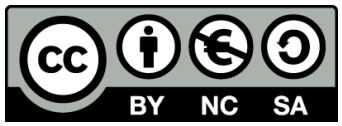
ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

# Αποθήκες Δεδομένων και Εξόρυξη Δεδομένων

**Ενότητα 2**: Επεξεργασία Δεδομένων

Αναστάσιος Γούναρης, Επίκουρος Καθηγητής
Τμήμα Πληροφορικής ΑΠΘ




Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

# Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# Επεξεργασία Δεδομένων

Επεξεργασία δεδομένων, μείωση μεγέθους δεδομένων, μέτρα απόστασης δεδομένων.



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

# Περιεχόμενα ενότητας

1. Μορφή και ιδιότητες δεδομένων.
2. Περίληψη δεδομένων.
3. Καθαρισμός δεδομένων.
4. Μετασχηματισμός δεδομένων.
5. Κβαντοποίηση δεδομένων.
6. Δειγματοληψία.
7. Μείωση αριθμού διαστάσεων.
8. Ομοιότητα – Απόσταση δεδομένων.

# Σκοποί ενότητας

- Περιγραφή της μορφής και των ιδιοτήτων των δεδομένων.
- Περιγραφή πράξεων επεξεργασίας δεδομένων.
- Γνωριμία με τα μέτρα απόστασης και ομοιότητας.

# Σύνολο δεδομένων

**Ιδιότητες/Χαρακτηριστικά**

| Κωδικός | Τίτλος | Σελίδες | Επίπεδο | Τιμή | Διαθεσιμό-τητα |
|---|---|---|---|---|---|
| 547745 | Προχωρημένη C++ | 427 | Υψηλό | 64.50 | ΟΧΙ |
| 547746 | Εισαγωγή στη Java | 256 | Χαμηλό | 57.20 | ΝΑΙ |
| 547747 | Εφαρμογές στην SQL | 357 | Μεσαίο | 45.30 | ΝΑΙ |
| **Διακριτή Αριθμητική** | **Διακριτή Συμβολική** | **Διακριτή Αριθμητική** | **Διακριτή Συμβολική** | **Συνεχής Αριθμητική** | **Διακριτή Συμβολική Δυαδική** |

**Διαστασιμότητα (Dimensionality): #Ιδιότητες = 6**

# Ιδιότητες δεδομένων

- Μια συμπληρωματική θεώρηση:
  - Ονομαστικές (nominal) **= ≠**
  - Διατεταγμένες (ordinal) **= ≠ < >**
  - Διαστήματος (interval) **= ≠ < > + -**
  - Αναλογίας (ratio) **= ≠ < > + - * /**

# Τύποι Δεδομένων

| A/A | Επιστροφή | Οικογ. κατάσταση | Εισόδημα | Απάτη |
|---|---|---|---|---|
| 1 | Ναι | Άγαμος | 125K | Όχι |
| 2 | Όχι | Διαζευγμ. | 95K | Ναι |
| 3 | Όχι | Άγαμος | 70K | Όχι |
| 4 | Όχι | Έγγαμος | 60K | Όχι |
| 5 | Όχι | Έγγαμος | 100K | Όχι |
| 6 | Ναι | Έγγαμος | 120K | Όχι |

| A/A | Αντικείμενα |
|---|---|
| 1 | Ψωμί, Αλεύρι, Γάλα |
| 2 | Μπύρα, Πάνες, Γάλα, Ψωμί |
| 3 | Μπύρα, Ψωμί |
| 4 | Μπύρα, Ψωμί, Πάνες, Γάλα |
| 5 | Αλεύρι, Πάνες, Γάλα |

| | Ομάδα | Αγώνας | Ήττα | Διαιτητής | Νίκη |
|---|---|---|---|---|---|
| Εγγρ1 | 2 | 0 | 0 | 1 | 1 |
| Εγγρ2 | 0 | 5 | 4 | 1 | 0 |
| Εγγρ3 | 2 | 1 | 0 | 2 | 2 |

# Περίληψη δεδομένων

- Μέση τιμή (mean)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Ενδιάμεση (median)

- Mode *η πιο συχνή τιμή*

- Διασπορά

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

# Ζητήματα με τα δεδομένα

- Ποιότητα.
  - Απαλοιφή θορύβου.
  - Εντοπισμός ανωμαλιών.
  - Ελλιπείς τιμές.
  - Μετασχηματισμός.
- Κβάντωση.
- Μείωση μεγέθους.
  - Γραμμών: Δειγματοληψία.
  - Στηλών: Ιδιοδιανύσματα, Επιλογή χαρακτηριστικών.
- Μέτρα απόστασης.

# Προβλήματα ποιότητας δεδομένων

- Αίτια:
  - Κακός σχεδιασμός:
    - Διαδικασίες συλλογής-δημιουργίας δεδομένων που δεν είχαν σχεδιασθεί για να διευκολύνουν το έργο της εξόρυξής τους.
  - Λάθη.
    - Ανθρώπινα, προβλήματα με συσκευές μέτρησης.
- Παράγοντες:
  - Θόρυβος (από ανθρώπινα λάθη ή προβλήματα συσκευών).
  - «Ανώμαλες» (outliers) και ασυνεπείς τιμές (από λάθη κατά την εισαγωγή και κακό σχεδιασμό).
  - Ελλιπείς τιμές (από κακό σχεδιασμό ή από λάθη κατά την εισαγωγή).

**garbage in, garbage out**

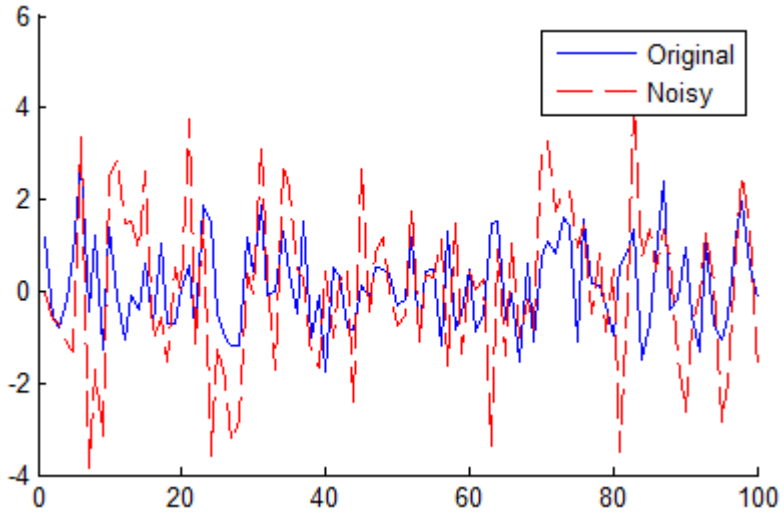
# Λύσεις

- Καθαρισμός.
- Μετασχηματισμός.

- Δεν υπάρχει πανάκεια.
- Δεν υπάρχουν γενικές λύσεις.
- 70%-80% του χρόνου της διαδικασίας ανακάλυψης γνώσης – εξόρυξης.

# Καθαρισμός θορύβου

- Θόρυβος: η τυχαία αλλοίωση τιμών.
- ή η παρείσφρηση τυχαίων αντικειμένων.

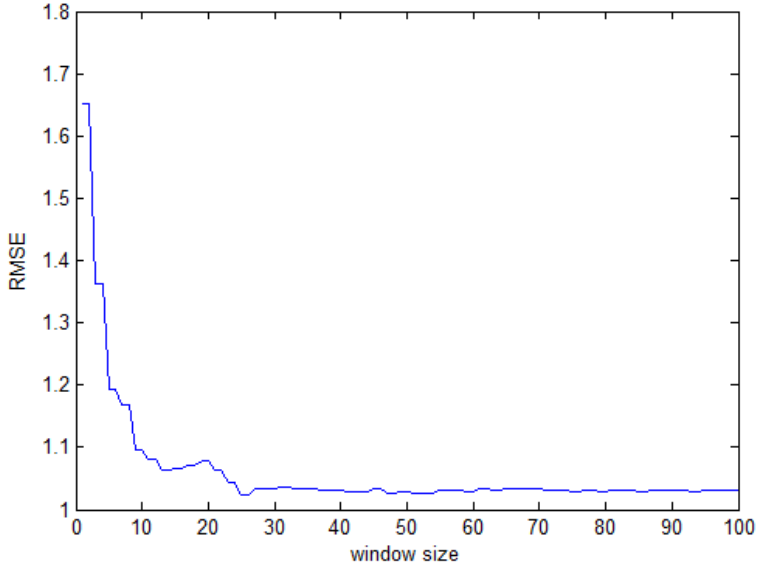


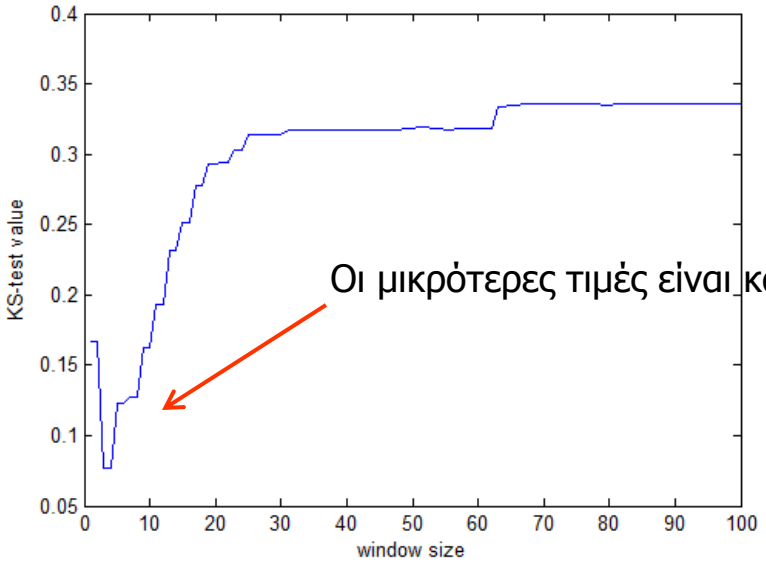

**Η αλλοίωση μετράται με τη ρίζα του μέσου τετραγωνικού λάθους.**

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}}$$

# Καθαρισμός κυλιόμενου μέσου

- Παράθυρο με μήκος w.
- Κάθε τιμή στη θορυβώδη σειρά τη θέτουμε ίση με το μέσο όρο των w/2 τιμών αριστερά και των w/2 τιμών δεξιά της.
- Αντίκτυπος της τιμής του w?



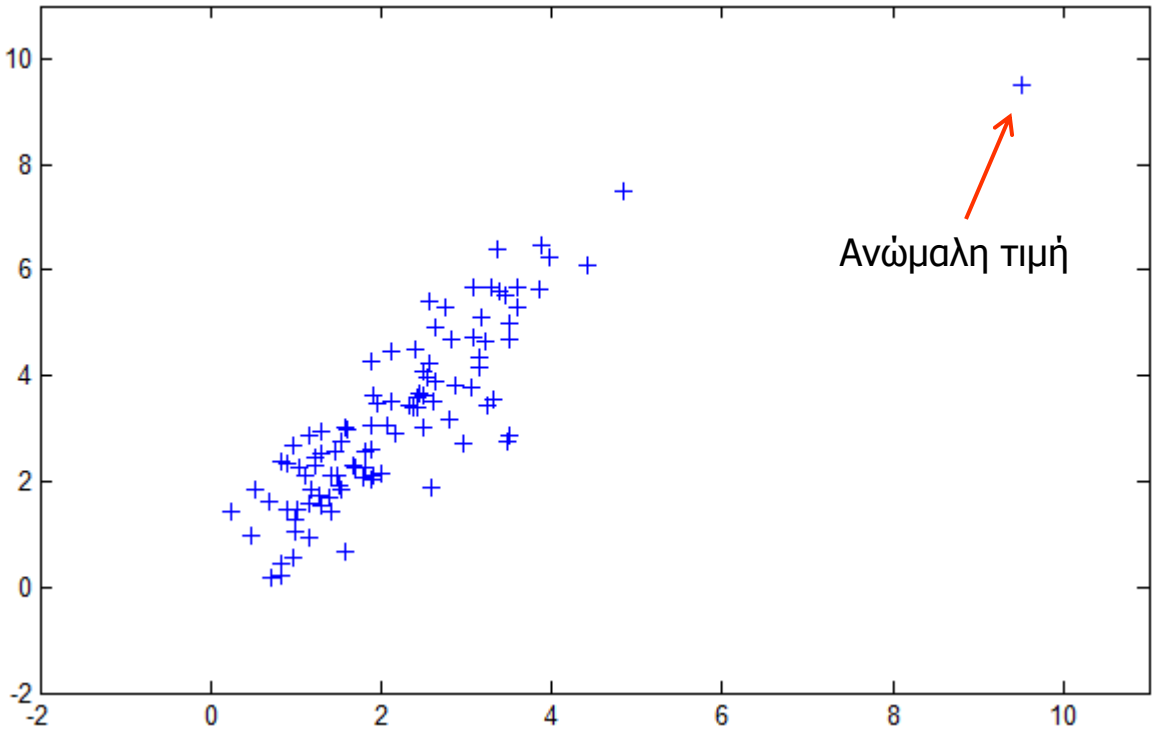




  - Δηλ. το μέτρο RMSE δεν είναι ενδεικτικό του πόσο τα αποθορυβοποιημένα δεδομένα ταιριάζουν με τα αρχικά.

# Ανώμαλες και Ασυνεπείς τιμές

- Ανώμαλες τιμές: δεν ακολουθούν την κατανομή των τιμών των υπολοίπων αντικειμένων.
- Τα αντικείμενα με ανώμαλες τιμές καλούνται outliers (ακρότατα, ακραία σημεία).
- Πιο συνηθισμένη αιτία: λάθη κατά την εισαγωγή τιμών.
- Σε κάποιες περιπτώσεις, όμως, μπορεί να μην υπάρχει λάθος, αλλά η διαφοροποίηση να δηλώνει κάτι ενδιαφέρον (π.χ., ανίχνευση οικονομικής απάτης).
- Οι ασυνεπείς (inconsistent) τιμές είναι οι τιμές που δεν έχουν νόημα, π.χ. μελλοντική ημερομηνία γέννησης.

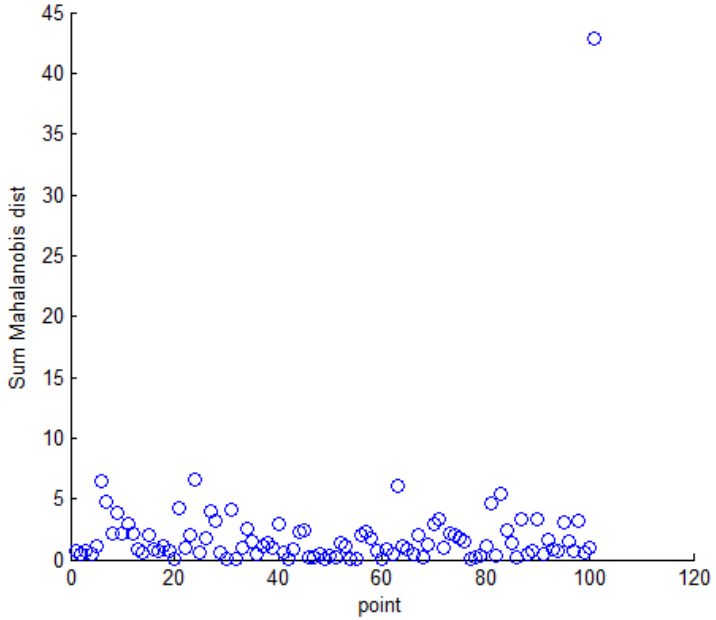
# Εντοπισμός ανώμαλων τιμών




100 σημεία με 2d κανονική κατανομή + 1 ανώμαλη τιμή

# Εντοπισμός βάσει αποστάσεων

- Υπάρχουν πολλές μέθοδοι καθορισμού του πότε ένα σημείο είναι εξωκείμενο. Δίνουμε ένα μόνο παράδειγμα:
- Μετράμε το άθροισμα της απόστασης κάθε σημείου από όλα τα υπόλοιπα.
- Ως μέτρο απόστασης παίρνουμε την απόσταση Mahalanobis.



# Ελλιπείς τιμές

- Αναφέρονται σε αντικείμενα με χαρακτηριστικά που η τιμή τους δεν είναι γνωστή.

- Αιτίες:
  - κάποια δεδομένα θεωρούνται προσωπικά, οπότε δεν μπορούμε να τα έχουμε στη διάθεσή μας.
  - προαιρετικά πεδία φόρμας, κ.λπ.

# Λύσεις

- Διαγραφή αντικειμένων με ελλιπείς τιμές.
  - Αν ο αριθμός των αντικειμένων με ελλιπείς τιμές είναι μεγάλος, τότε το σύνολο δεδομένων που προκύπτει μπορεί να μην είναι αντιπροσωπευτικό.
- Συμπλήρωση ελλιπών τιμών μέσω εκτίμησης.
  - συμπλήρωση με το μέσο όρο της τιμής, μέσω παλινδρόμησης, κ.α.
  - Λαμβάνοντας υπόψη την κλάση του αντικειμένου ή όχι.
- Διατήρηση ελλιπών τιμών.
  - Δεν κάνουμε κάποια ενέργεια αντιμετώπισης των ελλιπών τιμών.

Η ποιότητα του αποτελέσματος ενός αλγορίθμου επηρεάζεται αρνητικά από την ύπαρξη ελλιπών τιμών.

# Μετασχηματισμός δεδομένων

- Μετασχηματισμό δεδομένων ονομάζουμε την εφαρμογή κάποιας συνάρτησης επί των τιμών μίας ιδιότητας (μετασχηματίζουμε ξεχωριστά κάθε ιδιότητα).

- Συνήθως μετασχηματίζουμε τις τιμές μίας αριθμητικής ιδιότητας που έχει πολύ μεγαλύτερες ή μικρότερες τιμές

- Τυποποίηση (standardization, z-score)

$$X_{new} = \frac{X - \overline{X}}{\sigma}$$

- Μετασχηματισμός Min-max

$$X_{new} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} (X'_{\max} - X'_{\min}) + X'_{\min}$$

# Κβάντωση

- Κβάντωση: ο μετασχηματισμός μίας συνεχούς ιδιότητας σε διακριτή.
  - Διακριτοποίηση (discretization).
- Δυαδική μετατροπή: Η ειδική περίπτωση του μετασχηματισμού συνεχούς ιδιότητας σε δυαδική (binarization).
- Χρήσεις, λόγοι;
  - Απαιτήσεις Αλγορίθμων.
  - Απόδοση/Αποτελεσματικότητα μοντέλων.

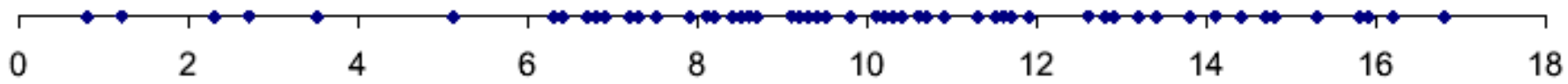
# Αλγόριθμοι Κβάντωσης

- Υπολόγισε το μέγεθος, **n**, του διακριτού πεδίου.
  - Π.χ., χρησιμοποιώντας τον τύπος Sturges:

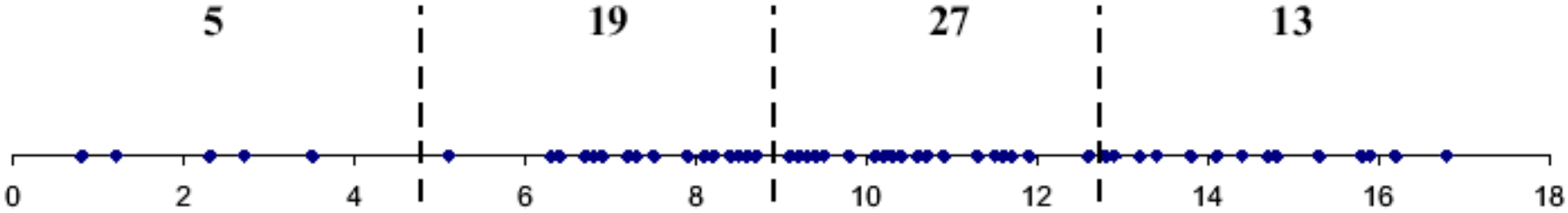
    $n = log_2 (M+1)$, *M: πλήθος διαφορετικών τιμών*
- Ταξινόμησε τις τιμές της συνεχούς μεταβλητής.
- Διαχώρισε τις τιμές αυτές σε n διακριτά διαστήματα, με καθορισμό **n-1** σημείων διαχωρισμού.
- Αντιστοίχησε ένα-προς-ένα κάθε διάστημα με μία εκ των n τιμών του διακριτού πεδίου.

# Ισο-ευρής διαχωρισμός

- Διαστήματα ίσου εύρους.




- Πρόβλημα στις ανομοιόμορφες (skewed) κατανομές



# Εύρος Διαστημάτων

- Έχοντας ορίσει το *n*, το εύρος W είναι:

$$W = \frac{X_{\max} - X_{\min}}{n}$$

- Εναλλακτικός ορισμός του n:

$$n = \frac{X_{\max} - X_{\min}}{W}$$

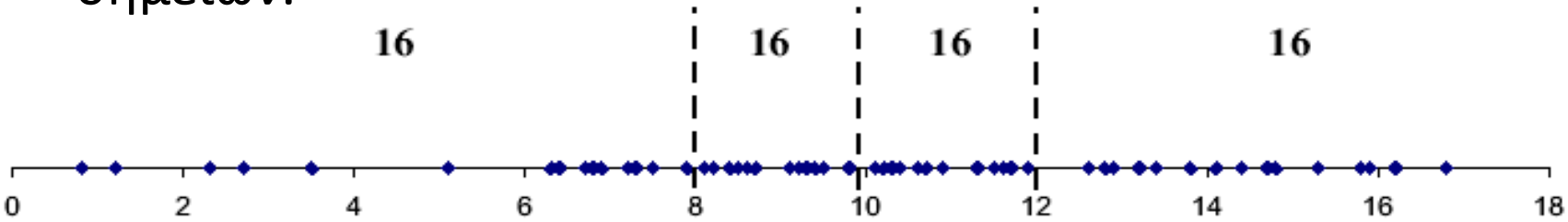
με

$$W = \frac{3.5(Q_3 - Q_1)}{\sqrt[3]{M}}, W = \frac{3.5\sigma}{\sqrt[3]{M}}$$

# Ισο-υψής διαχωρισμός

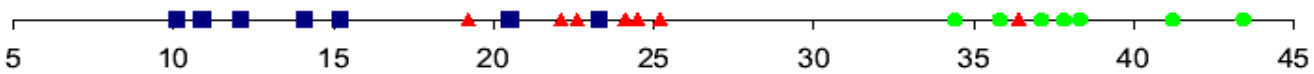
- Μεταβλητό εύρος των διαστημάτων.


- Κάθε διάστημα αντιπροσωπεύεται από ισοδύναμο αριθμό σημείων.

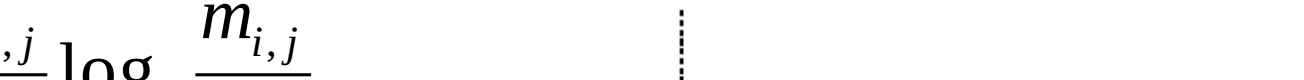


# Επιβλεπόμενη κβάντωση

Λαμβάνεται υπόψη η πληροφορία της κλάσης όπου ανήκει κάθε αντικείμενο (*k κλάσεις, n διαστήματα).*




$$e_i = -\sum_{j=1}^{k} \frac{m_{i,j}}{m_i} \log_2 \frac{m_{i,j}}{m_i}$$

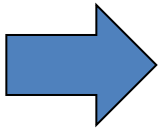



$$e = \sum_{i=1}^{n} \frac{m_i}{M} e_i$$

# Μείωση μεγέθους

Αρχικά δεδομένα

| Σ1 | Σ2 | Σ3 | Σ4 | Σ5 |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |



| Σ΄1 | Σ΄2 |
|---|---|
| | |
| | |
| | |
| | |
| | |

ιδιοδιανύσματα

δειγματοληψία

| Σ1 | Σ2 | Σ3 | Σ4 | Σ5 |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |

| Σ1 | Σ4 |
|---|---|
| | |
| | |
| | |
| | |
| | |

επιλογή χαρακτηριστικών

# Δειγματοληψία

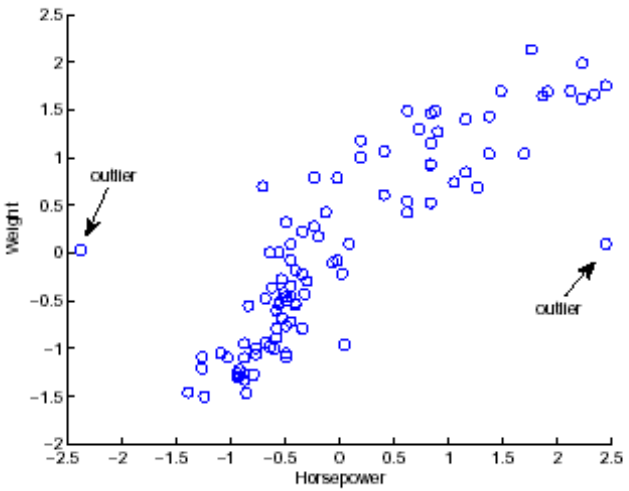
- Χρησιμοποιείται για επιλογή δεδομένων τόσο στην προεπεξεργασία όσο στην τελική ανάλυση.
- Στη στατιστική, προτιμάται όταν το κόστος **απόκτησης** ολόκληρου του συνόλου δεδομένων είναι πολύ υψηλό ή χρονοβόρο.
- Στην ΕΔ, προτιμάται όταν το κόστος **επεξεργασίας** ολόκληρου του συνόλου δεδομένων είναι πολύ υψηλό ή χρονοβόρο.
  - Κλιμακούμενοι αλγόριθμοι έχουν μικρότερη ανάγκη από δειγματοληψία.
- Η χρησιμοποίηση δείγματος δεν οδηγεί σε διαφορετικά αποτελέσματα σε σχέση με τη χρησιμοποίηση όλων των δεδομένων.
  - ΑΡΚΕΙ το δείγμα να είναι αντιπροσωπευτικό.
  - Δηλ. να έχει τις ίδιες ιδιότητες (από αυτές που μας ενδιαφέρουν) με το σύνολο.

# Τύποι Δειγματοληψίας

- Απλή τυχαία δειγματοληψία.
  - Με επανατοποθέτηση.
  - Χωρίς επανατοποθέτηση.
- Στρωματοποιημένη δειγματοληψία.

# Απλή τυχαία δειγματοληψία

- Μέγεθος δείγματος: Διατήρηση μέσης τιμής, ακρίβεια κατηγοριοποίησης.
- Παράπλευρο όφελος: απαλοιφή outliers.
- Διάστημα εμπιστοσύνης (με χρήση κεντρικού οριακού θεωρήματος).
  - Για μέση τιμή δείγματος να απέχει από τη μέση τιμή του αρχικού πληθυσμού $\mu$ το πολύ $\delta$, υπολογίζω μέγεθος δείγματος.



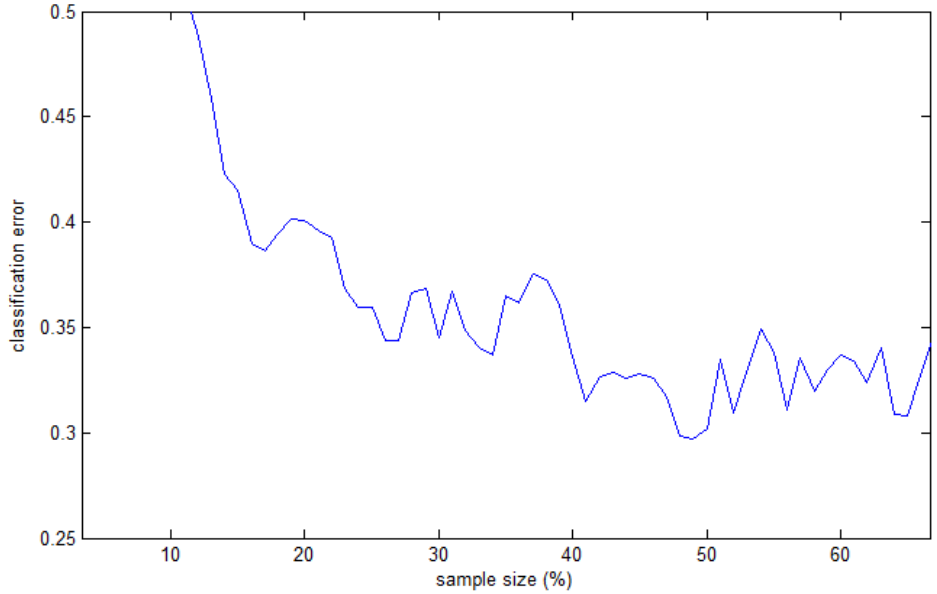

$$P(|\overline{X} - \mu| \leq \delta) = \alpha$$

$$n \approx \left\lceil z_a \frac{\sigma^2}{\delta^2} \right\rceil$$

$Z_a$: κανονικοποιημένη τιμή της κανονικής κατανομής που αντιστοιχεί στην τιμή $a$.

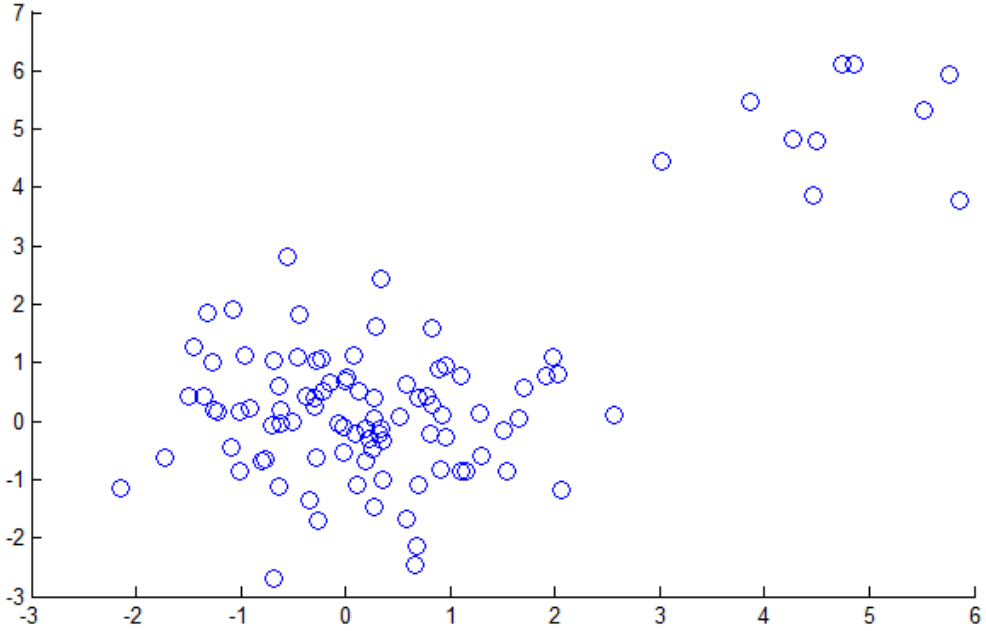
# Μέγεθος Δείγματος με Βάση την Ακρίβεια

- Παράδειγμα με το IRIS σύνολο δεδομένων.
- Τα μεγάλα δείγματα δεν σημαίνουν καλύτερα αποτελέσματα
- …λόγω του overfitting.
- Αλλά μικρά δείγματα μπορούν να χάσουν πληροφορία συνδιασποράς!



# Στρωματοποιημένη δειγματοληψία

- Διαφορετικά μεγέθη μεταξύ των ομάδων.
  - η τυχαία δειγματοληψία παράγει δείγμα στο οποίο δεν αντιπροσωπεύονται ικανοποιητικά όλες οι ομάδες.
- Στρωματοποιημένη δειγματοληψία.
  - εφαρμόζουμε τυχαία δειγματοληψία σε κάθε ομάδα ξεχωριστά.

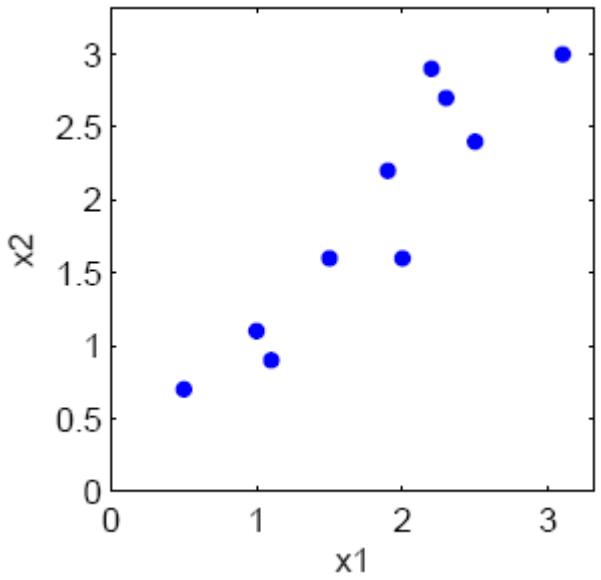


# Μείωση στηλών

- Η «κατάρα της διαστασιμότητας».
- Με μείωση αριθμού διαστάσεων:
  - απαλοιφή εξαρτήσεων – άσχετων χαρακτηριστικών.
  - δυνατότητα προβολής (για <= 3 διαστάσεις).
  - Καλύτερα, πιο κατανοητά και γρηγορότερα αποτελέσματα.
- 2 βασικές μέθοδοι.
  1. Δημιουργία νέων χαρακτηριστικών.
     - Ανάλυση Πρωταρχικών Συνιστωσών – Principal Component Analysis (PCA).
  2. Επιλογή χαρακτηριστικών.
     - Φιλτράρισμα.
     - Μαύρο κουτί.

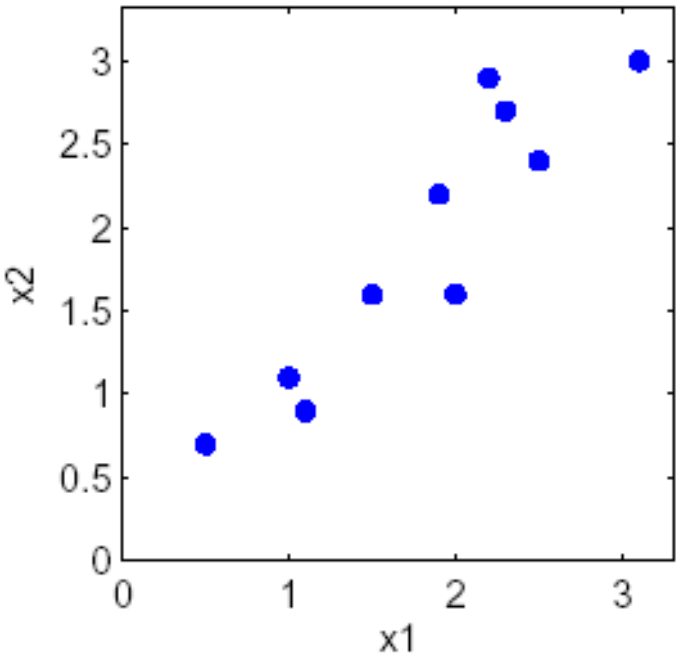
# Παράδειγμα PCA

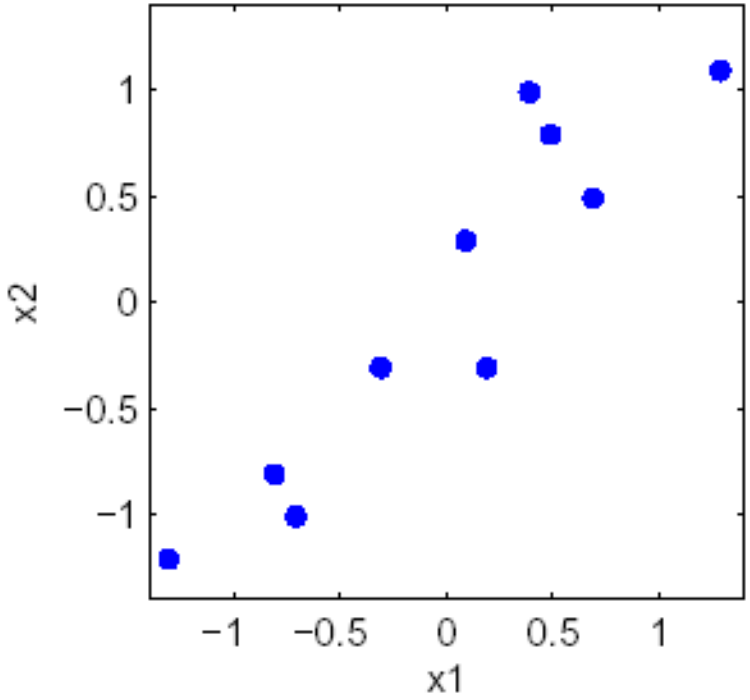
- Το ζητούμενο είναι να πάμε από 2 σε 1 διάσταση.
- Ακολουθούμε μία διαδικασία 5 βημάτων.




| X1 | X2 |
|---|---|
| 2.5 | 2.4 |
| 0.5 | 0.7 |
| 2.2 | 2.9 |
| 1.9 | 2.2 |
| 3.1 | 3.0 |
| 2.3 | 2.7 |
| 2 | 1.6 |
| 1 | 1.1 |
| 1.5 | 1.6 |
| 1.1 | 0.9 |

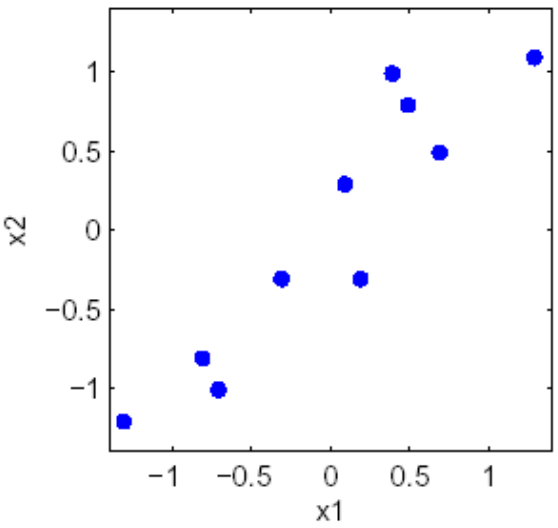
# Βήμα 1: αφαίρεση μέσης τιμής



# Βήμα 2: πίνακας συνδιασποράς

$$C_{i,j} = E[(X_i - \overline{X_i})(X_j - \overline{X_j})]$$




διασπορά

$$C = \begin{pmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{pmatrix}$$

Πόσο μεταβάλλονται μαζί τα X1,X2

# Βήμα 3: ιδιοδιανύσματα - ιδιοτιμές

$$C\mathbf{u} = \lambda\mathbf{u}$$

2 λύσεις στην εξίσωση

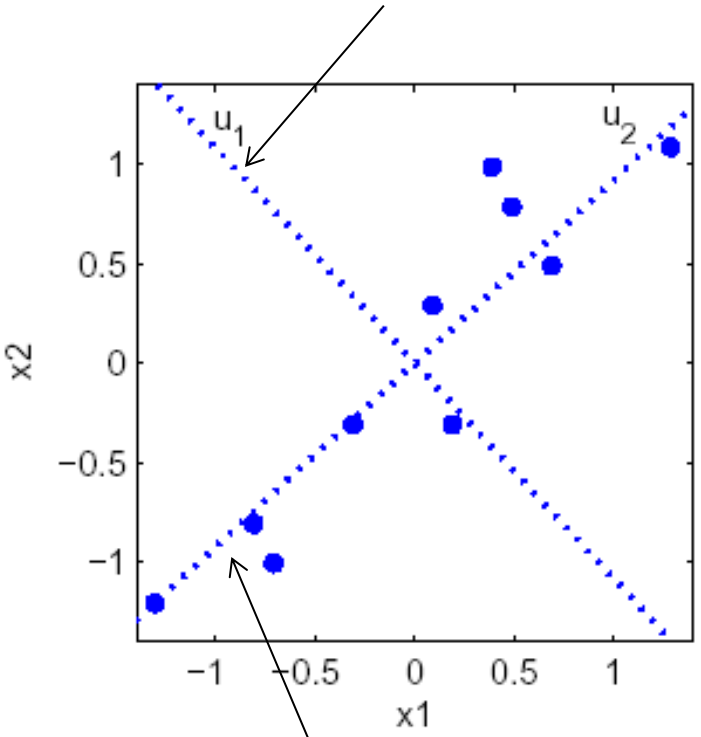
u1= $[-0.7352, 0.6779]^T$

λ1= 0.0491

u2= $[0.6779, 0.7352]^T$

λ2= 1.2840

μεγαλύτερη εναπομένουσα διασπορά.



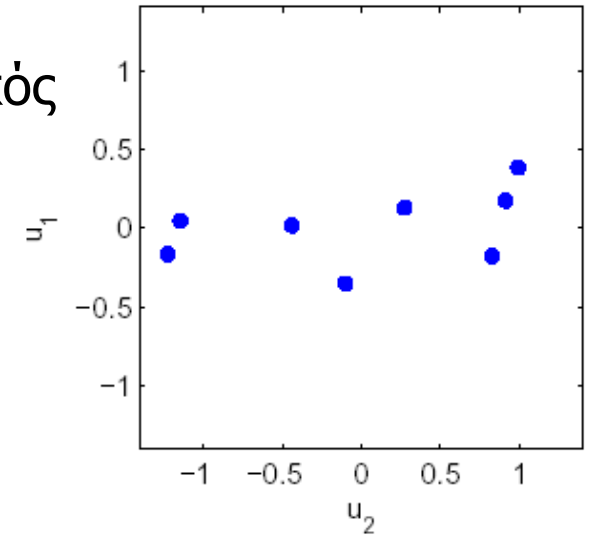

Κατεύθυνση με τη μεγαλύτερη διασπορά.

# Βήμα 4: επιλογή νέων ιδιοτήτων

1. Διάταξη ιδιοδιανυσμάτων βάσει ιδιοτιμών.
2. Τα αντικείμενα περιγράφονται ως γραμμικός συνδυασμός των ιδιοδιανυσμάτων, που ονομάζονται **κύριες συνιστώσες.**

$$\mathbf{U} = [\mathbf{u}_2, \mathbf{u}_1]$$

$$(x_1, x_2) \rightarrow (x_1, x_2)\mathbf{U}$$




$$C = \begin{pmatrix} 1.284 & 0 \\ 0 & 0.0491 \end{pmatrix}$$

Οι νέες διαστάσεις είναι ασυσχέτιστες.

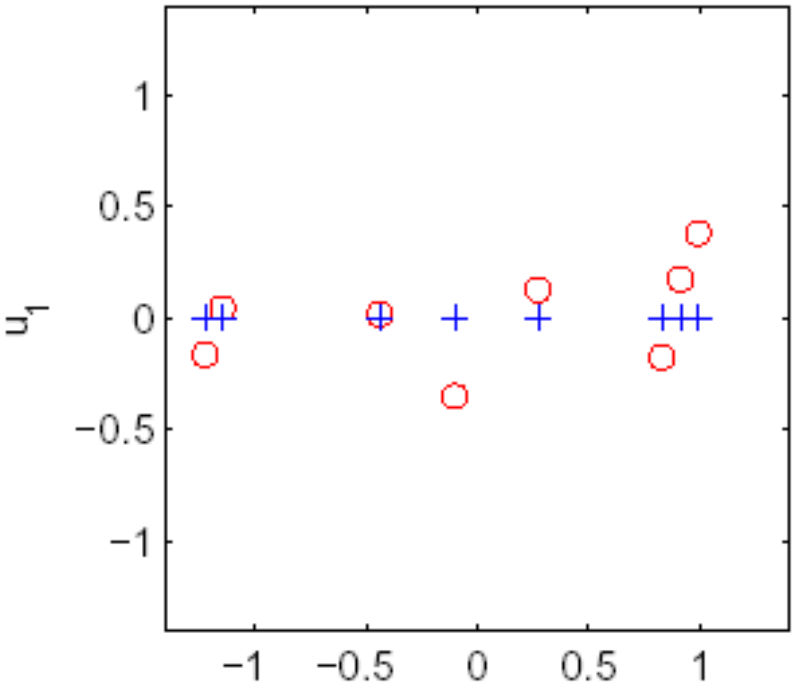
# Βήμα 5: μείωση διαστάσεων

Επιλέγουμε τα πρώτα k ιδιοδιανύσματα

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k, ..., \mathbf{u}_n]$$

$$\mathbf{U}' = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k]$$

$$(x_1, ..., x_n) \rightarrow (x_1, ..., x_n)\mathbf{U}'$$

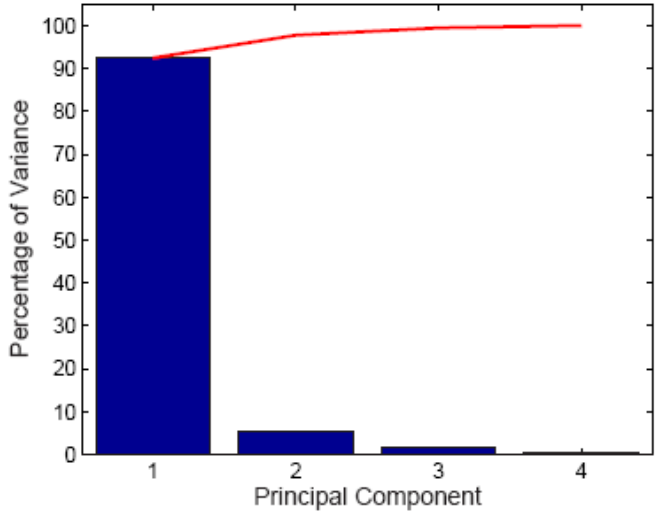



k = 1 (+)

**k = 2 (o)**

# Παράδειγμα: Iris

- 150 λουλούδια.
- 4 διαστάσεις (μήκος/πλάτος σέπαλου/πετάλου).
- 3 κλάσεις (Setosa, Versicolor, Virginica).
- Πως επιλέγουμε σε πόσες διαστάσεις θα γίνει η προβολή;



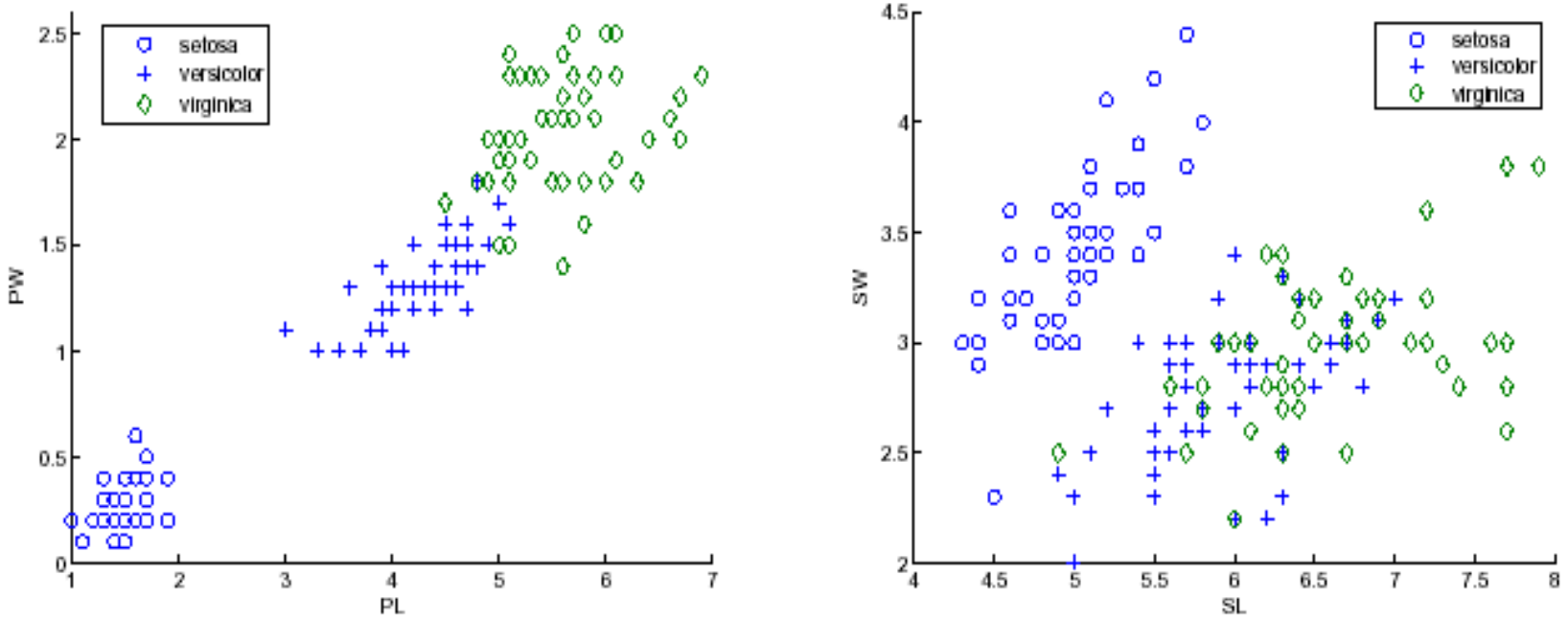
# Επιλογή χαρακτηριστικών

- Τα ιδιοδιανύσματα δεν αντιστοιχούν στις αρχικές διαστάσεις.
- Αν πρέπει να επιλέξουμε μεταξύ των αρχικών διαστάσεων;
  - Φιλτράρισμα: Επιλέγουμε ανεξάρτητα του αλγορίθμου που θα εφαρμοσθεί στη συνέχεια.
  - Μαύρο κουτί: Χρησιμοποιούμε τον αλγόριθμο σαν μαύρο κουτί, για την εύρεση του καταλληλότερου υποσυνόλου διαστάσεων.

# Φιλτράρισμα

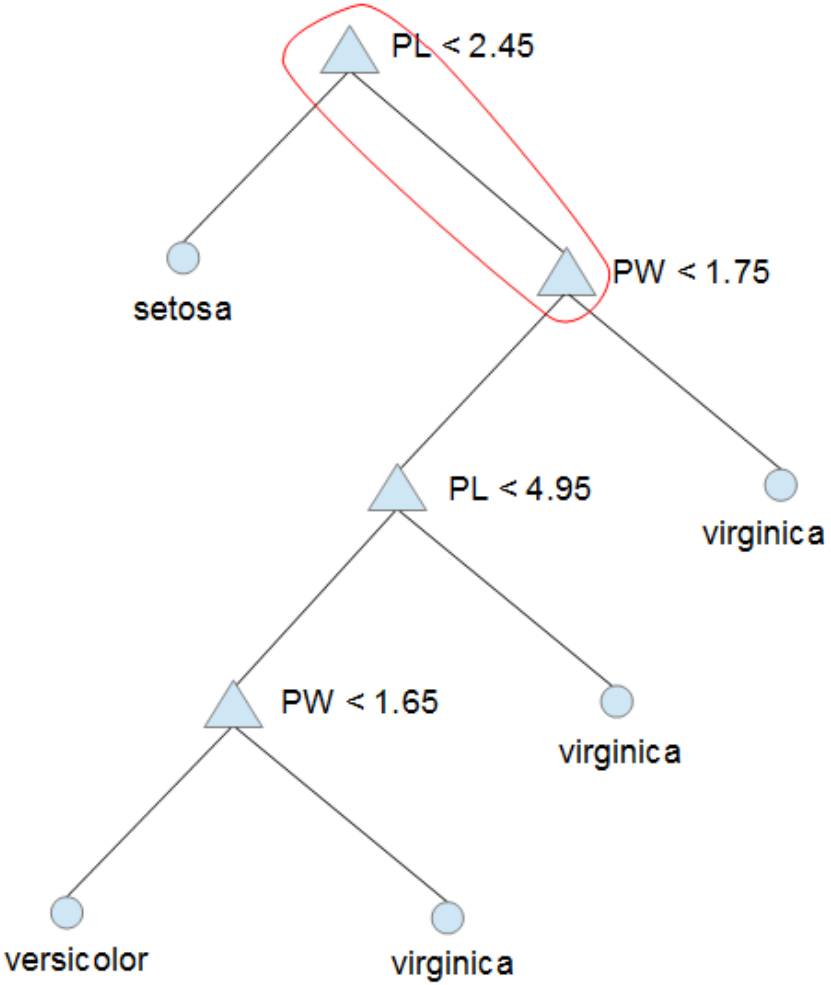
- 3 πρώτα βήματα όπως και στο PCA.
- Στο παράδειγμα IRIS:
  - Επιλέγουμε τα δύο πρώτα ιδιοδιανύσματα.
    - u1 = $(0.3614, -0.0845, 0.8567, 0.3583)^T$
    - u2 = $(-0.6566, -0.7302, 0.1734, 0.0755)^T$
  - Παίρνουμε το άθροισμα u1 + u2
    - = $(-0.2952, -0.8147, 1.0300, 0.4338)^T$
  - Η τρίτη και η τέταρτη θέση έχουν τη μεγαλύτερη τιμή.
  - Ένδειξη ότι η τρίτη και η τέταρτη ιδιότητα (μήκος -πλάτος πετάλου **PL, PW**) είναι οι καταλληλότερες για επιλογή.

# Παράδειγμα




Απεικόνιση του συνόλου δεδομένων iris

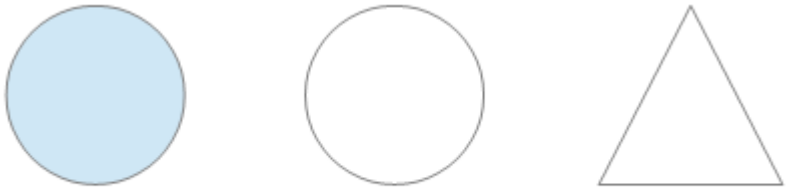
# Μαύρο κουτί


PL < 2.45
setosa
PW < 1.75
PL < 4.95
virginica
PW < 1.65
virginica
versicolor
virginica

# Μέτρα ομοιότητας/απόστασης

- Ιδιότητες μέτρων ομοιότητας.
  - Ανακλαστική $s(x, y) = 1 \Leftrightarrow x = y$
  - Συμμετρική $s(x, y) = s(y, x)$
- ΑΠΟΣΤΑΣΗ=*ΜΟΝΟΤΟΝΑ_ΦΘΗΝΟΥΣΑ*(ΟΜΟΙΟΤΗΤΑ).
- Ιδιότητες **μετρικού** απόστασης.
  - Θετικότητα $d(x, y) \geq 0$
  - Ανακλαστική $d(x, y) = 0 \Leftrightarrow x = y$
  - Συμμετρική $d(x, y) = d(y, x)$
  - Τριγωνική ανισότητα $d(x, y) \leq d(x, z) + d(z, y)$

# Μέτρα και Μετρικά: Τριγωνική ανισότητα




$$d(x,y) = \begin{cases} 1, \alpha\nu\ \delta\iota\alpha\phi\acute{\varepsilon}\rho\sigma\upsilon\nu\ \ \kappa\alpha\iota\ \sigma\varepsilon\ \chi\rho\acute{\omega}\mu\alpha\ \ \kappa\alpha\iota\ \sigma\varepsilon\ \sigma\chi\acute{\eta}\mu\alpha \\ \\ 0, \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

d(Α, Γ) = 1, d(Α, Β) = d(Β, Γ) = 0

d(Α, Γ) > d(Α, Β) + d(Β, Γ)

# Απόσταση Minkowski

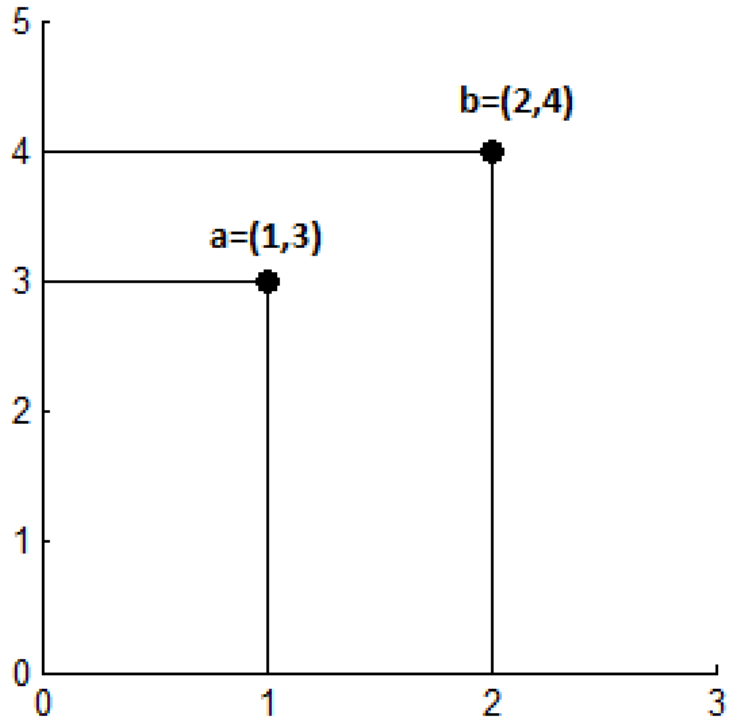
Γενίκευση ευκλείδειας απόστασης – όλα τα μέτρα είναι μετρικά.

$$d(x, y) = \sqrt[r]{\sum_{i=1}^{n} | x_i - y_i |^r}$$

r=1: Manhattan/block απόσταση.
r=2: Ευκλείδεια απόσταση.
r=άπειρο: υπέρτατη απόσταση.

$$d(x, y) = \max(| x_i - y_i |)$$

# Παράδειγμα




**Αποστάσεις μεταξύ a και b.**

Ευκλείδεια απόσταση:
$(1^2+1^2)^{1/2} = 1$

Απόσταση Manhattan:
1+1 = 2

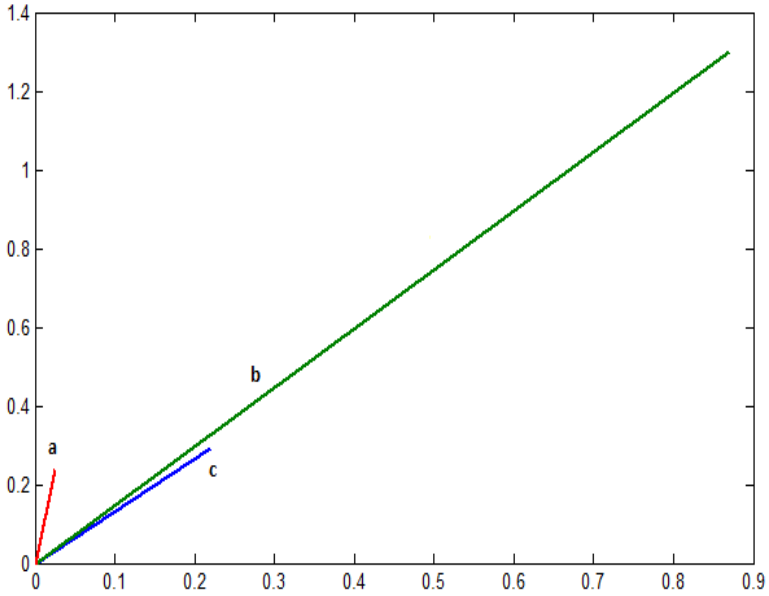
# Ομοιότητα cosine

$$s(x,y) = \frac{x \cdot y}{\| x \| \, \| y \|}$$

$$x \cdot y = \sum x_i y_i$$

$$\| x \| = \sqrt{\sum x_i^2}$$

| | x | y |
|---|---|---|
| a | 0.025 | 0.24 |
| b | 0.87 | 1.3 |
| c | 0.22 | 0.29 |




**Σύνοψη χαρακτηριστικών:**

- **Προτιμάται όταν το μέγεθος δεν παίζει ρόλο.**
- **Ανοχή-ανθεκτικότητα σε αραιά διανύσματα.**
- **Παραβιάζει όμως την ανακλαστική ιδιότητα!**
- **Αντίστοιχη είναι και η συσχέτιση Pearson:**

$$corr_{X,Y} = \frac{C_{X,Y}}{\sigma_X \sigma_Y}$$

# Απόσταση Mahalanobis

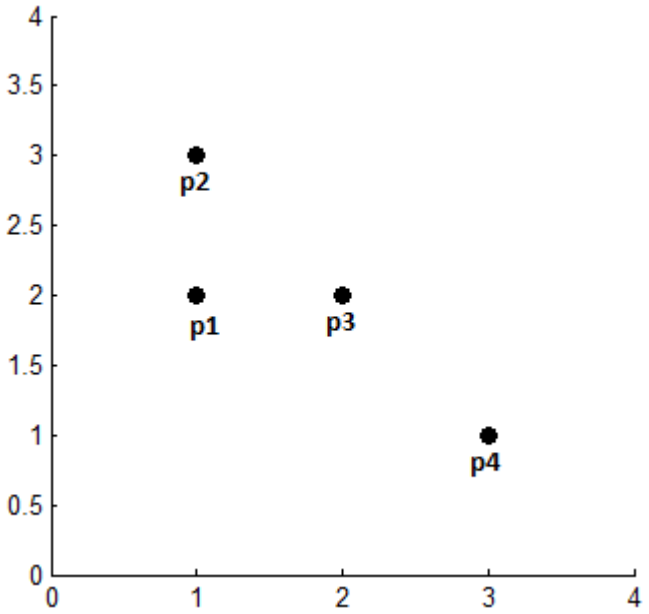
**Τα σημεία είναι κατανεμημένα σε σχήμα έλλειψης αντί για κύκλου!**

$$d(x,y) = \sqrt{(x-y)C^{-1}(x-y)^T}$$

$$C = \begin{pmatrix} 0.9167 & -0.6667 \\ -0.6667 & 0.6667 \end{pmatrix}$$

$$d(p_3, p_2) = \sqrt{1.5} = 1.2247$$

$$d(p_3, p_1) = 2$$




(με Ευκλείδεια το p1 είναι πιο κοντά στο p3 από ότι το p2).

# Απόσταση για δυαδικά δεδομένα

$$s(x,y) = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

$$s(x,y) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \quad \text{(Jaccard)}$$

# Σημείωμα Αναφοράς



Διαθέσιμο από τη δικτυακή διεύθυνση:http://eclass.auth.gr/courses/OCRS182/

# Σημείωμα Αδειοδότησης

ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

# Τέλος ενότητας

Επεξεργασία: Ανδρέας Κοσματόπουλος

Θεσσαλονίκη, Χειμερινό Εξάμηνο 2013-2014



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



# Σημειώματα

# Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.