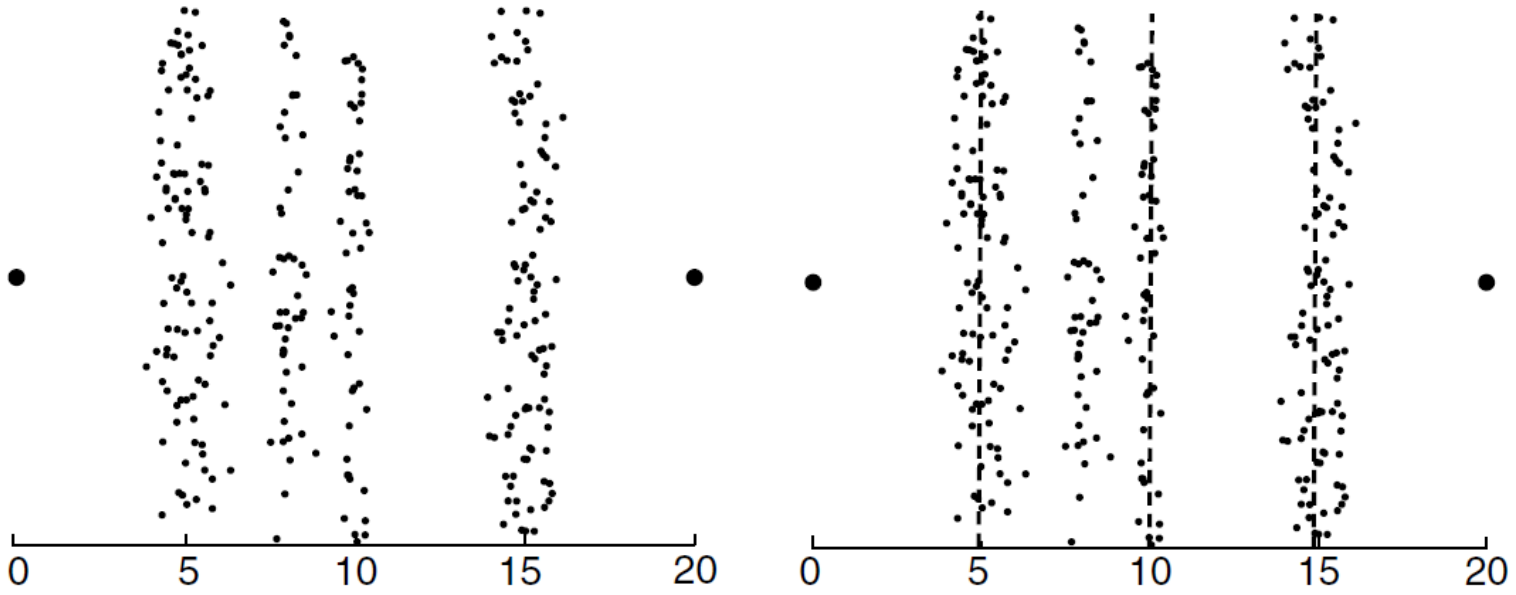
- Στη συνήθη **ανεπίβλεπτη διακριτοποίηση** το πεδίο ορισμού του γνωρίσματος διαμερίζεται:
  - α) είτε σε ισομήκη υποδιαστήματα,
  - β) είτε σε υποδιαστήματα με ίσο πλήθος προτύπων από το σύνολο δεδομένων (ίσες συχνότητες),
  - γ) είτε με ομαδοποίηση των προτύπων σε *n* ομάδες.
- Η στρατηγική της ίσης συχνότητας εξασφαλίζει συνήθως τη μεγαλύτερη ευστάθεια παρουσία ανωμαλιών.
- Η ομαδοποίηση δίνει συνήθως τα καλύτερα αποτελέσματα, πληρώνουμε όμως το ανάλογο υπολογιστικό κόστος.

Διακριτοποίηση

Προεπεξεργασία

Αρχικά μονοδιάστατα δεδομένα.

Διακριτοποίηση ίσου μήκους.



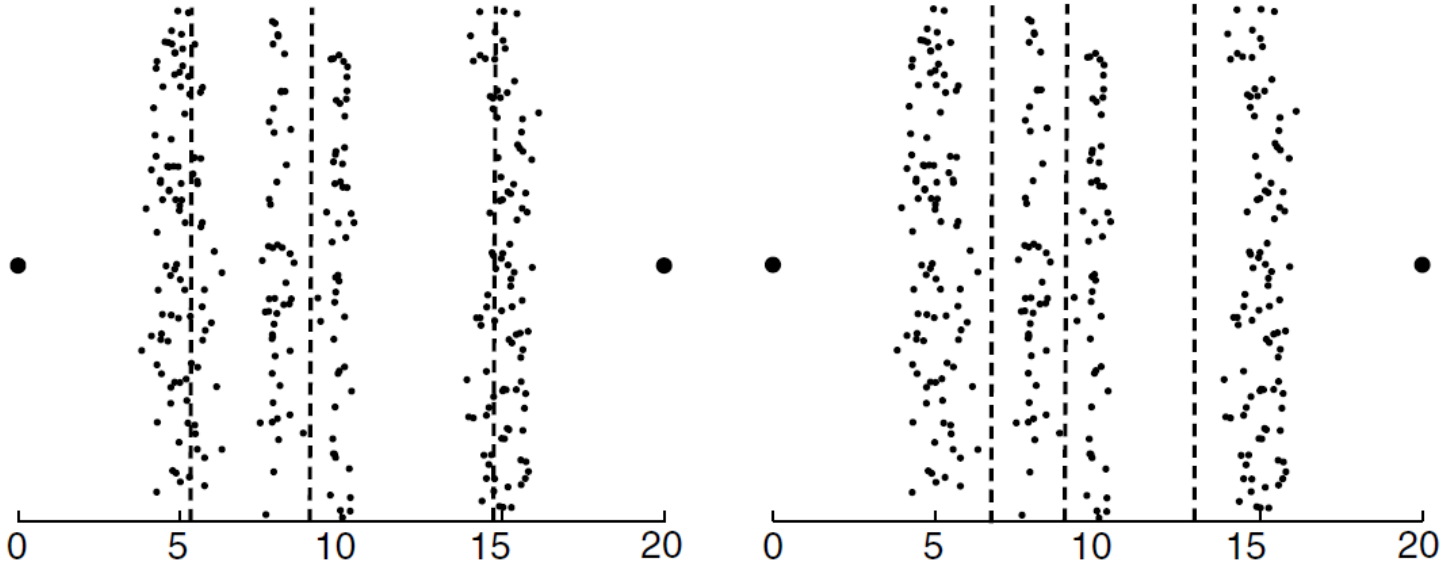
Πηγή: Tan, “Introduction to Data Mining”, 2006.

Διακριτοποίηση

Προεπεξεργασία

Διακριτοποίηση ίσης συχνότητας.

Διακριτοποίηση με ομαδοποίηση.



Πηγή: Tan, “Introduction to Data Mining”, 2006.

Προεπεξεργασία

Διακριτοποίηση

- Αν πρόκειται για σύνολο δεδομένων ταξινόμησης, για κάθε πρότυπο/στοιχείο έχουμε και μία διαθέσιμη ετικέτα κλάσης.
- Τότε, η διακριτοποίηση ενός γνωρίσματος μπορεί προαιρετικά να γίνει **επιβλεπόμενα**.
  - Ορίζεται κάποιο αριθμητικό μέτρο *εντροπίας* το οποίο μετρά την *καθαρότητα* κάθε υποδιαστήματος.
  - Καθαρότητα είναι ο βαθμός στον οποίον το υποδιάστημα περιέχει πρότυπα/στοιχεία του συνόλου δεδομένων από μόνο μία ή από περισσότερες κλάσεις.
  - Για κάθε διαμέριση σε υποδιαστήματα, υπολογίζουμε τον σταθμισμένο μέσο όρο της εντροπίας όλων των υποδιαστημάτων.
  - Το βάρος κάθε υποδιαστήματος είναι ανάλογο του πλήθους των προτύπων/στοιχείων που εμπίπτουν σε αυτό.

Προεπεξεργασία

Διακριτοποίηση

- Επιβλεπόμενη διακριτοποίηση.
  - Εν τέλει, στόχος είναι η εύρεση των σημείων διαμέρισης επί του πεδίου ορισμού του γνωρίσματος, τα οποία από κοινού δίνουν την ελάχιστη μέση εντροπία.
- Ένας απλός αλγόριθμος είναι η **επαναληπτική διχοτόμηση** του κάθε υποδιαστήματος, εκκινώντας από ένα μόνο υποδιάστημα ίσο με όλο το πεδίο ορισμού του γνωρίσματος.
  - Επιλέγεται κάθε φορά άπληστα ως σημείο διαμέρισης αυτό για το οποίο προκύπτει η ελάχιστη μέση εντροπία.
    - Άρα τα τοπικά καθαρότερα υποδιαστήματα.
  - Η διαδικασία επαναλαμβάνεται μέχρι να έχουν οριστεί συνολικά n υποδιαστήματα.
- Υψηλή εντροπία υποδιαστήματος ⟶ υποδιάστημα μικρής καθαρότητας.

Προεπεξεργασία

Διακριτοποίηση

- Γιατί η εντροπία;
  - Συνιστά μέτρο του κατά πόσον μία κατανομή πιθανοτήτων προσεγγίζει την ομοιόμορφη κατανομή.
  - Η ομοιόμορφη έχει τη μέγιστη δυνατή εντροπία, ενώ η κρουστική κατανομή την ελάχιστη δυνατή εντροπία.
  - Στην επιβλεπόμενη διακριτοποίηση, ο διακριτός δειγματικός χώρος της κατανομής είναι οι k κλάσεις.
  - Υποδιάστημα το οποίο περιέχει πρότυπα μόνο μίας κλάσης έχει εντροπία 0 και μέγιστη καθαρότητα.
  - Η εντροπία e_i του i-οστού υποδιαστήματος είναι:

$$e_i = \sum_{i=1}^{k} p_{ij} \log_2 p_{ij},$$

όπου p_{ij} είναι το ποσοστό των προτύπων με τιμή του επίμαχου γνωρίσματος ευρισκόμενη μες στο i-οστό υποδιάστημα τα οποία ανήκουν στην j-οστή κλάση.

Προεπεξεργασία

Διακριτοποίηση

- Γιατί η εντροπία;
  - Η μέση εντροπία e των n υποδιαστημάτων είναι:

$$e = \sum_{i=1}^{n} w_i e_i,$$

όπου w_i είναι το ποσοστό των προτύπων με τιμή του επίμαχου γνωρίσματος ευρισκόμενη μες στο i-οστό υποδιάστημα.

Προεπεξεργασία

Μετασχηματισμοί ανά γνώρισμα

- Πέρα από τη διακριτοποίηση, άλλου τύπου χειροκίνητοι μετασχηματισμοί ανά γνώρισμα είναι επίσης εφικτοί κατά την προεπεξεργασία.
- Π.χ., εφαρμογή απλών συναρτήσεων σε όλες τις τιμές ενός γνωρίσματος.
  - Ανεξάρτητα σε κάθε πρότυπο/στοιχείο του συνόλου δεδομένων.
- Γιατί; Εξαρτάται από την εκάστοτε εφαρμογή, τον τύπο του προς επίλυση προβλήματος και τον τύπο των δεδομένων.
- Απαιτείται προσοχή προκειμένου να μη χαθεί ενδεχομένως σημαντική πληροφορία.
  - Π.χ., η διάταξη.

Μετασχηματισμοί ανά γνώρισμα

- Προεπεξεργασία

- Παράδειγμα: Γνώρισμα που εκφράζει το πλήθος των byte τα οποία μεταδόθηκαν κατά τη διάρκεια μίας τηλεπικοινωνιακής συνεδρίας.

- Σε ένα πρόβλημα *ανίχνευσης εισβολών σε δίκτυα*, ίσως βοηθά η αντικατάσταση της τιμής αυτής από τον λογάριθμό της:
  - Π.χ., δύο πρότυπα με τιμή 10^8 και 10^9 (μεταφορές μεγάλων αρχείων) μοιάζουν περισσότερο μεταξύ τους από δύο πρότυπα με τιμή 10^1 και 10^4 (δύο πολύ διαφορετικές εφαρμογές).
  - Με τη λογαρίθμηση, η διαφορά των εφαρμογών γίνεται εμφανής:
    - *Πριν τη λογαρίθμηση*: η διαφορά μεγέθους των μεγάλων αρχείων επισκιάζει τη διαφορά των μικρών.
    - *Μετά τη λογαρίθμηση*: το 1 απέχει από το 4 περισσότερο από ότι το 8 από το 9.

Μετασχηματισμοί ανά γνώρισμα

Προεπεξεργασία

- Ειδική περίπτωση: **τυποποίηση** των *γκαουσιανά κατανεμημένων* πραγματικών τιμών ενός γνωρίσματος.
  - Αφαιρούμε από κάθε πρότυπο τη μέση τιμή του γνωρίσματος επί όλου του συνόλου δεδομένων και διαιρούμε το αποτέλεσμα με την τυπική απόκλιση.
  - Μετά την τυποποίηση, οι τιμές του γνωρίσματος επί όλου του συνόλου δεδομένων θα έχουν πλέον **μέση τιμή 0** και **διακύμανση 1**.
- Έτσι φέρνουμε με αυτόματο τρόπο όλα τα γνωρίσματα ενός συνόλου προτύπων στην ίδια τάξη μεγέθους, αν πριν δεν ήταν.
  - Διευκολύνουμε τη λειτουργία του αλγορίθμου εξόρυξης στο επόμενο στάδιο.

Προεπεξεργασία

Μετασχηματισμοί ανά γνώρισμα

- Διαφορετικά, ο αλγόριθμος εξόρυξης μπορεί να αγνοήσει τα γνωρίσματα με μικρή τάξη μεγέθους.
  - Ενδέχεται να εστιάσει αποκλειστικά σε εκείνες τις συνιστώσες των διανυσμάτων χαρακτηριστικών με την υψηλή τάξη μεγέθους.
    - **Έμμεση απώλεια πληροφορίας**!
- Η μέθοδος τυπικά είναι ορθή μόνο για γκαουσιανά κατανεμημένα γνωρίσματα.
  - Μπορούμε να την εφαρμόσουμε ευρετικά και σε άλλες περιπτώσεις.
  - Εναλλακτικά, ίσως πρέπει να αλλάξουμε τα γνωρίσματα ώστε να έχουν όλα την ίδια τάξη μεγέθους.
- Αντιμετώπιση ανωμαλιών κατά τον υπολογισμό της *μέσης τιμής* και της *τυπικής απόκλισης*:
  - Αντικατάσταση με *διάμεσο* και *απόλυτη τυπική απόκλιση*.

Προεπεξεργασία

Μετασχηματισμοί ανά γνώρισμα

- **Διάμεσος** (median): η τιμή για την οποία το 50% των δεδομένων είναι μικρότερα από αυτήν και το 50% των δεδομένων είναι μεγαλύτερα από αυτήν.
  - Εγγενώς πιο ευσταθής σε ανωμαλίες από τη μέση τιμή.

- **Απόλυτη τυπική απόκλιση**: Υπολογίζεται παρομοίως με την τυπική απόκλιση, αλλά ο τύπος της έχει απόλυτη τιμή αντί για ύψωση στο τετράγωνο.
  - Έτσι, οι ακραίες τιμές/ανωμαλίες δεν υπερτονίζονται μέσω του τετραγωνισμού.
  - Τύπος: $\frac{1}{N}\sum_{i=1}^{n}|x_i - \mu|$.

Thank you for your attention!

Q & A

Contact: imademlis@aueb.gr