Εξόρυξη Γνώσης από Δεδομένα

3: Προεπεξεργασία δεδομένων

Μέτρα ομοιότητας

Περιεχόμενα

- Προεπεξεργασία
  - Ενέργειες
  - Τεχνικές
- Μείωση διαστάσεων
- Μετασχηματισμοί
- Αναπαράσταση
- Μέτρα απόστασης

Γιατί;

- Τα δεδομένα που καλούμαστε να χειριστούμε δεν είναι πάντοτε «καθαρά»
  - Ατελή δεδομένα: λείπουν οι τιμές σε κάποια χαρακτηριστικά, λείπουν χαρακτηριστικά που χρειαζόμαστε, υπάρχουν τα αποτελέσματα μιας συνάθροισης αλλά όχι τα αρχικά δεδομένα
  - Θόρυβος: υπάρχουν λανθασμένες καταχωρήσεις, εξαιρέσεις (outliers)
  - Ασυνεπή δεδομένα: υπάρχουν αντικρουόμενα δεδομένα, π.χ. διαφορετικά ονόματα και κωδικοί για το ίδιο αντικείμενο
- Ελαττωματικά δεδομένα ➔ Κακά αποτελέσματα
  - Οι αποθήκες δεδομένων απαιτούν συνέπεια στην ολοκλήρωση των δεδομένων

Ποιότητα δεδομένων

- Ακρίβεια
- Πληρότητα
- Συνέπεια
- Επικαιρότητα
- Αξιοπιστία
- Προστιθέμενη αξία
- Ευκολία στην ερμηνεία
- Προσβασιμότητα

Ενέργειες

- Καθαρισμός δεδομένων (data cleaning)
  - Συμπλήρωση τιμών που λείπουν, διαχείριση θορύβου, απομάκρυνση εξαιρέσεων, επίλυση ασυνεπειών
- Ολοκλήρωση δεδομένων
  - Ολοκλήρωση πολλαπλών βάσεων δεδομένων ή αρχείων, πρέπει να επιλυθεί ο πλεονασμός δεδομένων
- Μετασχηματισμός δεδομένων
  - Κανονικοποίηση και άθροιση
- Μείωση δεδομένων
  - Μείωση της αναπαράστασης των δεδομένων σε όγκο με τρόπο που να παράγει τα ίδια (ή παρόμοια) αναλυτικά αποτελέσματα
- Μετατροπή των δεδομένων σε διακριτά (discretization)
  - Μείωση μέρους των δεδομένων (κυρίως των αριθμητικών) με απόδοση ιδιαίτερων τιμών (π.χ. λεκτικών)

Ελλιπείς τιμές

- Αιτίες
  - Πολλά πεδία (μη υποχρεωτικά) δεν συμπληρώνονται ως μη σημαντικά
  - Παλαιότερα δεδομένα διαγράφονται για να εξοικονομηθεί χώρος
- Λύσεις
  - Εισαγωγή προκαθορισμένων τιμών
  - Εισαγωγή τιμών που προκύπτουν από τις υπόλοιπες καταχωρήσεις: μέση τιμή συνολικά, μέση τιμή της κατηγορίας, πιθανή τιμή (;;)
  - Διαγραφή της συναλλαγής που έχει ελλείψεις

Θόρυβος

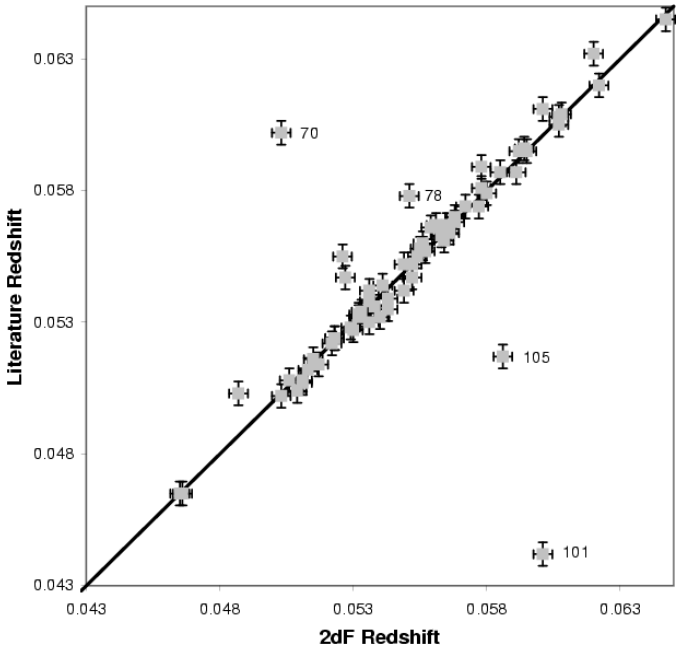
- Διακύμανση στις μετρήσεις μιας μεταβλητής
- Αιτίες
  - Τυχαίο λάθος
  - Προβλήματα στη μετάδοση των δεδομένων
  - Περιορισμοί της τεχνολογίας
- Λύσεις
  - Καδοποίηση (binning): Ταξινόμηση δεδομένων και τμηματοποίηση σε ομάδες ίσης συχνότητας (στην εμφάνιση των τιμών της μεταβλητής), εξομάλυνση (smoothing) με το μέσο όρο (της ομάδας ή συνολικά)
  - Παλινδρόμηση (regression): ταίριασμα των δεδομένων σε μια συνάρτηση παλινδρόμησης
  - Συσταδοποίηση (clustering): και απομάκρυνση των outliers
  - Εντοπισμός και αναγνώριση των εξαιρέσεων (από άνθρωπο)

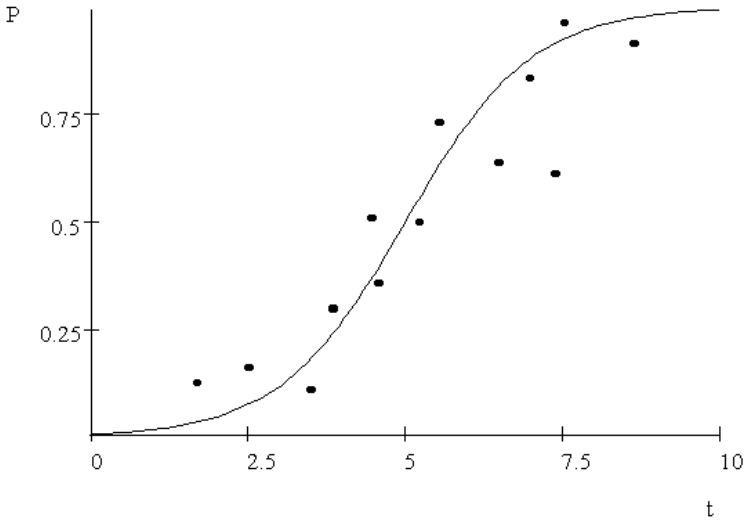
Καδοποίηση και εξομάλυνση

- Τιμές προϊόντος: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Ίσης συχνότητας (μεγέθους) κάδοι:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Εξομάλυνση με το μέσο:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 24, 24, 24, 24
  - Bin 3: 29, 29, 29, 29
- Εξομάλυνση με τα όρια:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Παλινδρόμηση

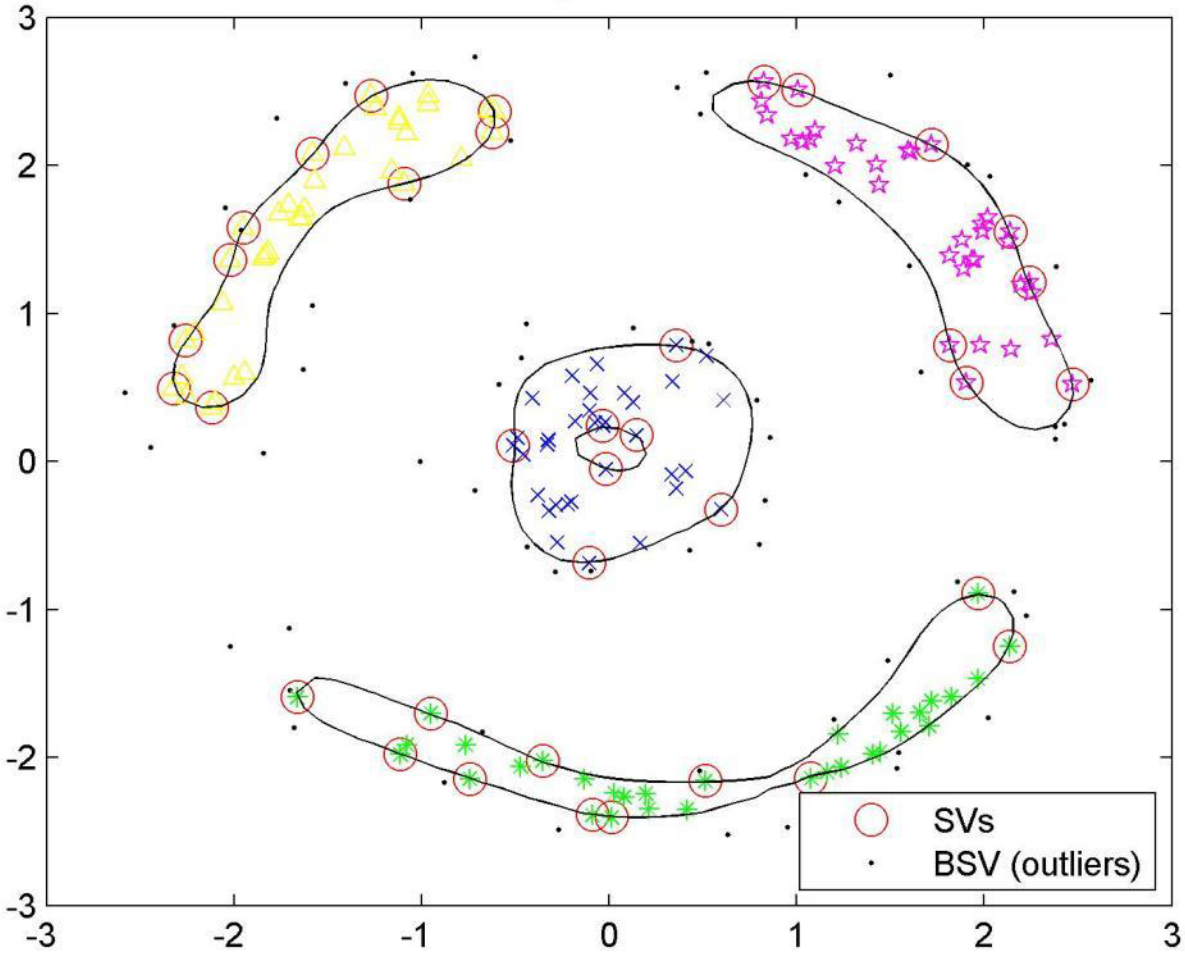
- Η συνάρτηση μπορεί να είναι γραμμική, σιγμοειδής κλπ.





- http://www.stat.uiuc.edu/courses/stat100//cuwu/datalist.html

Συσταδοποίηση



Ολοκλήρωση δεδομένων

- Συνδυάζει δεδομένα από πολλαπλές πηγές σε μια συμπαγή αποθήκη δεδομένων
- Ενιαίο σχήμα ολοκλήρωσης
- Προβλήματα:
  - Συγκρούσεις
  - Πλεονάζοντα δεδομένα
- Επίλυση συγκρούσεων
  - Μπορεί να έχουμε διαφορετικές τιμές για το ίδιο γνώρισμα της ίδιας οντότητας (π.χ. λόγω διαφορετικών μονάδων μέτρησης)

Πλεονασμός δεδομένων

- Εμφανίζεται κατά την ολοκλήρωση
- Αιτίες
  - Έλλειψη μοναδικού αναγνωριστικού: το ίδιο αντικείμενο έχει διαφορετικό αναγνωριστικό σε κάθε βάση
  - Χαρακτηριστικά που προκύπτουν από τις τιμές άλλων
- Λύσεις
  - Ανάλυση συσχέτισης μεταξύ των δεδομένων
  - Ταίριασμα με χρήση των μεταδεδομένων από κάθε πηγή
- Αποτέλεσμα: Βελτίωση ποιότητας/ταχύτητας

Ανάλυση συσχέτισης
Correlation analysis: Pearson's **r**

- Συντελεστής συσχέτισης (Pearson's product moment correlation coefficient)

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(\frac{X_i - \bar{X}}{\sigma_X}\right)\left(\frac{Y_i - \bar{Y}}{\sigma_Y}\right)\right]$$

- n : το πλήθος δειγμάτων
- $\bar{X}$:μέση τιμή
- σ_x: Τυπική απόκλιση (standard deviation)

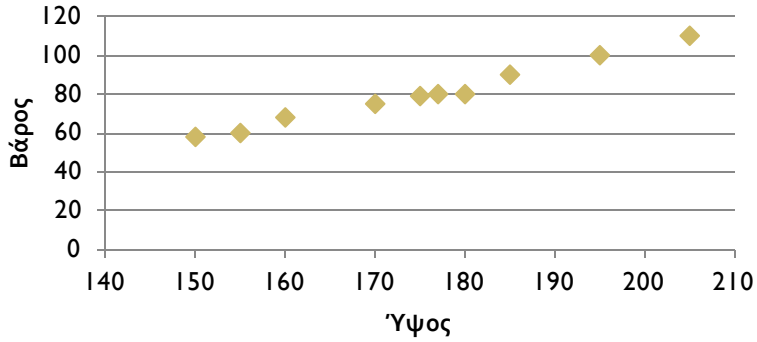
$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

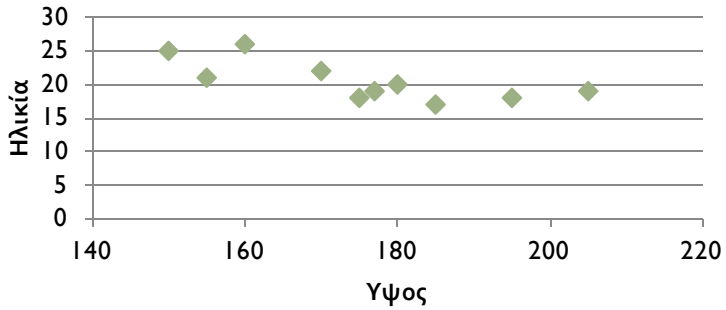
- Αν το r >0, τότε τα γνωρίσματα συσχετίζονται θετικά.
- Μεγάλες τιμές (θετικές ή αρνητικές) του r αναδεικνύουν συσχετιζόμενα γνωρίσματα

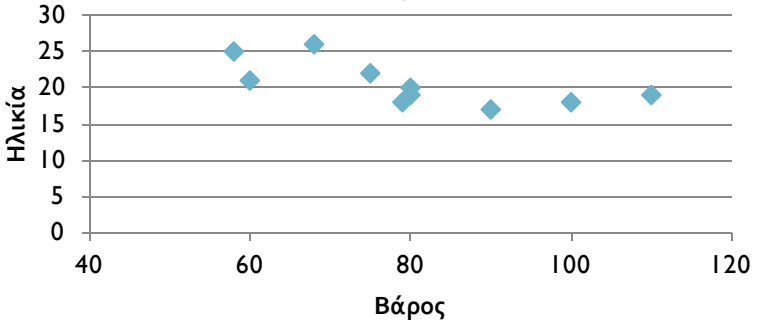
Παράδειγμα

α/α	ύψος	βάρος	ηλικία
1	180	80	20
2	170	75	22
3	175	79	18
4	160	68	26
5	150	58	25
6	205	110	19
7	195	100	18
8	185	90	17
9	155	60	21
10	177	80	19
Μέση τιμή	175,2	80	20,5
Τυπική απόκλιση	17,33205	16,51262	3,02765

r(υ,β)	r(υ,η)	r(β,η)
0,991933	-0,73685	-0,68897







- Το ύψος και το βάρος έχουν θετική συσχέτιση
- Το ύψος και η ηλικία έχουν αρνητική
- Το βάρος και η ηλικία έχουν αρνητική

Μείωση διαστάσεων/δεδομένων

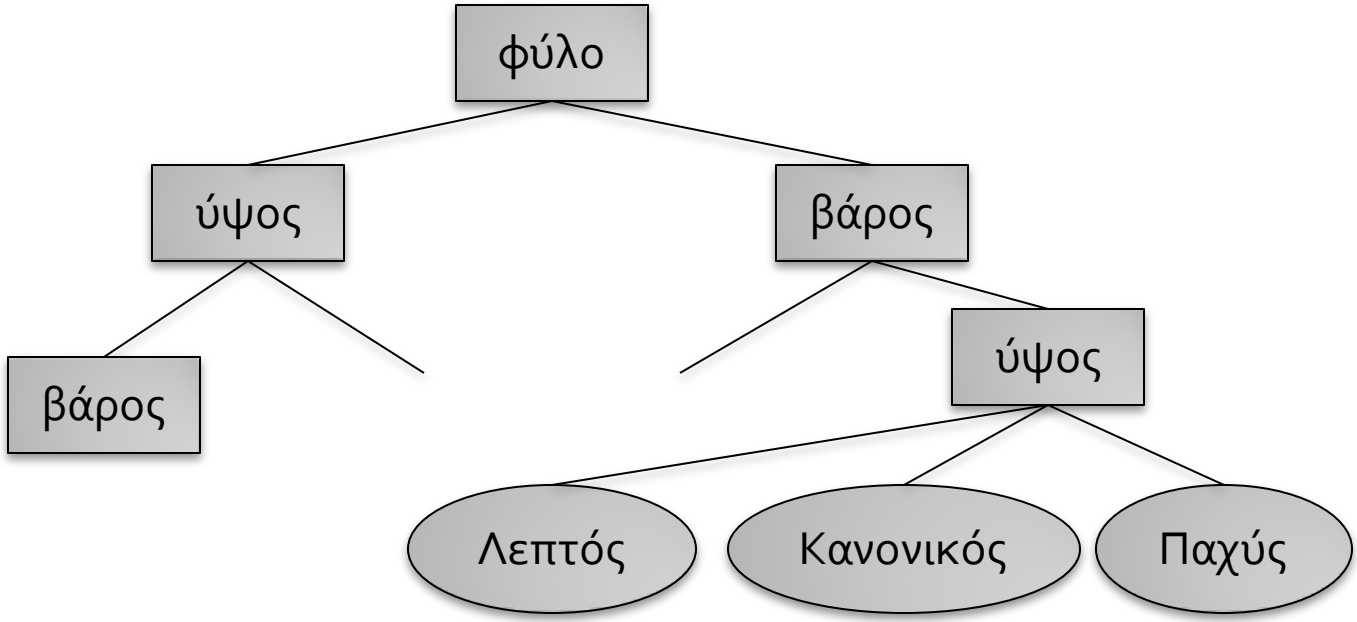
- Μια συλλογή δεδομένων μπορεί να έχει μέγεθος αρκετά Tb και να περιλαμβάνει εκατομμύρια συναλλαγές
- Μια πολύπλοκη ανάλυση/εξόρυξη δεδομένων μπορεί να καθυστερήσει πάρα πολύ
- Προσπαθούμε να βρούμε μια περιορισμένη αναπαράσταση του συνόλου δεδομένων που καταλαμβάνει πολύ μικρότερο όγκο αλλά παράγει τα ίδια αναλυτικά αποτελέσματα

Τεχνικές μείωσης

- Για μείωση διαστάσεων
  - Επιλογή του ελάχιστου συνόλου χαρακτηριστικών
    - Φροντίζουμε οι κατανομές των διαφορετικών ομάδων με βάση τα επιλεγμένα χαρακτηριστικά να είναι κοντά στις αρχικές κατανομές (που περιλαμβάνουν όλα τα χαρακτηριστικά)
- Μείωση πληθυσμού
  - Ταίριασμα δεδομένων με μοντέλα
  - Παλινδρόμηση, συσταδοποίηση, δειγματοληψία
- Άλλες τεχνικές
  - Συμπίεση δεδομένων
  - Αναπαράσταση με διακριτές τιμές (discretization) και εννοιολογική ιεραρχία

Παράδειγμα (μείωση διαστάσεων)

- Αρχικό σύνολο χαρακτηριστικών: {ύψος, περιφ.μέσης, μήκος χεριών, βάρος, ηλικία, φύλο}



- Μειωμένο σύνολο: {φύλο, ύψος, βάρος}

Μετασχηματισμοί

Τύποι μεταβλητών

- Κατηγορικές μεταβλητές (categorical ή nominal): δύο ή περισσότερες τιμές χωρίς συγκεκριμένη διάταξη (π.χ. τμήμα, πανεπιστήμιο, φύλο, κλπ)
- Τακτικές μεταβλητές (ordinal): δύο ή περισσότερες τιμές με συγκεκριμένη διάταξη αλλά άνισες αποστάσεις (π.χ. πτυχίο, μεταπτυχιακό, διδακτορικό ή πόσο ικανοποιημένος είστε από 1 ως 5)
- Μεταβλητές διαστήματος (interval): Οι τιμές έχουν διάταξη και οι αποστάσεις ανάμεσα στις κατηγορίες είναι ίσες (π.χ. θερμοκρασία, πίεση, pH κλπ)
- Μεταβλητές αναλογίας (ratio): Έχουν τις ιδιότητες μιας μεταβλητής διαστήματος και επιπλέον το 0.0 έχει ιδιαίτερη σημασία (π.χ. ύψος, βάρος κλπ). Επίσης διπλάσια τιμή σημαίνει διπλάσια ένταση (δε συμβαίνει το ίδιο με τη θερμοκρασία)

Επιτρεπτοί υπολογισμοί

- Ανάλογα με τον τύπο των μεταβλητών μπορούμε να υπολογίσουμε και συγκεκριμένες «χαρακτηριστικές» τιμές

	Nominal	**Ordinal**	**Interval**	**Ratio**
Frequency distribution	Ναι	Ναι	Ναι	Ναι
Median **Percentiles**	Όχι	Ναι	Ναι	Ναι
Add **Subtract**	Όχι	Όχι	Ναι	Ναι
Mean **standard deviation** **standard error of the mean**	Όχι	Όχι	Ναι	Ναι
ratio **coefficient of variation**	Όχι	Όχι	Όχι	Ναι
Μετασχηματισμοί	Οποιοδήποτε mapping	Κάθε mapping που διατηρεί τη διάταξη	Πολλαπλασιασμός ή προσθήκη σταθερής τιμής	Πολλαπλασιασμός με σταθερή τιμή

Μετασχηματισμοί δεδομένων

- Γιατί μετασχηματισμοί;
  - Τα δεδομένα που καταγράφουμε για τις μεταβλητές μας μπορεί να έχουν συνεχείς ή διακριτές τιμές, λεκτικές ή αριθμητικές ανάλογα με τη φύση του προβλήματος
  - Οι αλγόριθμοι που χρησιμοποιούμε δέχονται καθορισμένα είδη τιμών π.χ. πραγματικές τιμές. Τέτοιοι αλγόριθμοι απαιτούν μεταβλητές διαστήματος ή αναλογίας
- Παράδειγμα: Θέλουμε να επιλέξουμε φοιτητές για ένα μεταπτυχιακό με βάση το βαθμό πτυχίου, το τμήμα, το πανεπιστήμιο κλπ
  - Τμήμα: Πληροφορική, Οικονομικό, Μαθηματικό,… ή 1,2,3,…
  - Βαθμός: {7.2, 8.79, 5.54,…} ή {λίαν καλώς, άριστα, καλώς,…} ή {7, 9, 6, …}

Διαχείριση τιμών

- Στην πραγματικότητα έχουμε μικτά σύνολα μεταβλητών
- Οι περισσότεροι αλγόριθμοι υποστηρίζουν μεταβλητές διαστήματος (interval).
- Εξαίρεση αποτελούν τα δέντρα απόφασης
- Πως χειριζόμαστε τις μη αριθμητικές τιμές;
- Μια κατηγορική μεταβλητή με Μ τιμές
  - Δεν μπορεί να απεικονιστεί σε τιμές από 1 ως Μ γιατί δεν έχουν νόημα οι μετασχηματισμοί: π.χ. κ*επάγγελμα+λ*σπουδές
  - Μπορούμε να χρησιμοποιήσουμε Μ binary μεταβλητές. Αλλά έχουμε πρόβλημα για μεγάλες τιμές του Μ

Αναπαράσταση δεδομένων

- Ένα συχνό μοντέλο αναπαράστασης των δεδομένων ενός προβλήματος είναι το διανυσματικό
- Αν έχω d παραμέτρους/μεταβλητές θα έχω κι ένα διάνυσμα με d διαστάσεις

- Παράδειγμα:

  <θερμοκρασία, υγρασία, εποχή>

  <25, 90, καλοκαίρι>, <20, 50, άνοιξη>, …

Τεχνικές μετασχηματισμού

- Εξομάλυνση: απομάκρυνση θορύβου
- Άθροιση (aggregation): περίληψη
- Γενίκευση: εννοιολογική ιεραρχία
- Κανονικοποίηση
  - min-max Normalization [newmin..newmax]
  - z-score Normalization
  - Decimal scaling
- Δημιουργία νέων χαρακτηριστικών από τα υπάρχοντα

Κανονικοποίηση

- Min-max Normalization

$$newA = \frac{(oldA - minA)}{(maxA - minA)} * (newmax - newmin) + newmin$$

- z-score Normalization

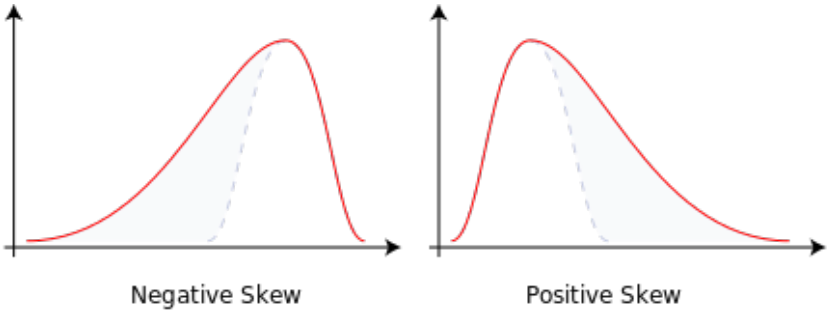
$$newA = \frac{oldA - \bar{A}}{\sigma_A}$$

- Decimal scaling

$$newA = \frac{oldA}{10^j}$$

όπου j ο μικρότερος ακέραιος για τον οποίο Max(|newA|)<1

Deskewing



- Συχνά οι τιμές που λαμβάνουν κάποιες μεταβλητές έχουν μια έμφαση προς ένα συγκεκριμένο υπο-πεδίο τιμών (π.χ. το εισόδημα)
- Μετασχηματισμός για τιμές στο 0..1

$$\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p).$$

$$\operatorname{logit}^{-1}(\alpha) = \frac{1}{1+\exp(-\alpha)} = \frac{\exp(\alpha)}{\exp(\alpha)+1}$$

Διακριτοποίηση - Discretization

- Οι τιμές θερμοκρασίας που καταγράφονται από ένα μετεωρολογικό σταθμό είναι συνεχόμενες σε μια κλίμακα από -40 μέχρι +50. Ομοίως οι τιμές υγρασίας, βροχόπτωσης κλπ.
- Η μετατροπή τους σε διακριτές τιμές μπορεί να γίνει με αντικατάσταση των επιμέρους πεδίων τιμών με λεκτικά (-40..-10→πολύ κρύο, -10..5→ κρύο κλπ)
- Η ώρα καταγραφής (timestamp) επίσης είναι ένα συνεχόμενο μέγεθος που μπορεί να μετατραπεί σε διακριτό, π.χ. πρωί/βράδυ, ημέρα, εβδομάδα, εποχή κλπ

Συνοψίζοντας

- Η προεπεξεργασία των δεδομένων είναι σημαντική για την ποιότητα των αποτελεσμάτων
- Περιλαμβάνει
  - Καθαρισμό και ολοκλήρωση των δεδομένων
  - Μείωση διαστάσεων και επιλογή χαρακτηριστικών
  - Μετατροπή τιμών σε διακριτές (discretization)

Μέτρα απόστασης

Βασική απαίτηση

- Ένας τύπος απόστασης d(x,y) είναι μετρική (μέτρο) απόστασης όταν
  - $d(x, y) \geq 0$
  - $d(x, y) = 0$ αν και μόνο αν $x = y$
  - $d(x, y) = d(y, x)$
  - $d(x, z) \leq d(x, y) + d(y, z)$ (τριγωνική ανισότητα)

Πόσο απέχουν;

Sex	Age	Time Addr	ResStat	occup	Time Emp	Time Bank	House Exp	Decision
M	50	0.5	owner	unemploye	0	0	00145	reject
M	19	10	rent	labourer	0.8	0	00140	reject
F	52	15	owner	creative_	5.5	14	00000	accept
M	22	2.5	rent	creative_	2.6	0	00000	accept
M	29	13	owner	driver	0.5	0	00228	reject

Ευκλείδεια απόσταση

- Συμβολισμοί: n αντικείμενα με d χαρακτηριστικά (i=1..n)

  $x(i)=(x_1(i), x_2(i),\ldots, x_d(i))$

- Το πιο γνωστό μέτρο απόστασης είναι η Ευκλείδεια απόσταση:

$$d_E(i,j) = \sqrt{\sum_{k=1}^{d} \left(x_k(i) - x_k(j)\right)^2}$$

- Έχει νόημα όταν όλες οι μεταβλητές έχουν μετρηθεί στις ίδιες μονάδες (commensurate variables)
- Ισχύει όταν οι μεταβλητές συνεισφέρουν ανεξάρτητα στο μέτρο απόστασης (ορθογώνιος χώρος)

Κλιμάκωση & Βάρη

- Για μεταβλητές σε διαφορετικά μέτρα χρησιμοποιούμε κλιμάκωση με βάση την τυπική απόκλιση

$$d_{E_{scaled}}(i,j) = \sqrt{\sum_{k=1}^{d} \frac{\left(x_k(i) - x_k(j)\right)^2}{\sigma_k^2}}$$

- όπου σ_k η τυπική απόκλιση

$$\sigma_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_k(i) - \bar{x}_k)^2}$$

- Αν γνωρίζουμε τη σημαντικότητα κάθε μεταβλητής βάζουμε βάρη

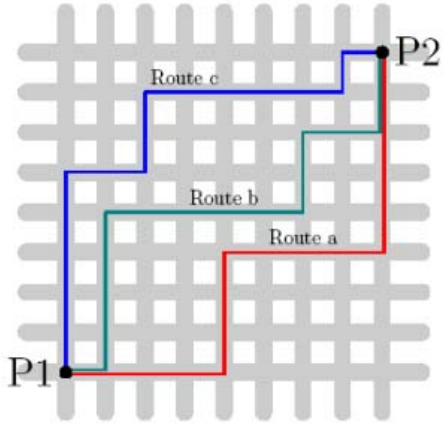
$$d_{E_{weighted}}(i,j) = \sqrt{\sum_{k=1}^{d} w_k\left(x_k(i) - x_k(j)\right)^2}$$

Επεκτάσεις

- P-norm ή απόσταση L_p(p>=1):

$$d(i,j) = \left(\sum_{k=1}^{d} |x_k(i) - x_k(j)|^p \right)^{1/p}$$

- Για p=2 έχουμε την Ευκλείδεια
- Για p=1 έχουμε την απόσταση Manhatan



$$d_{manhattan}(i,j) = \sum_{k=1}^{d} |x_k(i) - x_k(j)|$$

Εξάρτηση μεταξύ μεταβλητών

- Για να εξετάσουμε τη γραμμική εξάρτηση μπορούμε να χρησιμοποιήσουμε τις μετρικές covariance ή correlation
- Covariance: Μετρά κατά πόσο δύο μεγέθη μεταβάλλονται μαζί
- Έστω δύο μεταβλητές X και Y και n αντικείμενα με τιμές x(1), …, x(n) και y(1), …, y(n). Η covariance των X και Y είναι:

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y})$$

- Εξαρτάται από το εύρος του πεδίου τιμών των X και Y

Correlation

- Συνήθως διαιρούμε με το standard deviation

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y})}{\left(\sum_{i=1}^{n}(x(i)-\bar{x})^2 \sum_{i=1}^{n}(y(i)-\bar{y})^2\right)^{1/2}} \qquad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

Rank correlation coefficients

- Spearman correlation coefficient
  - Για δύο βαθμονομήσεις n στοιχείων, όπου $d_i = x_i - y_i$ η διαφορά θέσης κάθε στοιχείου

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

- Kendall rank correlation coefficient

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}.$$

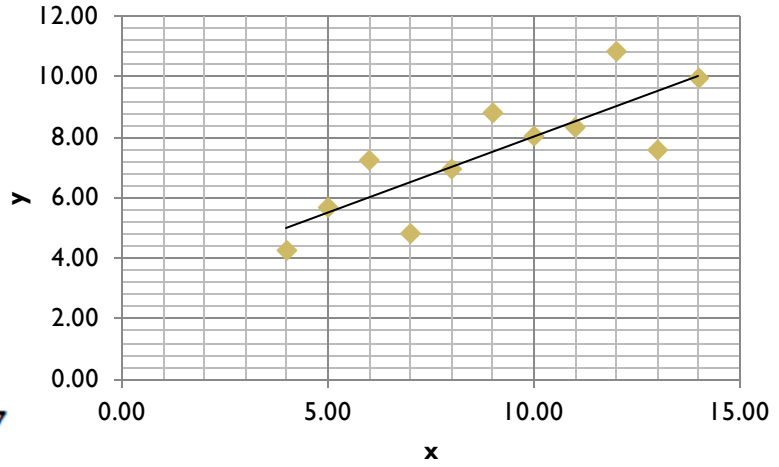
Παράδειγμα

- Δεδομένα

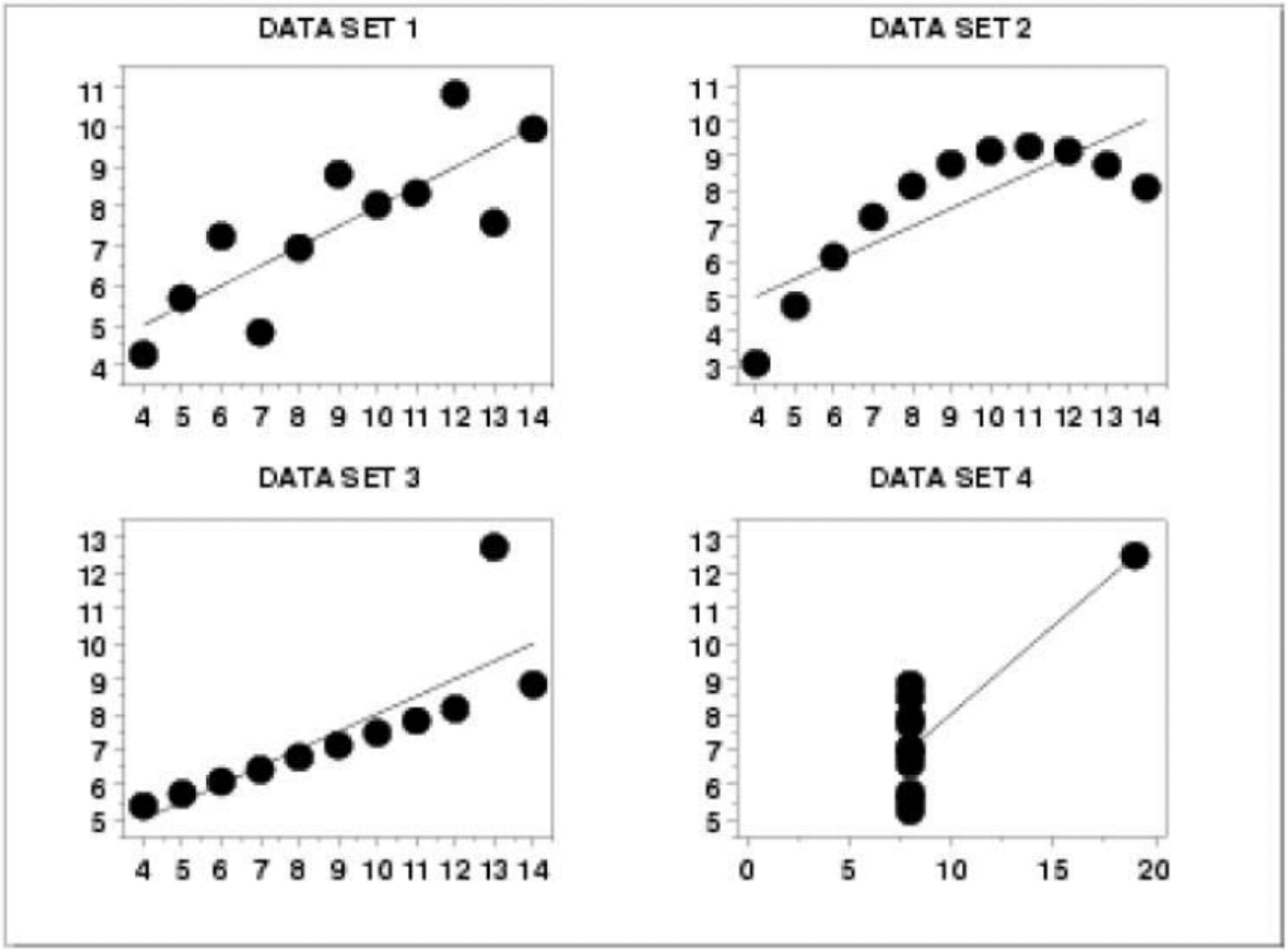
X	10,00	8,00	13,00	9,00	11,00	14,00	6,00	4,00	12,00	7,00	5,00
Y	8,04	6,95	7,58	8,81	8,33	9,96	7,24	4,26	10,84	4,82	5,68

Αποτελέσματα

$N = 11$
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept = 3
Slope = 0.5
Residual standard deviation = 1.237
Correlation = 0.816



Οπτικοποίηση



Πίνακες correlation/covariance

- Αν έχουμε d μεταβλητές, μπορούμε να υπολογίσουμε d^2 συνδυασμούς συσχέτισης
- Ο πίνακας συσχέτισης Σ είναι d x d
  - d στοιχεία στη διαγώνιο που περιέχουν τις διακυμάνσεις (variances) κάθε μεταβλητής
  - Συμμετρικός: covariance(i,j)=covariance(j,i)

Απόσταση Mahalanobis

- Λαμβάνει υπόψη της το scaling κάθε άξονα
- Διορθώνει τη συσχέτιση διαφορετικών χαρακτηριστικών
- Υποθέτει ότι όλα τα ζεύγη μεταβλητών είναι σχεδόν linearly correlated

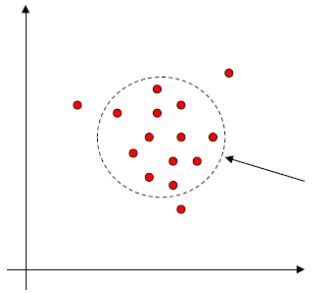
$$d_{MH}(x,y) = \sqrt{(x-y)^T \sum^{-1} (x-y)}$$

Vector difference in d-dimensional space

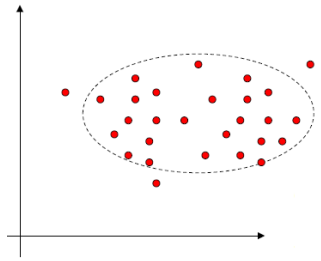
Inverse covariance matrix

Παραλλαγές

- Αν
  - Ο covariance matrix είναι διαγώνιος και ισοτροπικός
  - Όλες οι διαστάσεις έχουν ίσες variance
  - Η Mahalanobis εκφυλίζεται σε Ευκλείδεια



- Αν
  - Ο covariance matrix είναι διαγώνιος και μη-ισοτροπικός
  - Οι διαστάσεις δεν έχουν ίσες variance
  - Η Mahalanobis εκφυλίζεται σε Ευκλείδεια με βάρη



Για binary vectors

	j=1	j=0
i=1	n_{11}	n_{10}
i=0	n_{01}	n_{00}

- Matching coefficient $$\frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}$$

- Jaccard similarity coefficient

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

Άλλες μετρικές απόστασης

- Για κατηγορικές μεταβλητές
  - Ο αριθμός των τιμών που ταιριάζουν δια το συνολικό αριθμό διαστάσεων
- Αποστάσεις μεταξύ string διαφορετικού μήκους
  - Σημασιολογικές αποστάσεις
- Αποστάσεις μεταξύ εικόνων και κυματομορφών
  - Μας ενδιαφέρει να μην επηρεάζονται από μετασχηματισμούς (π.χ. μετακινήσεις, κλιμάκωση, περιστροφή κλπ)