Αναγνώριση Προτύπων

Προ-επεξεργασία δεδομένων

Βασικές έννοιες

Ανδρέας Λ. Συμεωνίδης

Αν. Καθηγητής

Τμήμα Ηλεκτρολόγων Μηχ/κών

&

Μηχ/κών Υπολογιστών, Α.Π.Θ.

Email: asymeon@eng.auth.gr

issel
Intelligent Systems & Software Engineering Labgroup

ARISTOTLE UNIVERSITY OF THESSALONIKI

Διάρθρωση διάλεξης

- Τύποι χαρακτηριστικών
- Ιδιότητες χαρακτηριστικών
- Θόρυβος, Ελλειπούσες τιμές
- Μείωση διαστάσεων
- Ομοιότητα δεδομένων

Ο όρος «Δεδομένα»

- Μια συλλογή από σημεία με τα χαρακτηριστικά τους
- Ένα **χαρακτηριστικό** είναι μια ιδιότητα ενός αντικειμένου
  - Παραδείγματα: το χρώμα των ματιών, η θερμοκρασία, κλπ.
  - Ένα χαρακτηριστικό είναι επίσης γνωστό ως μεταβλητή, συνιστώσα, ή πεδίο
- Μια συλλογή από χαρακτηριστικά περιγράφουν ένα **αντικείμενο**
  - Επίσης γνωστό ως εγγραφή, σημείο, δείγμα, ή στιγμιότυπο

Χαρακτηριστικά

Αντικείμενα

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Τύποι χαρακτηριστικών

- Οι διαφορετικοί τύποι δεδομένων:
  - **Κατηγορικές (Nominal)**
    - Παράδεγμα: χρώμα ματιών
  - **Τακτικές (Ordinal)**
    - Παράδειγμα: ταξινομήσεις (π.χ., βαθμοί εξετάσεων στην κλίμακα 1-10)
  - **Διαστημάτων (Interval)**
    - Παράδειγμα: Θερμοκρασία σε Κελσίου ή Fahrenheit.
  - **Αναλογίας (Ratio)**
    - Παράδειγμα: Χρόνος

Ιδιότητες των τιμών των χαρακτηριστικών

- Ο τύπος ενός χαρακτηριστικού εξαρτάται από το ποιές από τις παρακάτω ιδιότητες παρουσιάζει:
  - Διακριτότητα (Distinctness): = ≠
  - Τάξη (Order): < >
  - Πρόσθεση (Addition): + -
  - Πολλαπλασιασμός (Multiplication): * /

  - Nominal χαρακτηριστικά → διακριτότητα
  - Ordinal χαρακτηριστικά → διακριτότητα & τάξη
  - Interval χαρακτηριστικά → διακριτότητα, τάξη & πρόσθεση
  - Ratio χαρακτηριστικά → και τις 4 ιδιότητες

Ιδιότητες των τιμών των χαρακτηριστικών

Τύπος	Πράξεις	Πράξεις	Μετασχηματισμός
Nominal	= ≠	mode, entropy, contingency correlation, χ^2 test	Οποιοσδήποτε
Ordinal	= ≠ < >	median, percentiles, rank correlation, run tests, sign tests	*new_value* = *f*(*old_value*) *f:* monotonic function
Interval	= ≠ < > + -	mean, standard deviation, Pearson's correlation, *t* and *F* tests	*new_value* = a*old_value + b a και b σταθερές
Ratio	= ≠ < > + - * /	geometric mean, harmonic mean, percent variation	*new_value* = a * *old_value*

Διαφορετικές αναπαραστάσεις δεδομένων

- Το πεδίο εφαρμογής καθορίζει και τη μορφή των δεδομένων!!!
- Εγγραφές (Record data)
  - Πίνακας δεδομένων (Data Matrix)
  - Δεδομένα από κείμενο (Document Data)
  - Δεδομένα συναλλαγών (Transaction Data)
- Γράφοι (Graph data)
  - Δεδομένα διαδικτύου (World Wide Web data)
  - Δομές μορίων (Molecular Structures)
- Ταξινομημένα (Ordered data)
  - Χωρικά δεδομένα (Spatial Data)
  - Χρονικά δεδομένα (Temporal Data)
  - Ακολουθιακά δεδομένα (Sequential Data)
  - Γενετικά δεδομένα (Genetic Sequence Data)

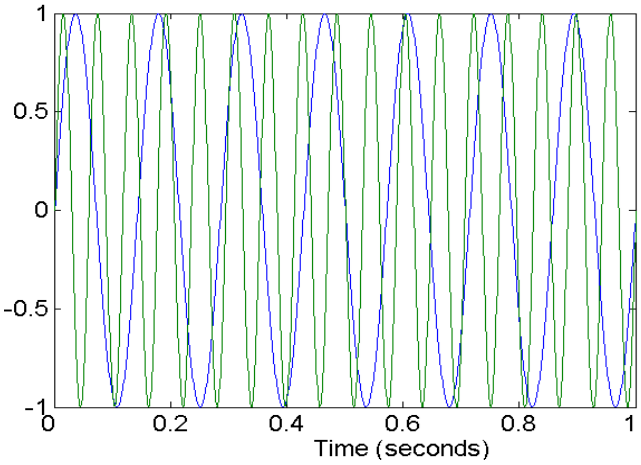
Βασικές ιδιότητες των σετ δεδομένων

- Διαστασιμότητα (Dimensionality)
  - Η κατάρα της διαστασιμότητας (Curse of Dimensionality)
- Αραιότητα (Sparsity)
  - Μόνο η παρουσία μετρά
- Ανάλυση (Resolution)
  - Τα πρότυπα βασίζονται στην κλίμακα

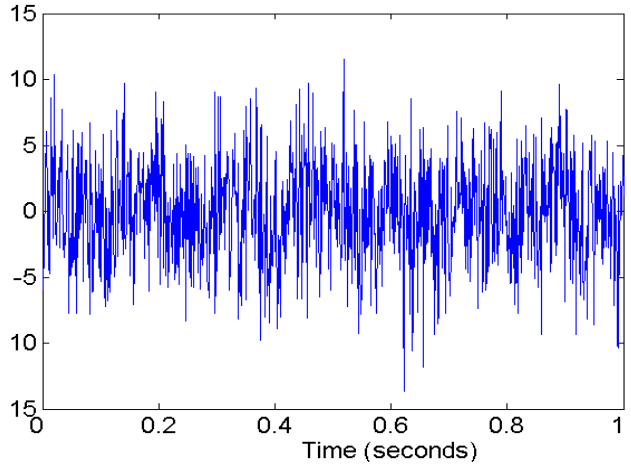
Ποιότητα δεδομένων

- Θόρυβος και εξωκείμενες τιμές
- Ελλειπούσες τιμές (Missing values)
  - Στη φάση της δημιουργίας του μοντέλου
  - Στη φάση της ταξινόμησης μιας νέας εγγραφής
- Διπλές εγγραφές
  - Σημαντικό θέμα όταν τα δεδομένα έρχονται από ετερεογενείς πηγές
  - Απαιτείται καθαρισμός δεδομένων

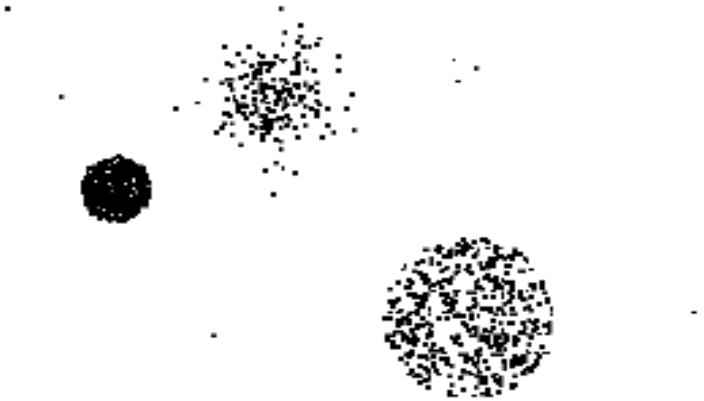
Θόρυβος και Εξωκείμενες τιμές



Two Sine Waves



Two Sine Waves + Noise



Ελλειπούσες τιμές

- Λόγοι ύπαρξης ελλειπουσών τιμών
  - Πληροφορία που δε συλλέγεται
    (π.χ. Χρήστες συμπληρώνουν μόνο τα υποχρεωτικά πεδία σε μια φόρμα)
  - Χαρακτηριστικά που δεν εφαρμόζονται σε όλα τα δεδομένα
    (π.χ., το ετήσιο εισόδημα δεν εφαρμόζεται σε περιπτώσεις μικρών παιδιών)
- Διαχείριση ελλειπουσών τιμών
  - Διαγραφή των εγγραφών με ελλειπείς τιμές
  - Πρόβλεψη των ελλειπουσών τιμών
  - Αγνόηση των ελλειπουσών τιμών κατά την ανάλυση
  - Αντικατάσταση με όλες τις πιθανές τιμές (βεβαρυμένες από την πιθανότητά τους)

Διαχείριση κατά την εκπαίδευση (1)

Tid	Refund	Marital Status	Taxable Income	Class
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	?	Married	90K	Yes

Ελλειπούσα τιμή

Πριν το διαχωρισμό:

Entropy(Parent)

= -0.3 log(0.3)-(0.7)log(0.7) = 0.8813

	Class = Yes	Class = No
Refund=Yes	**0**	**3**
Refund=No	**2**	**4**
Refund=?	**1**	**0**

Διαχωρισμός στο Refund:

Entropy(Refund=Yes) = 0

Entropy(Refund=No)

= -(2/6)log(2/6) – (4/6)log(4/6) = 0.9183

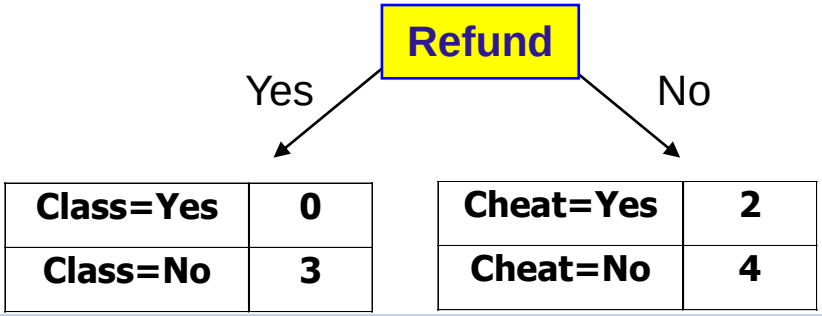
Entropy(Children)

= 0.3 (0) + 0.6 (0.9183) = 0.551

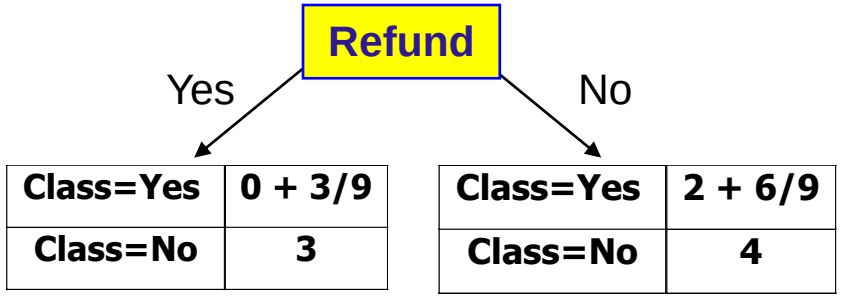
Gain = 0.9 × (0.8813 – 0.551) = 0.3303

Διαχείριση κατά την εκπαίδευση (2)

Tid	Refund	Marital Status	Taxable Income	Class
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



Tid	Refund	Marital Status	Taxable Income	Class
10	?	Married	90K	Yes



P(Refund=Yes) = 3/9

P(Refund=No) = 6/9

Ταξινομήστε την εγγραφή στο αριστερό παιδί με βάρος = 3/9 και στο δεξί με βάρος = 6/9

Προεπεξεργασία δεδομένων

- Συνάθροιση (Aggregation)
- Δειγματοληψία (Sampling)
- Μείωση διαστάσεων (Dimensionality Reduction)
- Επιλογή χαρακτηριστικών (Feature subset selection)
- Δημιουργία χαρακτηριστικών (Feature creation)
- Διακριτοποίηση (Discretization) και Ψηφιοποίηση (Binarization)
- Μετασχηματισμός χαρακτηριστικών (Attribute Transformation)

Συνάθροιση

- Συνδυασμός δυο ή περισσότερων χαρακτηριστικών (ή εγγραφών) σε ένα (μια)
- Στόχος:
  - Μείωση δεδομένων
    - Μείωση του αριθμού των χαρακτηριστικών ή εγγραφών
  - Αλλαγή κλίμακας
    - Πόλεις συναθροίζονται σε περιοχές, χώρες κτλ
  - Πιο “σταθερά” δεδομένα
    - Τα συναθροισμένα δεδομένα έχουν συνήθως μικρότερες διακυμάνσεις

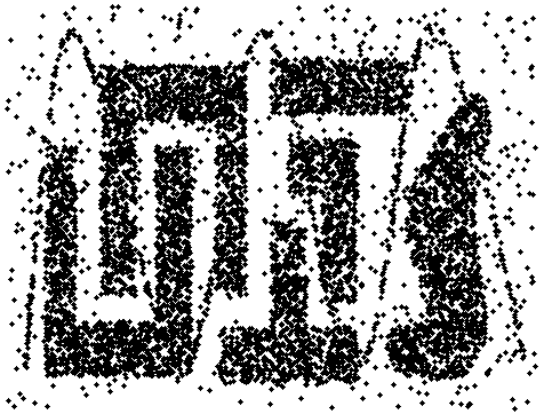
Δειγματοληψία

- Η δειγματοληψία είναι η βασική τεχνική για επιλογή δεδομένων.
- Χρησιμοποιείται και στο στάδιο της προκαταρκτικής εξέτασης, και στην τελική φάση ανάλυσης των δεδομένων.
- Η χρήση ολόκληρου του σετ δεδομένων είναι πολλές φορές υπολογιστικά και χρονιά απαγορευτική.
- Βασική αρχή για αποτελεσματική δειγματοληψία είναι η παρακάτω:
  - Το αποτέλεσμα ενός δείγματος θα είναι όσο καλό θα ήταν και το αποτέλεσμα ολόκληρου του σετ δεδομένων, εάν το δείγμα είναι αντιπροσωπευτικό.
  - Ένα δείγμα είναι αντιπροσωπευτικό αν εσωκλείει τις ίδιες ιδιότητες (ενδιαφέροντος) με το αρχικό σετ δεδομένων.

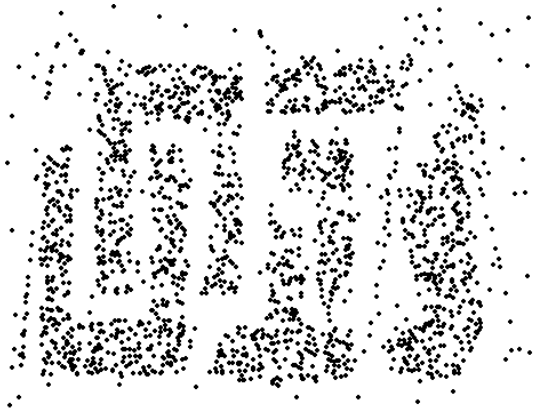
Τύποι δειγματοληψίας

- Απλή, τυχαία δειγματοληψία (Simple Random Sampling)
  - Ίση πιθανότητα να επιλεγεί οποιαδήποτε εγγραφή
- Δειγματοληψία χωρίς αντικατάσταση (Sampling without replacement)
  - Όταν επιλέγεται μια εγγραφή, αφαιρείται από τον πληθυσμό
- Δειγματοληψία με αντικατάσταση (Sampling with replacement)
  - Όταν επιλέγεται μια εγγραφή, μένει στον πληθυσμό (μια εγγραφή μπορεί να επιλεγεί περισσότερες απόμια φορές)
- Τμηματική δειγματοληψία (Stratified sampling)
  - Διαχωρισμός του σετ δεδομένων σε τμήματα – επιλογή δειγμάτων τυχαία από τα τμήματα

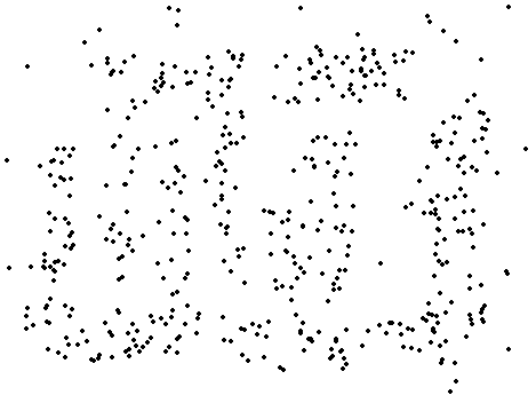
Μέγεθος του δείγματος



8000 σημεία



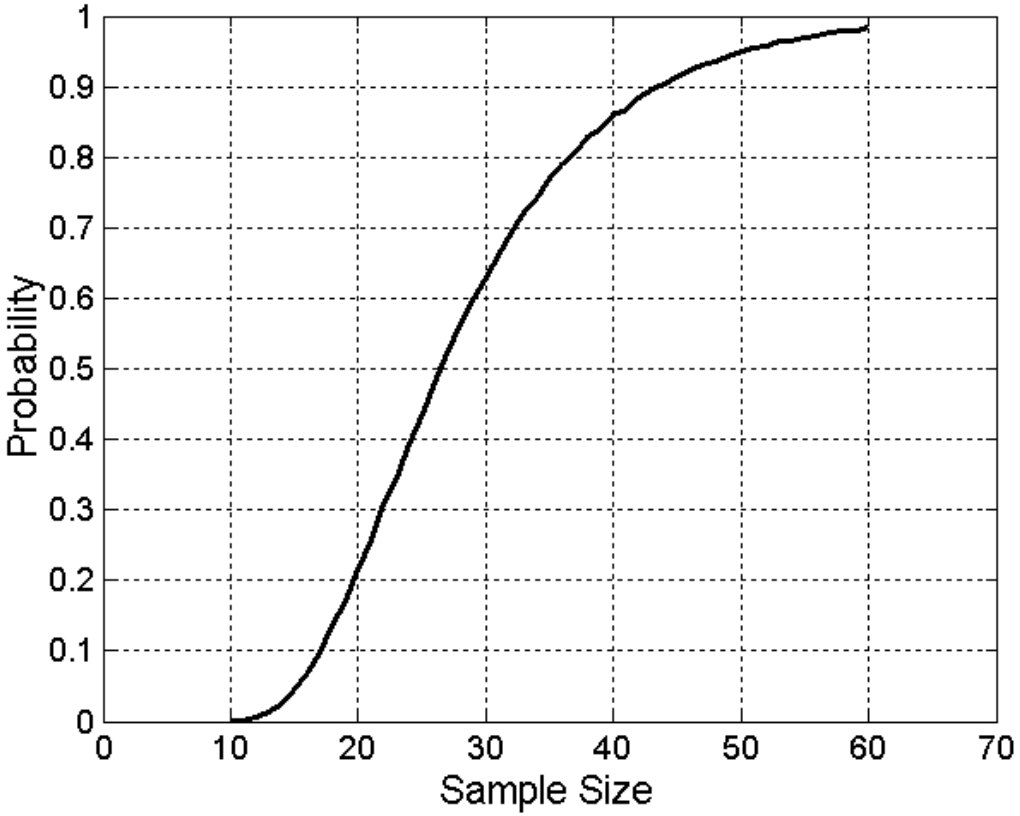
2000 σημεία



500 σημεία

Μέγεθος του δείγματος

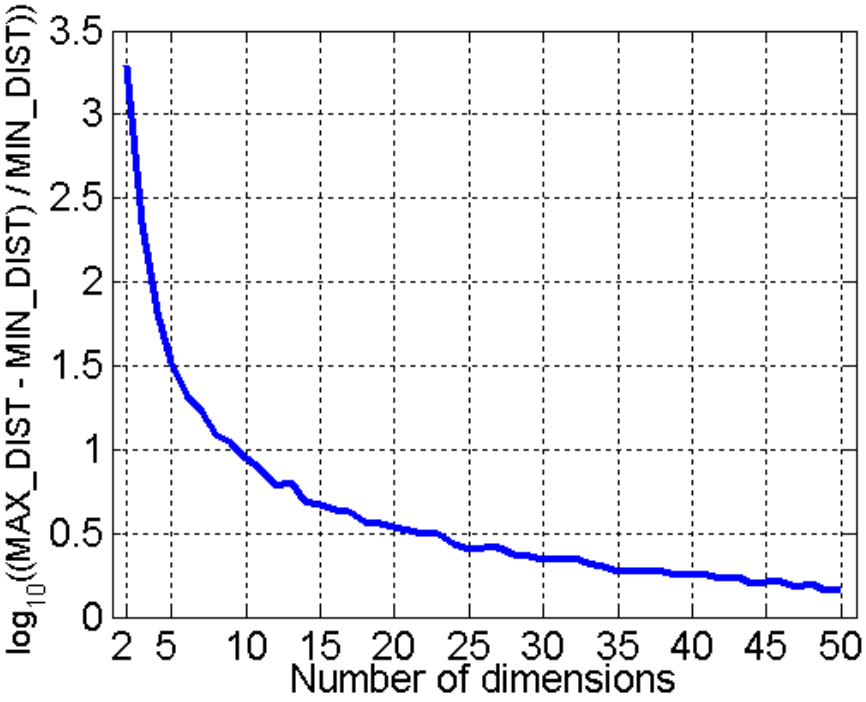
- **Ποιο είναι το απαραίτητα μέγεθος δείγματος, ώστε να έχουμε τουλάχιστον ένα σημείο από κάθε κλάση;**



Η κατάρα της διαστασιμότητας

- Όταν αυξάνονται οι διαστάσεις, τα δεδομένα γίνονται αυξανόμενα αραιά

- Οι ορισμοί πυκνότητας και απόστασης ανάμεσα στα σημεία τα οποία είναι σημαντικά κριτήρια για ομαδοποίηση και ανίχνευσης outliers, χάνουν το νόημά τους



- Δημουργήστε τυχαία 500 σημεία
- Υπολογίστε την max και min απόσταση ανάμεσα σε οποιοδήποτε ζεύγος σημείων

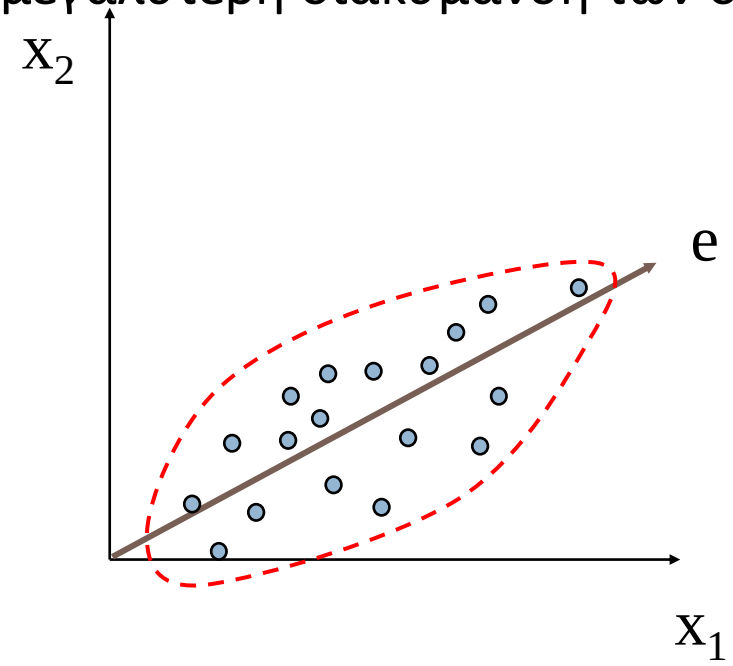
Μείωση διαστάσεων

- **Στόχος:**
  - Αποφυγή της κατάρας της διαστασιμότητας
  - Μείωση του χρόνου και μνήμης που χρειάζεται για την εφαρμογή των αλγορίθμων
  - Βοηθά στην ευκολότερη οπτικοποίηση δεδομένων
  - Μπορεί να βοηθήσει στην απαλοιφή μη σχετικών χαρακτηριστικών και τη μείωση του θορύβου

- **Τεχνικές:**
  - Principle Component Analysis
  - Singular Value Decomposition
  - Άλλες: επιβλεπόμενες και μη-γραμμικές τεχνικές

Μείωση διαστάσεων: PCA

- Στόχος είναι η εύρεση μιας προβολής η οποία περιλαμβάνει όσο το δυνατόν μεγαλύτερη διακύμανση των δεδομένων



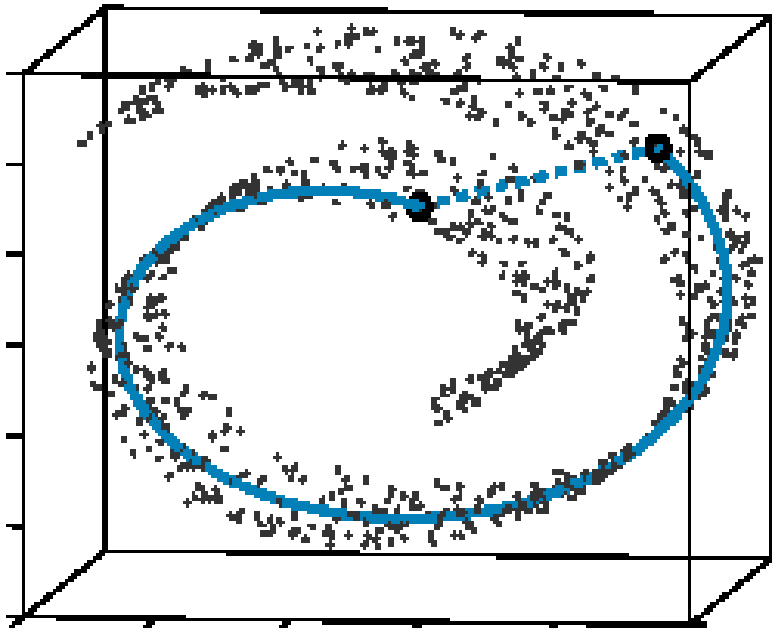
- Εύρεση των ιδιοδιανυσμάτων του πίνακα συμμεταβλητότητας
- Τα ιδιοδιανύσματα ορίζουν τον νέο χώρο του προβλήματος

Μείωση διαστάσεων: PCA

Dimensions = 206

Μείωση διαστάσεων: ISOMAP

By: Tenenbaum, de Silva, Langford (2000)



- Δημιουργία ενός γράφου γειτνίασης
- Για κάθε ζεύγος σημείων στο γράφο, υπολογισμός της κοντινότερης διαδρομής ανάμεσα στα σημεία

Επιλογή χαρακτηριστικών

- Διαφορετική προσέγγιση μείωσης των διαστάσεων των δεδομένων
- Περιττά χαρακτηριστικά
  - Ύπαρξη πολλαπλής πληροφορίας για το ίδιο θέμα περισσότερες τις μιας φορές
  - Παράδειγμα: ποσοστό και ποσό έκπτωσης ενός προϊόντος δεδομένης της τιμής αγοράς του
- Μη σχετικά χαρακτηριστικά
  - Δεν περιέχουν πληροφορία χρήσιμη για το πρόβλημα προς επίλυση
  - Παράδειγμα: Το ΑΕΜ των φοιτητών είναι άσχετο με τη διαδικασία πρόβλεψης του τελικού βαθμού τους

Δημιουργία χαρακτηριστικών

- Δημιουργία νέων χαρακτηριστικών που απεικονίζουν χρήσιμη πληροφορία καλύτερα από τα υπάρχοντα χαρακτηριστικά
- Τρεις προσεγγίσεις:
  - Εξαγωγή χαρακτηριστικών (Feature Extraction)
    - Συνδεδεμένη με το πεδίο εφαρμογής
  - Απεικόνιση των δεδομένων σε νέο χώρο (βλέπε SVMs)
  - Κατασκευή χαρακτηριστικών (Feature Construction)
    - Συνδυασμός χαρακτηριστικών

Ομοιότητα και Ανομοιότητα

- **Ομοιότητα (Similarity)**
  - Αριθμητική μετρική του κατά πόσο όμοια είναι δυο αντικείμενα
  - Συνήθως στο εύρος [0,1]
  - Είναι υψηλότερη όταν τα αντικείμενα «μοιάζουν» περισσότερο
- **Ανομοιότητα (Dissimilarity)**
  - Αριθμητική μετρική του κατά πόσο ανόμοια είναι δυο αντικείμενα
  - Η ελάχιστη ανομοιότητα είναι συνήθως 0 – το πάνω όριο διαφέρει.
  - Είναι υψηλότερη όταν τα αντικείμενα «μοιάζουν» περισσότερο
- **Εγγύτητα (Proximity)**
  - Αναφέρεται σε ομοιότητα ή ανομοιότητα

Ομοιότητα/ανομοιότητα για απλά χαρακτηριστικά

p και *q* είναι οι τιμές του χαρακτηριστικού για 2 αντικείμενα

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Ομοιότητα ανάμεσα σε Δυαδικά διανύσματα

- Στην περίπτωση που *p* και *q* έχουν μόνο δυαδικά χαρακτηριστικά:
- Υπολογίστε τις ομοιότητες με βάση τις παρακάτω ποσότητες

  M_{01} = ο αριθμός των χαρακτηριστικών για τα οποία p=0, q=1

  M_{10} = ο αριθμός των χαρακτηριστικών για τα οποία p=1, q=0

  M_{00} = ο αριθμός των χαρακτηριστικών για τα οποία p=0, q=0

  M_{11} = ο αριθμός των χαρακτηριστικών για τα οποία p=1, q=1
- Συντελεστής Simple Matching

  SMC = number of matches / number of attributes

  $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$
- Συντελεστής Jaccard

  J = number of 11 matches / number of not-both-zero attributes values

  $= (M_{11}) / (M_{01} + M_{10} + M_{11})$

SMC versus Jaccard

p = 1 0 0 0 0 0 0 0 0 0 q = 0 0 0 0 0 0 1 0 0 1

$M_{01} = 2$, $M_{10} = 1$, $M_{00} = 7$, $M_{11} = 0$

$SMC = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$

$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$

Ομοιότητα συνημίτονου

- Εάν d_1 και d_2 δυο διανύσματα κειμένου, τότε:

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$$

όπου • το εσωτερικό γινόμενο και $||d||$ το μέτρο του d.

- Παράδειγμα:

d_1 = 3 2 0 5 0 0 0 2 0 0

d_2 = 1 0 0 0 0 0 0 1 0 2

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$$\cos(d_1, d_2) = .3150$$

Εκτεταμένος συντελεστής Jaccard (Tanimoto)

- Παραλλαγή του συντελεστή Jaccard για συνεχόμενα χαρακτηριστικά
- Εκφυλλίζεται σε Jaccard για δυαδικά χαρακτηριστικά

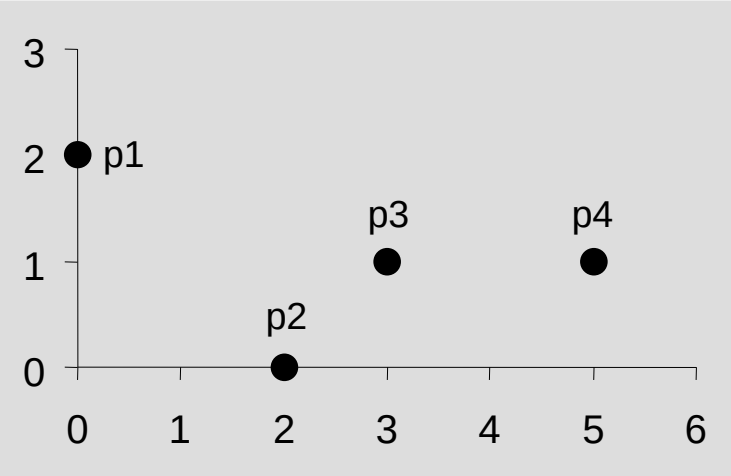
$$T(d_1, d_2) = (d_1 \bullet d_2) / (||d_1||^2 + ||d_2||^2 - d_1 \bullet d_2)$$

Ευκλείδεια απόσταση

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

- Η κανονικοποίηση είναι απαραίτητη, αν διαφέρουν οι κλίμακες.

Ευκλείδεια απόσταση



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Απόσταση Minkowski

- Γενίκευση της Ευκλείδειας απόστασης

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

όπου r παράμετρος, n ο αριθμός των χαρακτηριστικών, και p_k και q_k είναι, αντίστοιχα, οι τιμές των k^{th} χαρακτηριστικών των p και q.

Απόσταση Minkowski: με βάση το r

- r = 1. Manhattan, taxicab, L_1 norm.
  - Χαρακτηριστικό παράδειγμα είναι η απόσταση Hamming (ο αριθμός των bits που είναι διαφορετικά ανάμεσα σε δυο δυαδικά διανύσματα)
- r = 2. Ευκλείδεια απόσταση
- $r \to \infty$. "supremum" (L_{max} norm, L_∞ norm) απόσταση
  - Η μέγιστη διαφορά ανάμεσα σε οποιοδήποτε τμήμα των διανυσμάτων
- Σημείωση: μη συγχέεται το r με το n (οι αποστάσεις αυτές ορίζονται για όλες τις διαστάσεις).

Απόσταση Minkowski

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

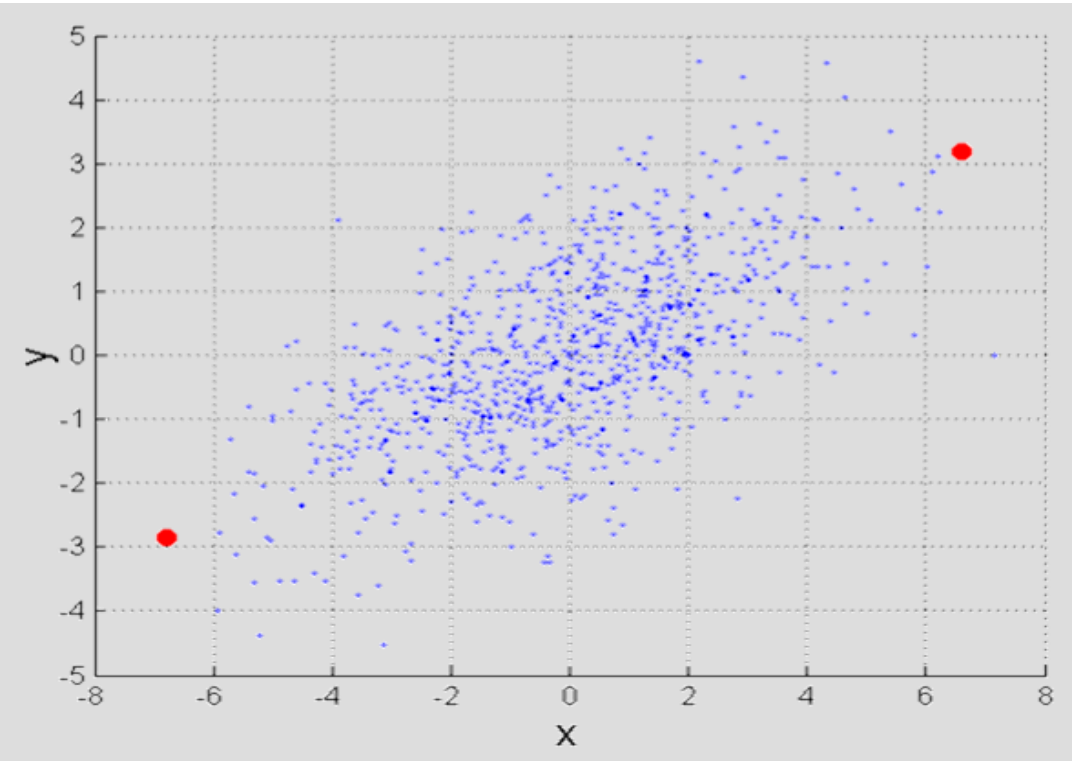
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Απόσταση Mahalanobis

$$mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$

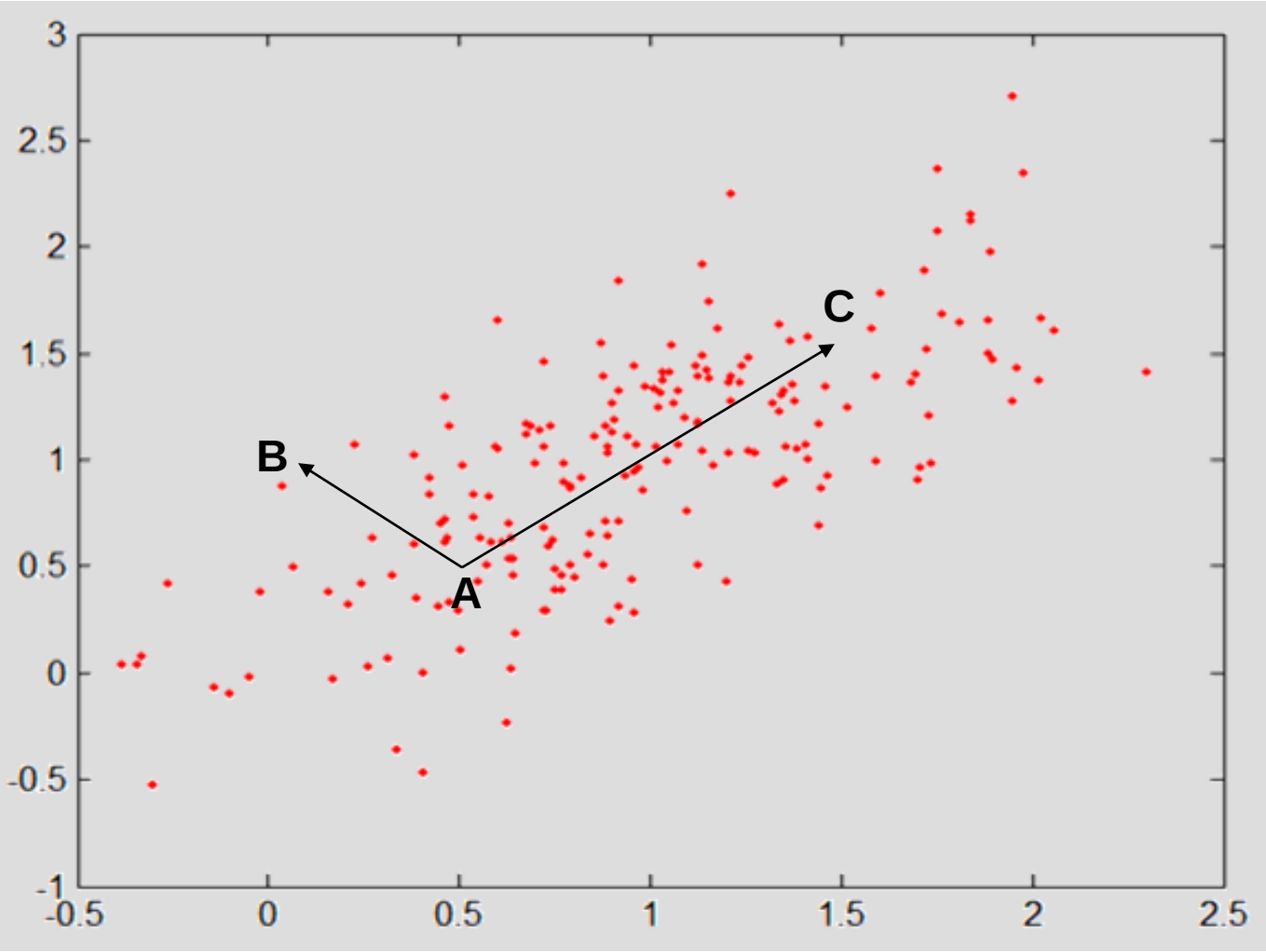


Το Σ είναι ο πίνακας συμμεταβλητότητας των δεδομένων εισόδου X

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

Για τα κόκκινα σημεία, η Ευκλείδεια απόσταση είναι 14.7, και η Mahalanobis 6.

Απόσταση Mahalanobis



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Συσχέτιση (Correlation)

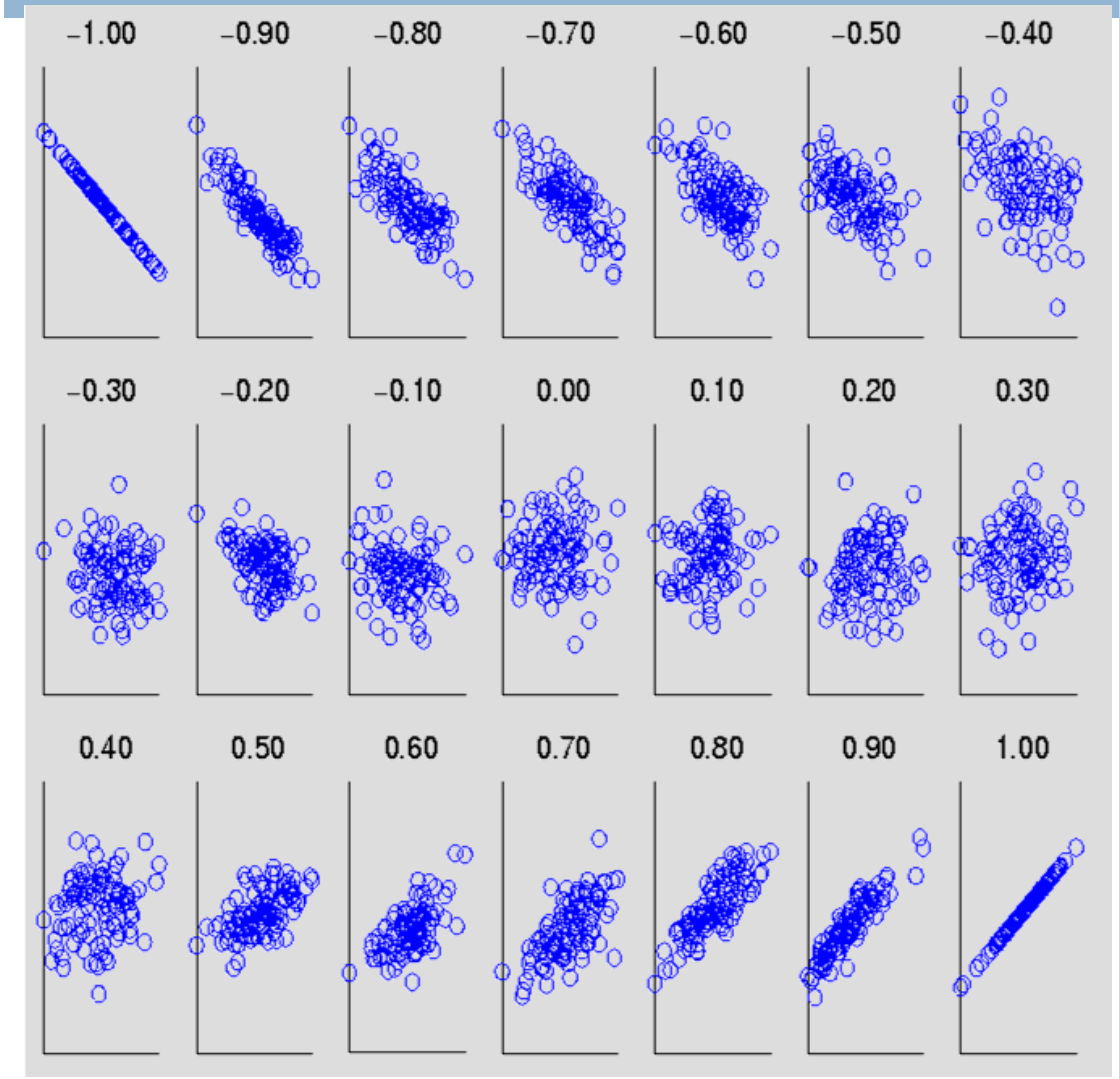
- Μετρά τη γραμμική σχέση ανάμεσα σε αντικείμενα
- Για τον υπολογισμό της συσχέτισης, κανονικοποιούμε τα p και q, και παίρνουμε το εσωτερικό γινόμενο

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Οπτική αποτίμηση της συσχέτισης



Δείχνουν την ομοιότητα από −1 σε 1.