ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Αποθήκες Δεδομένων και Εξόρυξη Γνώσης από Δεδομένα

Ειδική Κατηγοριοποίηση – Bayesian Κατηγοριοποίηση

Μανώλης Μαραγκουδάκης

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

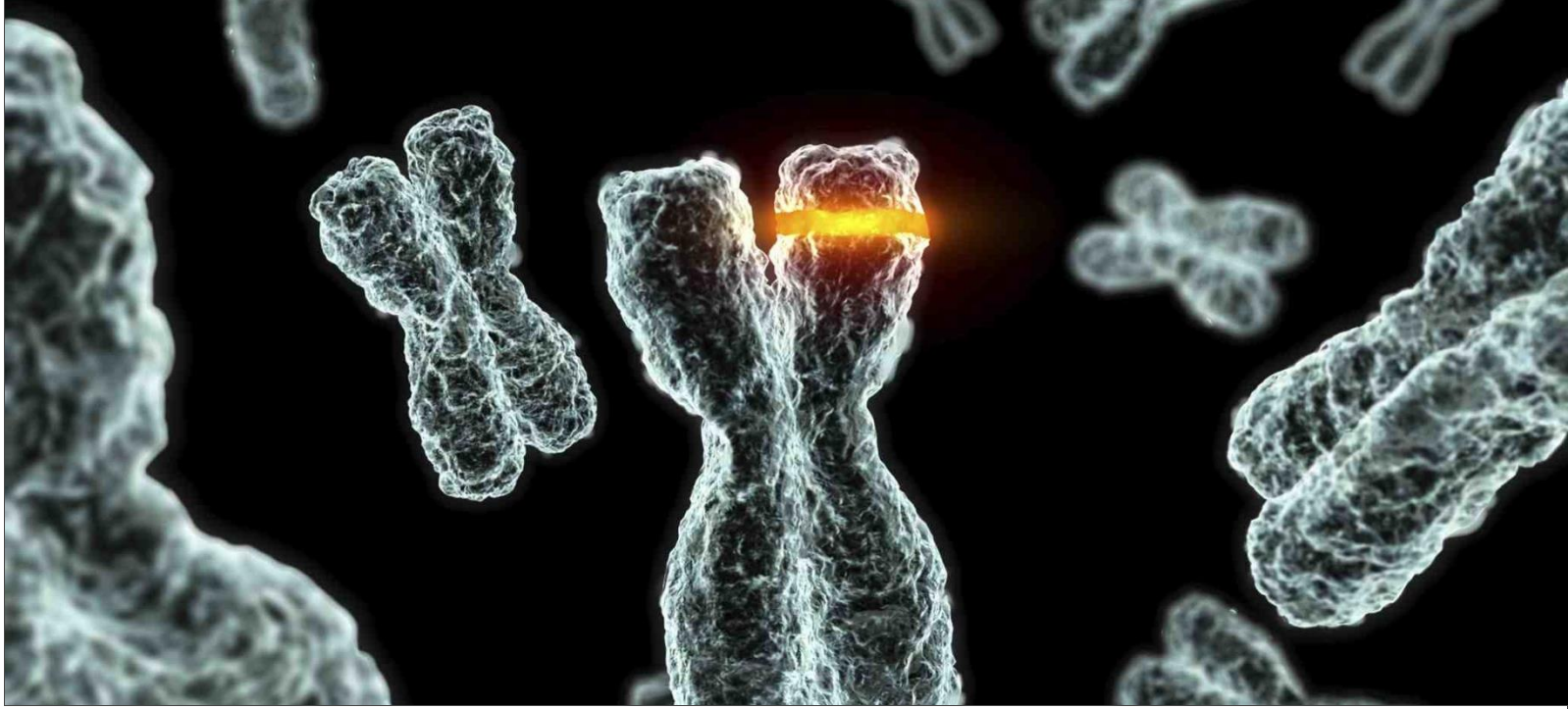
Άδειες Χρήσης

- 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 Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη 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.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



Ενότητα: Bayesian κατηγοριοποίηση

Κατηγοριοποίηση με Bayes

- Ένα πιθανοτικό πλαίσιο για κατηγοριοποίηση
- Υπό συνθήκη πιθανότητα: $P(C \mid A) = \frac{P(A,C)}{P(A)}$
- Θεώρημα Bayes:

$$P(A \mid C) = \frac{P(A,C)}{P(C)}$$

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Παράδειγμα του θεωρήματος Bayes

- Δεδομένων ότι:
  - Ένας γιατρός γνωρίζει οτι η μηνιγγίτιδα (M) προκαλεί αγκύλωση στο λαιμό (S) σε 50% των περιπτώσεων
  - Η εκ των προτέρων πιθανότητα ενός ασθενή να έχει μηνιγγίτιδα είναι 1/50.000
  - Η εκ των προτέρων πιθανότητα ενός ασθενή να έχει αγκύλωση στο λαιμό είναι 1/20
- Εάν ένας ασθενής παρουσιάσει αγκύλωση στο λαιμό, ποια είναι η πιθανότητα να έχει μηνιγγίτιδα;

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Κατηγοριοποίηση με Bayes

- Θεωρήστε κάθε ιδιότητα και την ετικέτα της κλάσης ως τυχαίες μεταβλητές

- Δοσμένης μιας εγγραφής με ιδιότητες (A_1, A_2,…,A_n)
  - Ο στόχος είναι η πρόβλεψη της κλάσης C
  - Ειδικότερα, θέλουμε να βρούμε την τιμή του C που μεγιστοποιεί την πιθανότητα $P(C| A_1, A_2,…,A_n)$

- Μπορεί η πιθανότητα $P(C| A_1, A_2,…,A_n)$ να εκτιμηθεί από τα δεδομένα;

Κατηγοριοποίηση με Bayes

- Προσέγγιση:
  - Υπολογισμός της μεταγενέστερης πιθανότητας $P(C \mid A_1, A_2, \ldots, A_n)$ για όλες τις τιμές C με βάση το θεώρημα Bayes

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - Επιλογή της τιμής του C που μεγιστοποιεί την πιθανότητας $P(C \mid A_1, A_2, \ldots, A_n)$
  - Ανάλογη της επιλογής της τιμής του C που μεγιστοποιεί την $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$

- Πως θα εκτιμήσουμε την $P(A_1, A_2, \ldots, A_n \mid C)$;

Κατηγοριοποίηση με Naïve Bayes

- Αν υποθέσουμε ότι οι ιδιότητες A_i είναι ανεξάρτητες δοσμένης της κλάσης:
  - $P(A_1, A_2, \ldots, A_n | C) = P(A_1 | C_j)\, P(A_2 | C_j) \ldots P(A_n | C_j)$
  - Μπορούμε να υπολογίσουμε την $P(A_i | C_j)$ για όλα τα A_i και C_j.
  - Ένα νέο παράδειγμα κατηγοριοποιείται ως C_j αν η παράσταση $P(C_j) \prod P(A_i | C_j)$ είναι μέγιστη.

Πως υπολογίζουμε τις πιθανότητες από τα δεδομένα;

Tid	Refund	Marital Status	Taxable Income	Evade
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

- Κλάση: $P(C) = N_c/N$
  - Π.χ., P(No) = 7/10,
    P(Yes) = 3/10
- Για διακριτές ιδιότητες:

$$P(A_i \mid C_k) = |A_{ik}| / N_c \qquad k$$

  - όπου $|A_{ik}|$ είναι ο αριθμός των παραδειγμάτων που έχουν την ιδιότητα A_i και ανήκει στην κλάση C_k
  - Παράδειγμα:

    P(Status=Married|No) = 4/7
    P(Refund=Yes|Yes)=0

Πως υπολογίζουμε τις πιθανότητες από τα δεδομένα;

- Για συνεχείς ιδιότητες:
  - Διακριτοποίηση της κλίμακας σε διαστήματα
    - Παραβιάζει την υπόθεση ανεξαρτησίας
  - Διπλός διαχωρισμός: $(A < v)$ or $(A > v)$
    - επιλογή μόνο ενός από τους δυο διαχωρισμούς ως νέα ιδιότητα
  - Εκτίμηση της πυκνότητα πιθανότητας:
    - Υποθέτουμε ότι οι ιδιότητες ακολουθούν την κανονική κατανομή
    - Χρησιμοποιούμε τα δεδομένα για να εκτιμήσουμε τις παραμέτρους της κατανομής (π.χ., μέση τιμή και διασπορά)
    - όταν γίνει γνωστή η κατανομή, μπορεί να χρησιμοποιηθεί για να υπολογισθεί η πιθανότητα $P(A_i|c)$

Πως υπολογίζουμε τις πιθανότητες από τα δεδομένα;

Tid	Refund	Marital Status	Taxable Income	Evade
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

- Κανονική κατανομή

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - Μια για κάθε ζεύγος (A_i,c_i)
- Για (Income=120K, Class=No):
  - αν Class=No
    - Μέση τιμή = 110
    - Διασπορά= 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Παράδειγμα Naïve Bayes

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × 1.2 × 10^{-9} = 0

Αφού P(X|No)P(No) > P(X|Yes)P(Yes)

Τότε P(No|X) > P(Yes|X)

=> Class = No

Κατηγοριοποίηση με Naïve Bayes

- Αν κάποια από τις πιθανότητες είναι μηδενική, όλη η παράσταση γίνεται μηδενική
- Εκτίμηση πιθανότητας:

$$\text{Αρχική: } P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace: } P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m-estimate: } P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: αριθμός κλάσεων

p: εκ των προτέρων πιθανότητα

m: παράμετρος

Παράδειγμα Naïve Bayes

A: ιδιότητες

M: mammals

N: non-mammals

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Naïve Bayes (Περίληψη)

- Εύρωστη σε μεμονωμένα σημεία θορύβου
- Χειρίζεται τις τιμές που λείπουν με το να αγνοούν το παράδειγμα κατά τη διάρκεια του υπολογισμού των πιθανοτήτων
- Εύρωστη σε μη σχετικές ιδιότητες
- Δεν ισχύει πάντα η υπόθεση περί ανεξαρτησίας
  - Χρησιμοποιούνται άλλες τεχνικές όπως τα δίκτυα Bayes - Bayesian Belief Networks (BBN)

Εισαγωγή- Bayesian Inference

- Πρόβλημα:
  - Σε μια ειδική κλινική το 0.15 των ασθενών έχει τον ιό HIV. Αν κάποιος ασθενής κάνει ένα αιματολογικό τεστ και έχει τον ιό, θα βρεθεί θετικό με πιθανότητα 0.95, ενώ αν δεν έχει τον ιό το τεστ θα είναι θετικό με πιθανότητα 0.02.
  - Αν ένας ασθενής βρεθεί θετικός στο τεστ, ποια είναι η πιθανότητα ο ασθενής να:
    - a) έχει τον ιό
    - b) μην έχει τον ιό
  - Αν ένας ασθενής βρεθεί αρνητικός στο τεστ, ποια είναι η πιθανότητα ο ασθενής να:
    - a) έχει τον ιό
    - b) μην έχει τον ιό

Εισαγωγή

- Λύση:

Ας δώσουμε στα γεγονότα τα εξής ονόματα:

H = ο ασθενής έχει τον ιό

P = το αποτέλεσμα του τεστ είναι θετικό

Από το δεδομένα έχουμε: $P(P \mid H) = 0.95$ $P(P \mid \overline{H}) = 0.02$

***P*(*H*) = 0.15**

Και μας ζητούν τα ακόλουθα:

a) *P(H | P)*

b) $P(\overline{H} \mid P)$

c) $P(H \mid \overline{P})$

d) $P(\overline{H} \mid \overline{P})$

Εισαγωγή

- A) $P(H \mid P) = \frac{P(P \mid H)P(H)}{P(P)}$

$P(P) = P(H \wedge P) + P(\overline{H} \wedge P)$

- Από το δεύτερο αξίωμα των πιθανοτήτων έχουμε:
- $P(H \wedge P) = P(P \mid H)P(H)$
  - και
    $P(\overline{H} \wedge P) = P(P \mid \overline{H})P(\overline{H})$
- $P(P) = P(P \mid H)P(H) + P(P \mid \overline{H})P(\overline{H})$
- Άρα
- και επομένως:
- $P(H \mid P) = \frac{(P \mid H)P(H)}{(P \mid H)P(H) + P(P \mid \overline{H})P(\overline{H})}$ =0.8934

Εισαγωγή

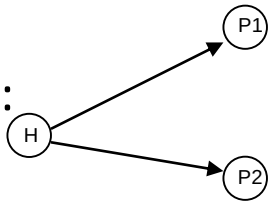
- B) $P(\overline{H}\,|\,P) = 1 - P(H\,|\,P)$ = 0.1066
- C) $P(H\,|\,\overline{P}) = \frac{P(\overline{P}\,|\,H)P(H)}{P(\overline{P})}$ = 0.008923
- D) $P(\overline{H}\,|\,\overline{P}) = 1 - P(H\,|\,\overline{P})$ = 0.99107

Δίκτυα Bayes

- Μέχρι στιγμής είδαμε πως η Bayesian θεωρία πιθανοτήτων μπορεί να συσχετίσει δυο γεγονότα. Π.χ. την πιθανότητα ένας που οδηγάει μια Ferrari να είναι πλούσιος. Όμως η θεωρία Bayes μπορεί να συσχετίσει πολλά γεγονότα, δένοντας τα σε ένα δίκτυο.

- Ας ξαναδούμε το προηγούμενο παράδειγμα πάλι:
  - Ας υποθέσουμε ότι ο ασθενής ξανακάνει ένα ακόμη τεστ, ***ανεξάρτητο*** από το άλλο (δηλ. αν έχει γίνει κάποιο σφάλμα στο πρώτο τεστ δεν σημαίνει ότι αυτό θα επηρεάσει την πιθανότητα σφάλματος του δεύτερου)

- Το ακόλουθο διάγραμμα μας δείχνει την παραπάνω σχέση:



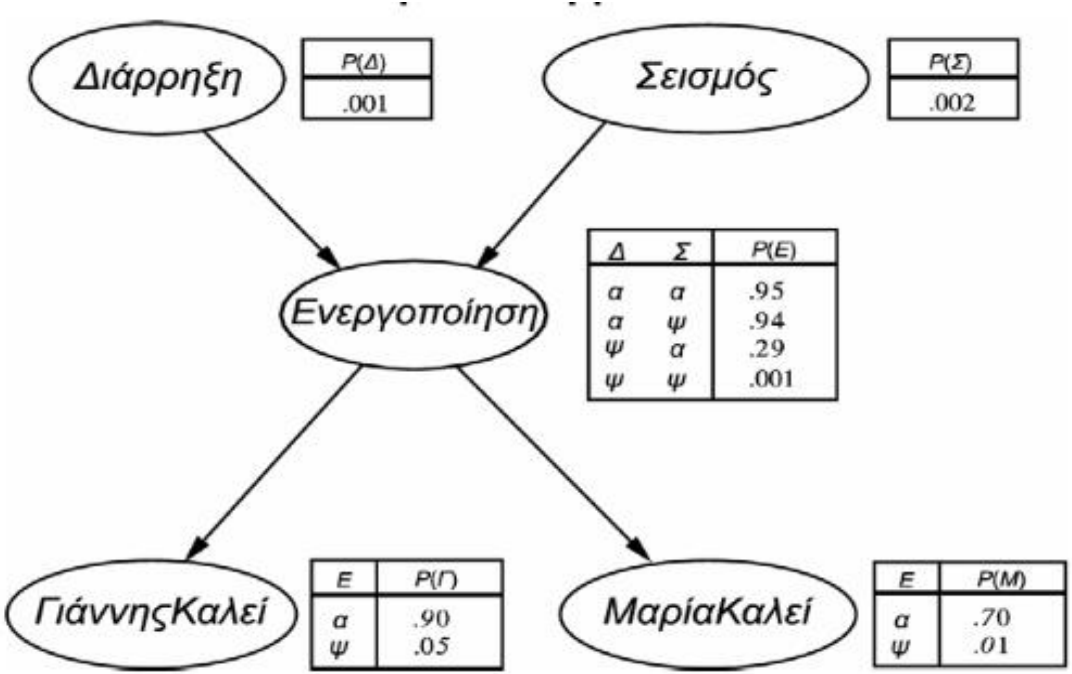
Δίκτυα Bayes

- Έστω ότι ο ασθενής κάνει 2 τεστ και βγαίνουν και τα 2 θετικά. Ποια η πιθανότητα να έχει τον ιό;
- $$P(H \mid P1 \wedge P2) = \frac{P(P1 \wedge P2 \mid H)P(H)}{P(P1 \wedge P2)}$$
  - $$P(P1 \wedge P2) = P(P1 \wedge P2 \mid H)P(H) + P(P1 \wedge P2 \mid \overline{H})P(\overline{H})$$
  - $P(P1 \wedge P2|H) = (P1|H)P(P2|H)$
  - $$P(H \mid P1 \wedge P2) = \frac{0.95 \times 0.95 \times 0.15}{0.135715} = 0.99749$$
- Στο προηγούμενο παράδειγμα με το ένα τεστ θετικό, η πιθανότητα ήταν 0.8934, τώρα είναι πολύ μεγαλύτερη.
- Τα δυο τεστ έδωσαν μεγαλύτερη πεποίθηση (belief) στη γνώση μας. (σσ. Λέγονται και ***Bayesian Belief Networks***)

Δίκτυα Bayes

1. Ένα σύνολο τυχαίων μεταβλητών (διακριτών ή συνεχών) σχηματίζει τους κόμβους του γραφήματος.
2. Ένα σύνολο κατευθυνόμενων συνδέσμων ή βελών συνδέει ζευγάρια κόμβων. Αν υπάρχει βέλος από τον κόμβο *X* στον κόμβο *Y*, τότε λέμε ότι ο *X* είναι *γονέας* του *Y*.
3. Ο κάθε κόμβος X_i έχει μια υπό συνθήκη κατανομή πιθανότητας $P(X_i|\mathit{Γονείς}(X_i))$.
4. Το γράφημα δεν έχει καθόλου κατευθυνόμενους κύκλους.

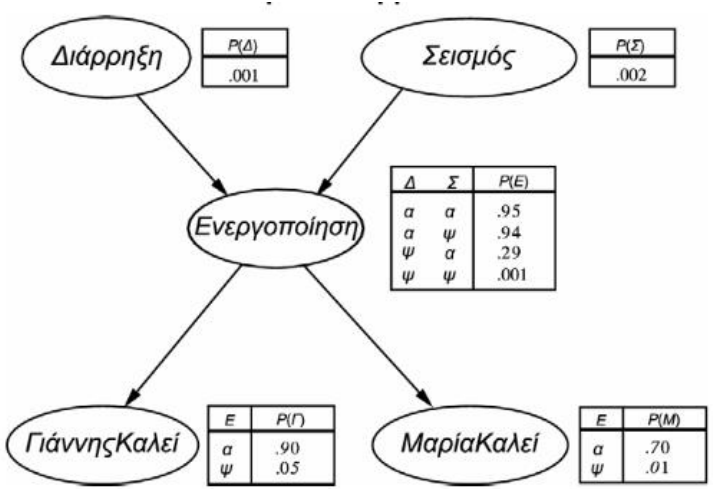
Παράδειγμα


Διάρρηξη
P(Δ)
.001
Σεισμός
P(Σ)
.002
Ενεργοποίηση
Δ Σ P(Ε)
α α .95
α ψ .94
ψ α .29
ψ ψ .001
ΓιάννηςΚαλεί
Ε P(Γ)
α .90
ψ .05
ΜαρίαΚαλεί
Ε P(Μ)
α .70
ψ .01

Joint Probability Distribution

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \gamma o \nu \varepsilon \acute{\iota} \varsigma(x_i))$$

- Για παράδειγμα:
  - $P(\gamma \wedge \mu \wedge \varepsilon \wedge \neg\delta \wedge \neg\sigma)$
  - $= P(\gamma \mid \varepsilon)\, P(\mu \mid \varepsilon)\, P(\varepsilon \mid \neg\delta \wedge \neg\sigma)\, P(\neg\delta)\, P(\neg\sigma)$
  - $= 0{,}90 \times 0{,}70 \times 0{,}001 \times 0{,}999 \times 0{,}998 = 0{,}00062$



Από πού προκύπτει;

- Κανόνας αλυσίδας:

$$P(x_1, x_2, \ldots, x_n) = P(x_n \mid x_{n-1}, \ldots, x_1) \cdot P(x_{n-1} \mid x_{n-2}, \ldots, x_1)$$

$$\cdot \ldots \cdot P(x_2 \mid x_1) \cdot P(x_1) =$$

$$\prod_{i=1}^{n} P(x_i \mid x_{i-1}, \ldots, x_1)$$

- Εάν *Γονείς*(X_i) $\subseteq \{X_{i-1}, \ldots, X_1\}$ τότε:
  - **P**($X_i \mid X_{i-1}, \ldots, X_1$) = **P** ($X_i \mid$ *Γονείς*(X_i))
- Για παράδειγμα:
  - **P**(*ΜαρίαΚαλεί* | *ΓιάννηςΚαλεί*, *Ενεργοποίηση*, *Σεισμός*, *Διάρρηξη*) = **P**(*ΜαρίαΚαλεί* | *Ενεργοποίηση*)

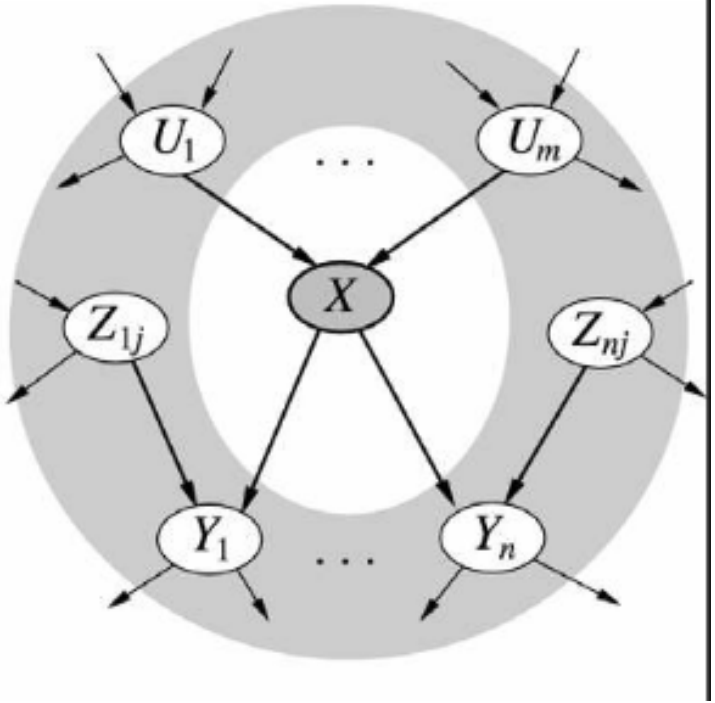
Συμπαγής Αναπαράσταση

- Για *n* Boolean μεταβλητές, κάθε μία από τις οποίες έχει *k* το πολύ γονείς:
  - Η πλήρης συνδυασμένη κατανομή απαιτεί 2^n αριθμούς.
    - Για n=10 είναι 1024.
  - Το δίκτυο Bayes απαιτεί $n \cdot 2^k$ αριθμούς.
    - Για n=10 και k=3 είναι 80.

- *Τοπικά δομημένα συστήματα* ή *αραιά συστήματα*: Κάθε υποστοιχείο αλληλεπιδρά μόνο με ένα φραγμένο πλήθος άλλων στοιχείων, ανεξάρτητα από το συνολικό πλήθος των στοιχείων

Conditional Independence Assumption

- Ένας κόμβος είναι υπό συνθήκη ανεξάρτητος από όλους τους υπόλοιπους κόμβους του δικτύου, με δεδομένους τους γονείς του, τα παιδιά του, και τους γονείς των παιδιών του — δηλαδή, με δεδομένο το **κάλυμμα Markov** (Markov blanket) για τον κόμβο αυτόν.



Inference-Συμπερασμός

Η πιθανότητα του ερωτήματος ***X*** δοθέντος του συμβάντος ***e*** είναι:

$$\mathbf{P}(X|\mathbf{e}) = \alpha \cdot \mathbf{P}(X,\mathbf{e}) = \alpha \cdot \sum_{y} \mathbf{P}(X,\mathbf{e},\mathbf{y})$$

όπου *α* παράγοντας κανονικοποίησης της πιθανότητας [0-1], Y οι μη συσχετιζόμενες μεταβλητές

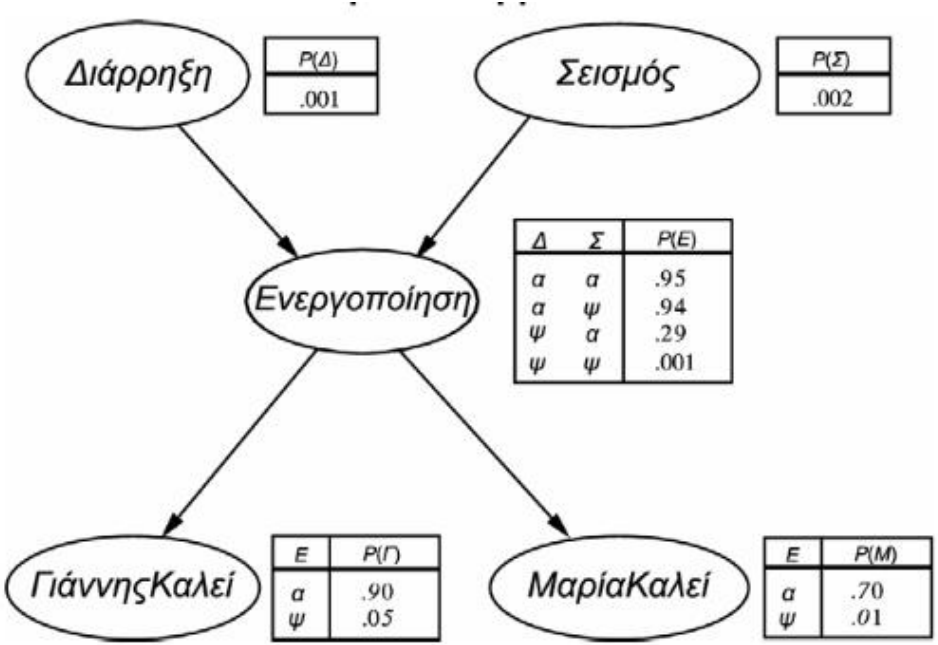
- Για *Διάρρηξη* = *αληθές* έχουμε:

$$\mathbf{P}(\delta|\gamma,\mu) = \alpha \cdot \sum_{\sigma}\sum_{\varepsilon} P(\delta)P(\sigma)P(\varepsilon|\delta,\sigma)P(\gamma|\varepsilon)P(\mu|\varepsilon)$$

- Πολυπλοκότητα υπολογισμού: $O(n2^n)$ στη χειρότερη περίπτωση.
- Βγάζοντας κάποιους όρους έξω από τα αθροίσματα έχουμε:

$$\mathbf{P}(\delta|\gamma,\mu) = \alpha \cdot P(\delta) \cdot \sum_{\sigma} P(\sigma) \sum_{\varepsilon} P(\varepsilon|\delta,\sigma)P(\gamma|\varepsilon)P(\mu|\varepsilon)$$

- Η πολυπλοκότητα μπορεί σαν βελτιωθεί μέχρι και $O(2^n)$.



Συμπερασμός

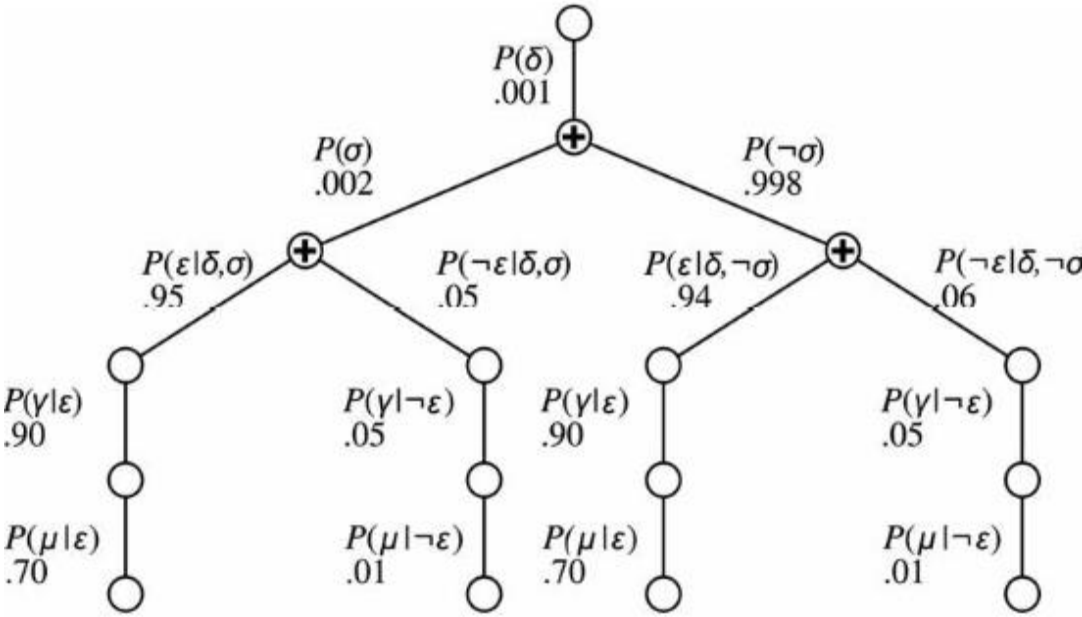
- Για *Διάρρηξη* = *αληθές* έχουμε:

$$\mathbf{P}(\delta|\gamma,\mu) = \alpha \cdot \sum_{\sigma}\sum_{\varepsilon} P(\delta)P(\sigma)P(\varepsilon|\delta,\sigma)P(\gamma|\varepsilon)P(\mu|\varepsilon)$$

- Πολυπλοκότητα υπολογισμού: $O(n2^n)$ στη χειρότερη περίπτωση.
- Βγάζοντας κάποιους όρους έξω από τα αθροίσματα έχουμε:

$$\mathbf{P}(\delta|\gamma,\mu) = \alpha \cdot P(\delta) \cdot \sum_{\sigma} P(\sigma) \sum_{\varepsilon} P(\varepsilon|\delta,\sigma)P(\gamma|\varepsilon)P(\mu|\varepsilon)$$

- Η πολυπλοκότητα μπορεί σαν βελτιωθεί μέχρι και $O(2^n)$.



Κατηγοριοποίηση με βάση τα παραδείγματα

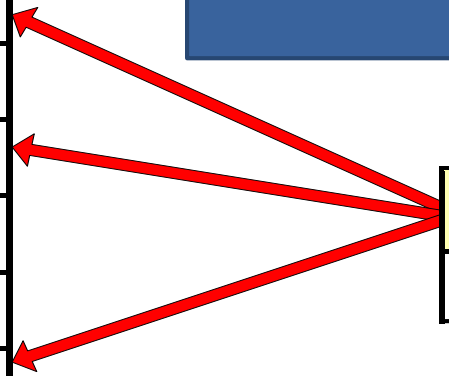
- Αποθήκευση των παραδειγμάτων εκπαίδευσης
- Χρήση τους για την πρόβλεψη της κλάσης ενός νέου παραδείγματος

Σύνολο αποθηκευμένων παραδειγμάτων

Atr1	……	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

Νέο παράδειγμα

Atr1	……	AtrN

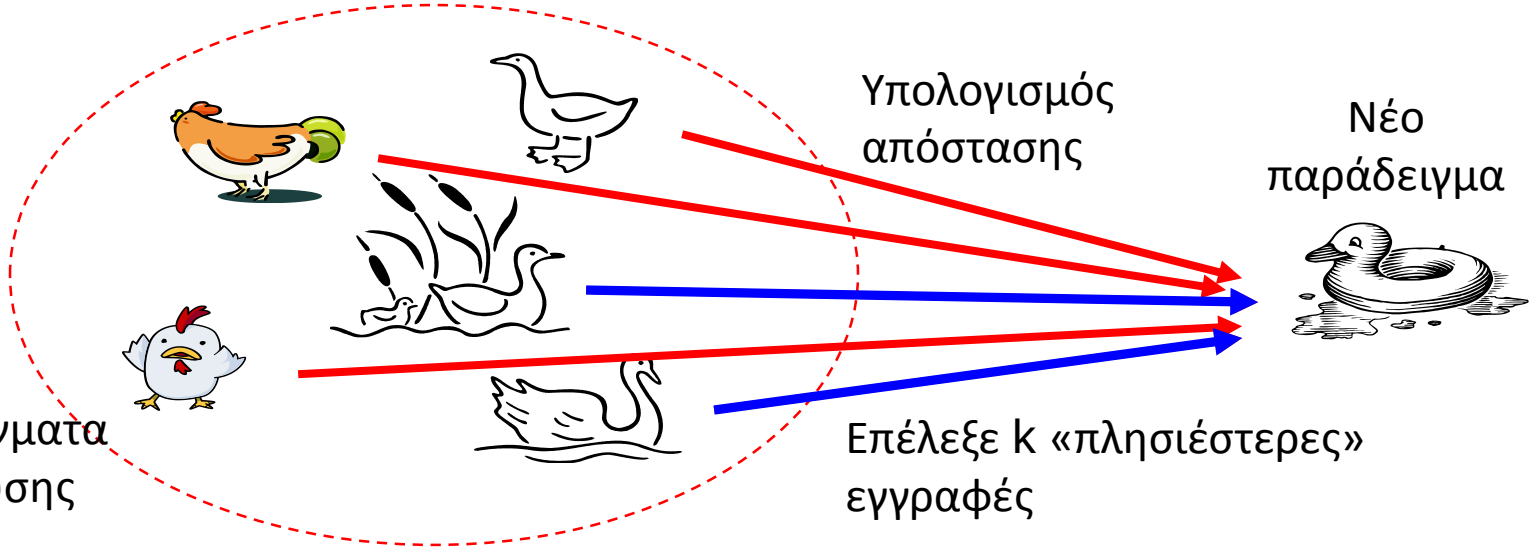


Κατηγοριοποίηση με βάση τα παραδείγματα

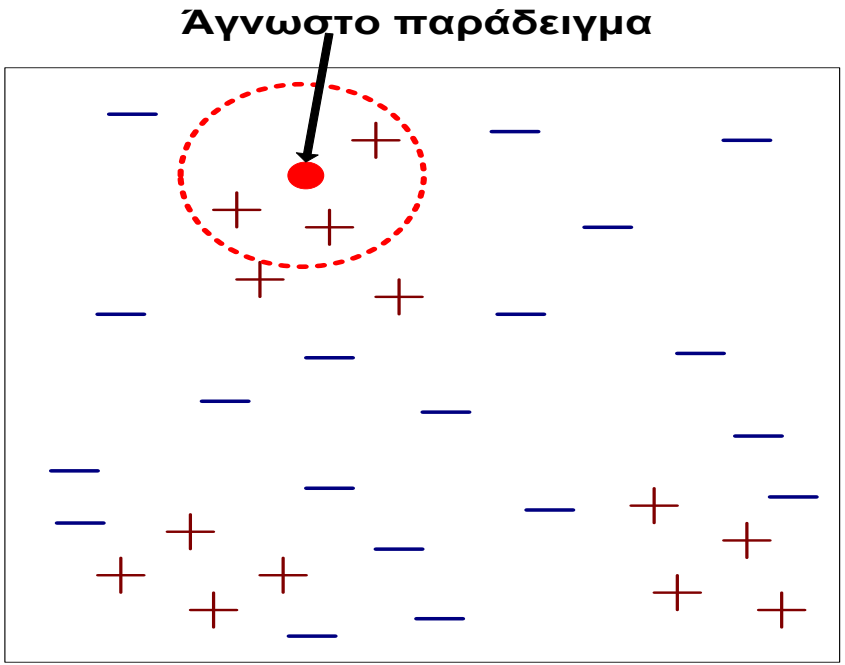
- Παραδείγματα:
  - Rote-learner
    - απομνημονεύει όλο το σύνολο εκπαίδευσης και πραγματοποιεί κατηγοριοποίηση μόνο εάν οι ιδιότητες μιας εγγραφής ταιριάζουν ακριβώς με κάποιο παράδειγμα εκπαίδευσης
  - Πλησιέστερος γείτονας - Nearest neighbor
    - χρησιμοποιεί τα k “πλησιέστερα” σημεία (πλησιέστεροι γείτονες) για την κατηγοριοποίηση

Κατηγοριοποίηση με βάση τα παραδείγματα

- Βασική ιδέα
  - Εάν περπατάει σαν πάπια και κάνει την πάπια (!) τότε είναι πιθανόν να είναι πάπια

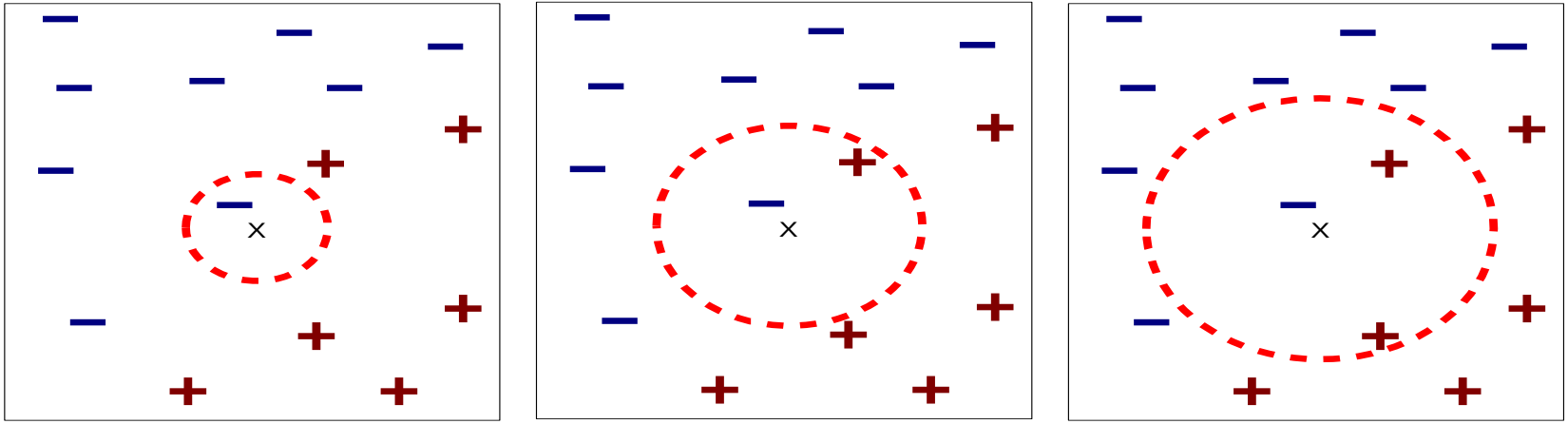


Κατηγοριοποίηση πλησιέστερου γείτονα



- Απαιτεί τρία πράγματα
  - Το σύνολο των αποθηκευμένων εγγραφών
  - Ένα μέτρο απόστασης για τον υπολογισμό της απόστασης μεταξύ των εγγραφών
  - Τον αριθμό των γειτόνων που θα ληφθούν υπόψη (k)
- Για την ταξινόμηση ενός νέου παραδείγματος:
  - Υπολογισμός της απόστασης με τα άλλα παραδείγματα εκπαίδευσης
  - Αναγνώριση των k πλησιέστερων γειτόνων
  - Χρήση των ετικετών κλάσης των γειτόνων για να καθορισθεί η ετικέτα κλάσης του νέου παραδείγματος (π.χ. με πλειοψηφία)

Ορισμός πλησιέστερου γείτονα



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

οι k-πλησιέστεροι γείτονες μιας εγγραφής x είναι σημεία που έχουν την k μικρότερη απόσταση στο x

Κατηγοριοποίηση πλησιέστερου γείτονα

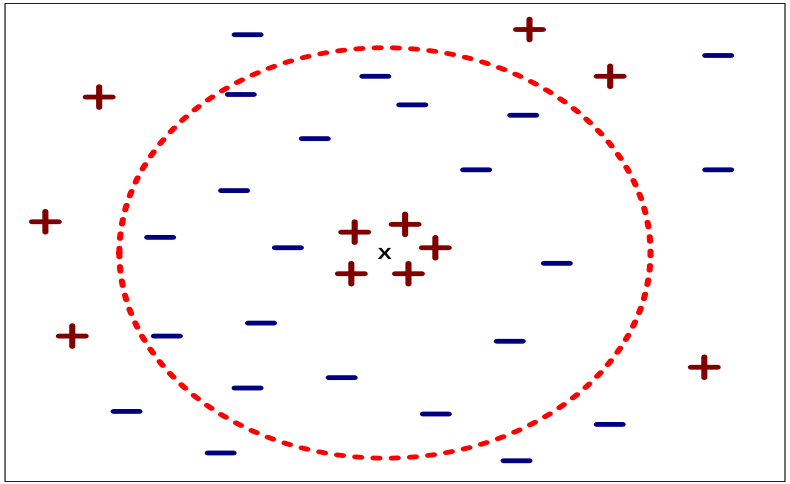
- Υπολογισμός της απόστασης 2 σημείων
  - Ευκλείδεια απόσταση

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Καθορισμός της κλάσης από τη λίστα των πλησιέστερων γειτόνων
  - Η πλειοψηφία των κλάσεων των k πλησιέστερων γειτόνων
  - Καθορισμός βάρους πλειοψηφία με βάση την απόσταση
    - βάρος, $w = 1/d^2$

Κατηγοριοποίηση πλησιέστερου γείτονα

- Επιλέγοντας την τιμή k:
  - Εάν πολύ μικρό, ευαίσθητο σε σημεία θορύβου
  - Εάν πολύ μεγάλο, παρεισφρέουν στην γειτνίαση και άλλες κλάσεις



Κατηγοριοποίηση πλησιέστερου γείτονα

- Ευθυγράμμιση ιδιοτήτων
  - Οι ιδιότητες μπορεί να χρειάζονται αλλαγή κλίμακας για να αποτραπεί η περίπτωση μια ιδιότητα να υπερκαλύπτει τις άλλες στην απόσταση
  - παράδειγμα:
    - το ύψος ενός ατόμου μπορεί να κυμαίνεται από 1.5m έως1.8m
    - το βάρος από 45kg έως 150kg
    - το εισόδημα από 8.000€ έως 1.000.000 €

Κατηγοριοποίηση πλησιέστερου γείτονα

- Πρόβλημα με την Ευκλείδεια απόσταση:
  - Υψηλή διαστατικότητα δεδομένων
    - «κατάρα» της διαστατικότητας
  - Μπορεί να οδηγήσει σε ακριβώς αντίθετα αποτελέσματα

1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1

d = 1.4142

vs

1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

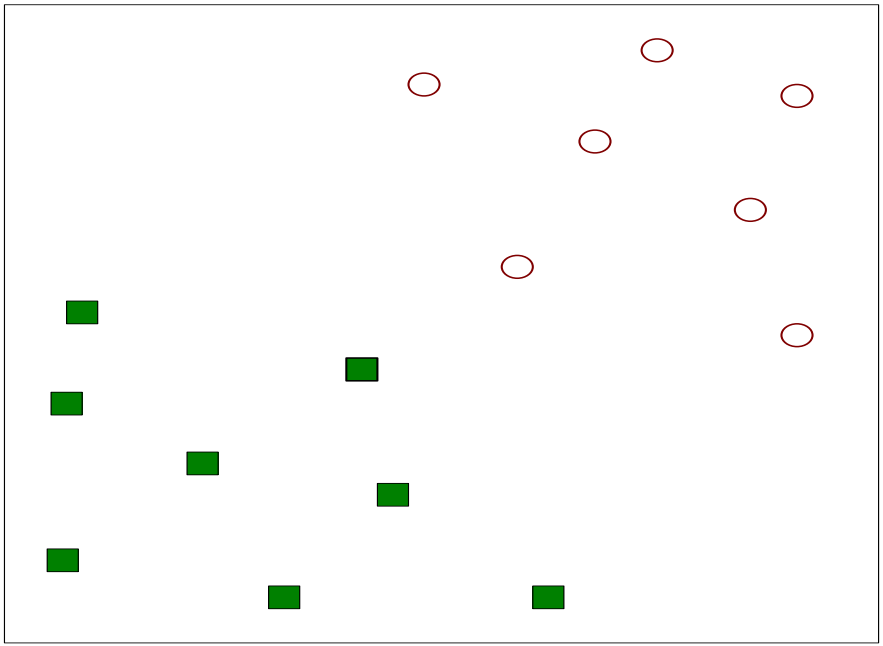
d = 1.4142

- λύση: κανονικοποίηση των διανυσμάτων

Κατηγοριοποίηση πλησιέστερου γείτονα

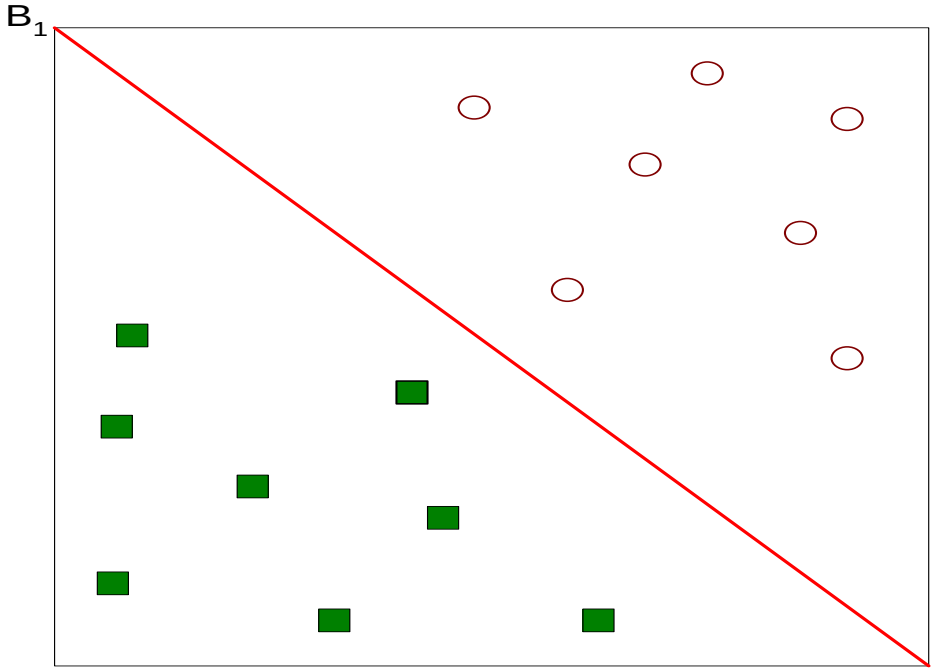
- Οι k-NN ταξινομητές είναι «lazy learners»
  - Δεν φτιάχνουν κάποιο σταθερό μοντέλο
    - Αντίθετα με τα δέντρα αποφάσεων (eager learner)
  - Η κατηγοριοποίηση νέων παραδειγμάτων είναι σχετικά δαπανηρή

Μηχανές διανυσμάτων υποστήριξης Support Vector Machines



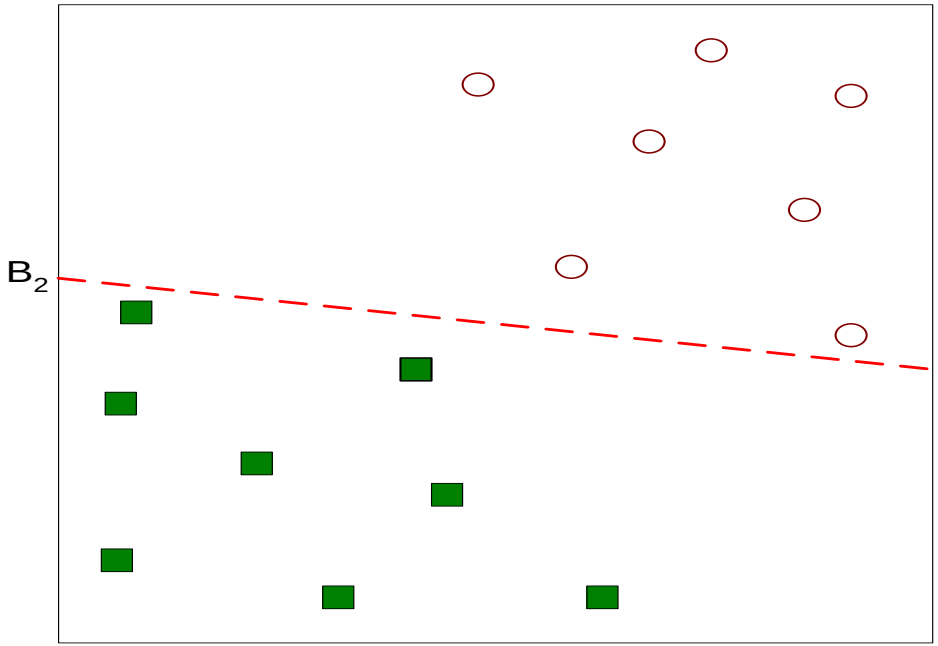
- Εύρεση ενός γραμμικού ορίου απόφασης (hyperplane) που διαχωρίζει τα δεδομένα

Support Vector Machines



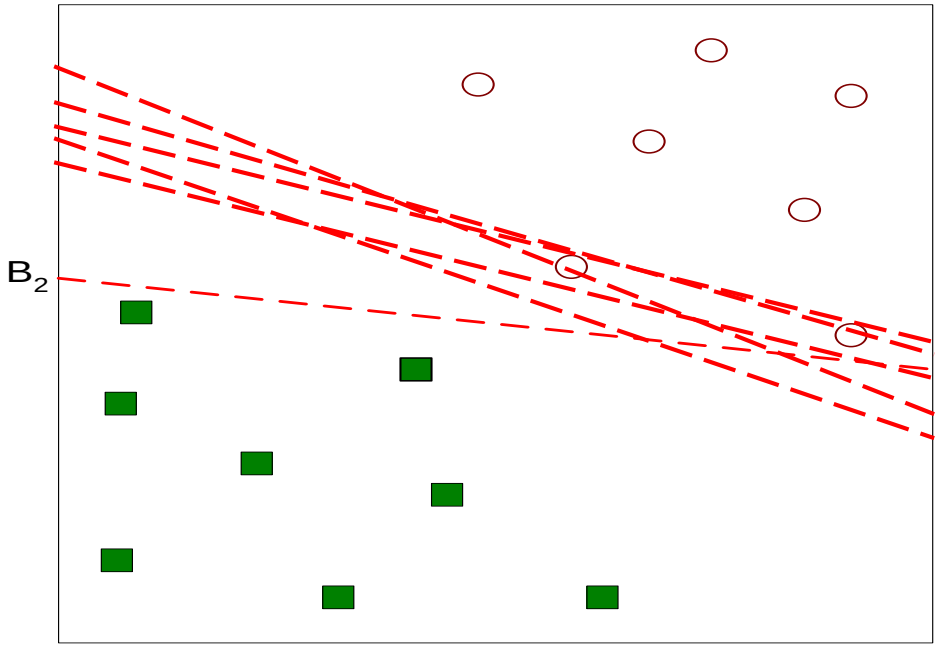
- Μια πιθανή λύση

Support Vector Machines



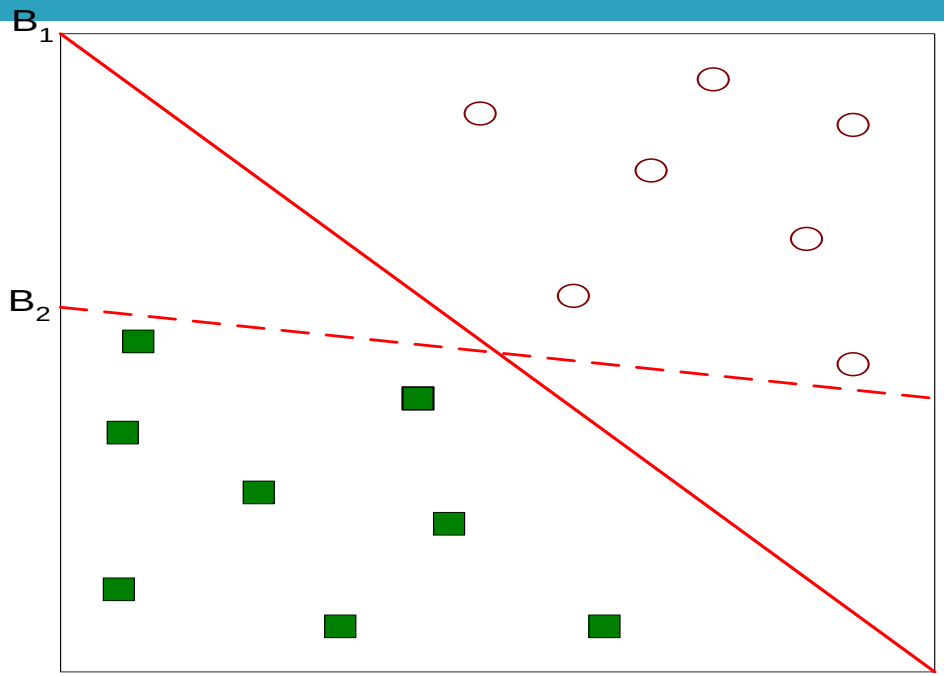
- Μια άλλη εναλλακτική λύση

Support Vector Machines



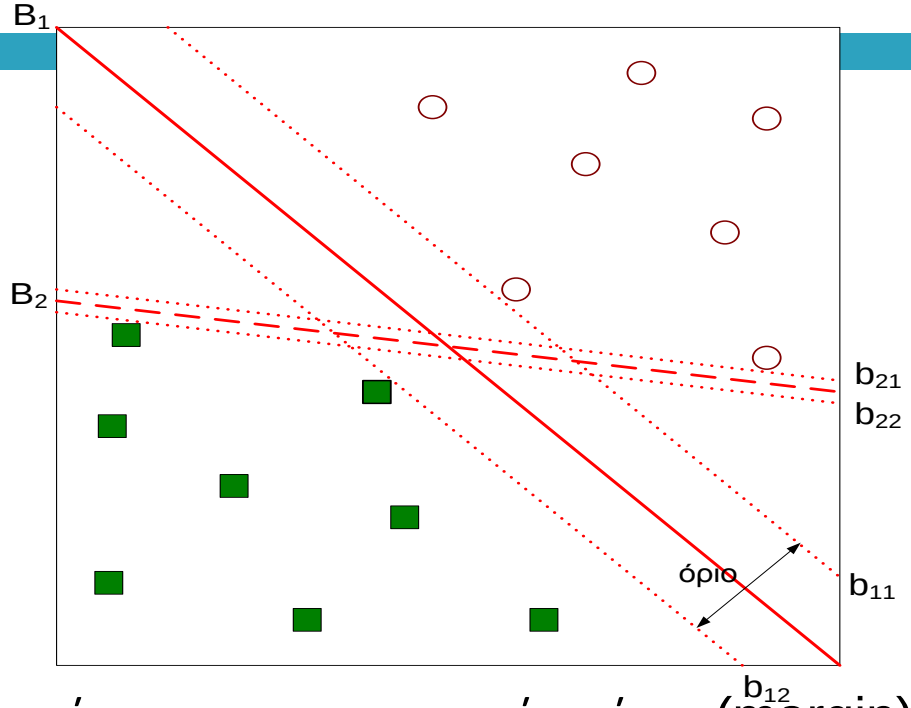
- Άλλες πιθανές λύσεις

Support Vector Machines



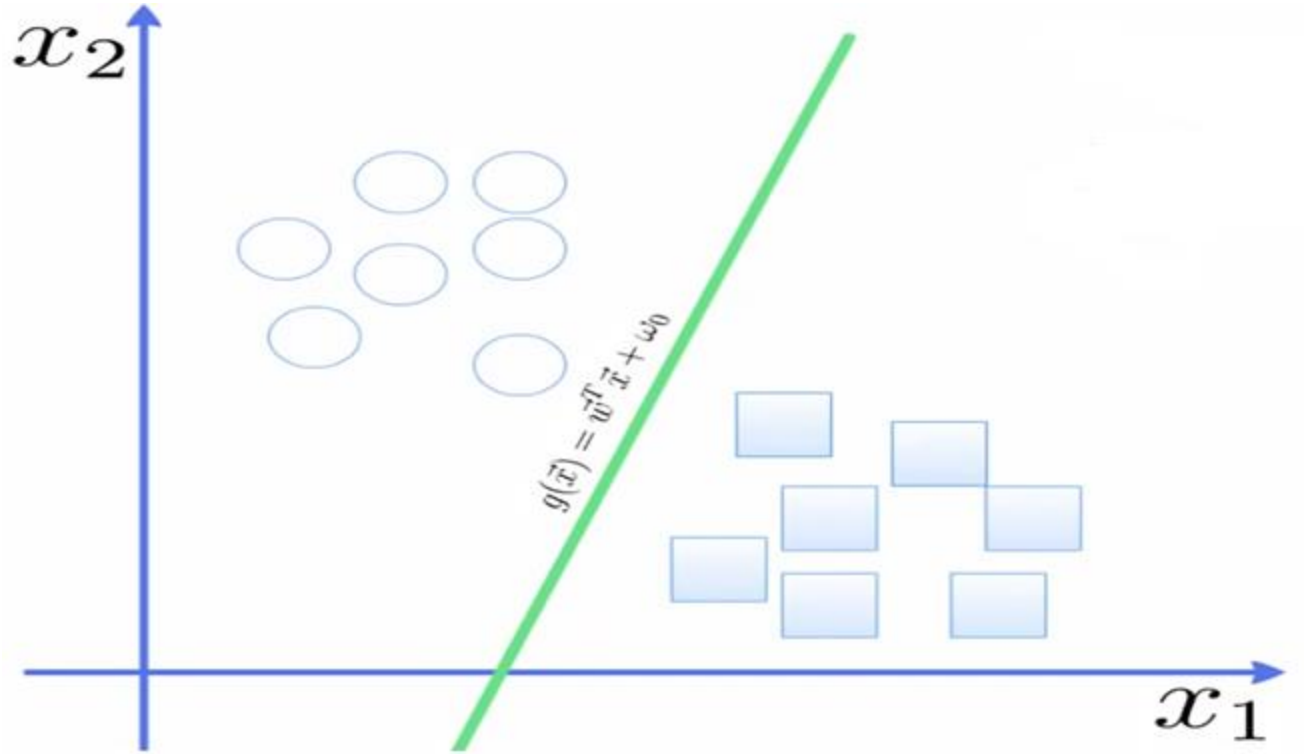
- Ποια είναι καλύτερη λύση; Τη B1 ή B2?
- Πως ορίζεται η έννοια *καλύτερη*;

Support Vector Machines



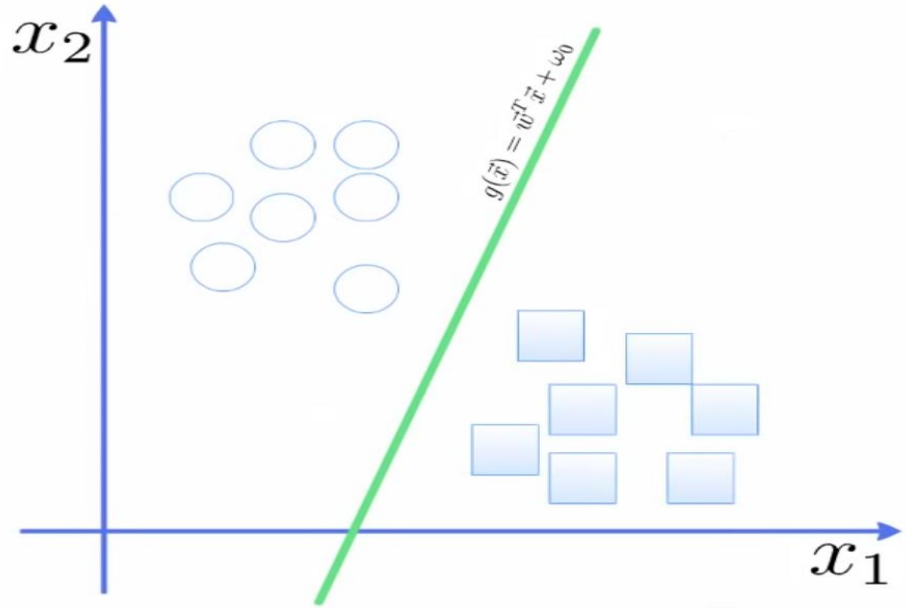
- Εύρεση του ορίου που μεγιστοποιεί το όριο (margin) που διαχωρίζει τα δεδομένα
  - Άρα B1 καλύτερη του B2

Support Vector Machines



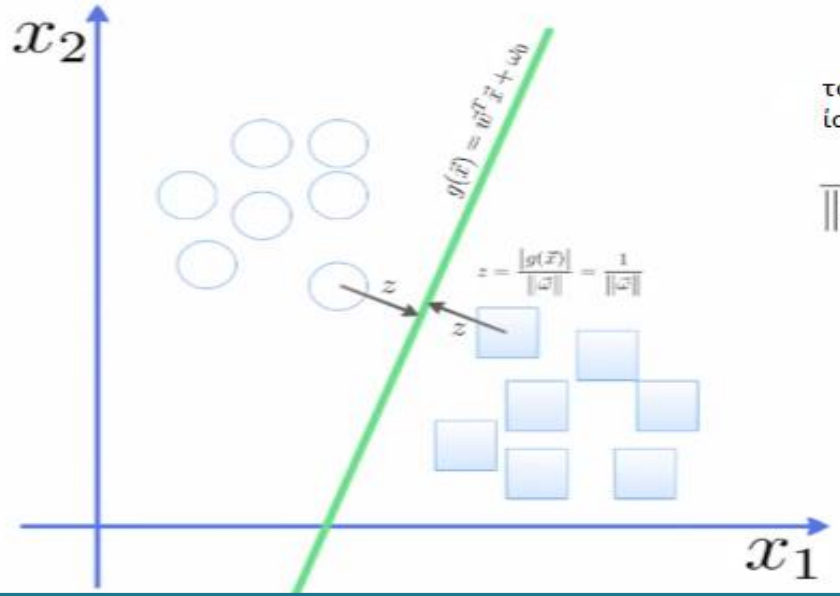
Support Vector Machines

$$g(\vec{x}) \geqslant 1, \quad \forall \vec{x} \in \text{class } 1$$
$$g(\vec{x}) \leqslant -1, \quad \forall \vec{x} \in \text{class } 2$$



Support Vector Machines

$$g(\vec{x}) \geqslant 1, \quad \forall \vec{x} \in \text{class } 1$$
$$g(\vec{x}) \leqslant -1, \quad \forall \vec{x} \in \text{class } 2$$



το συνολικό όριο είναι ίσο με:

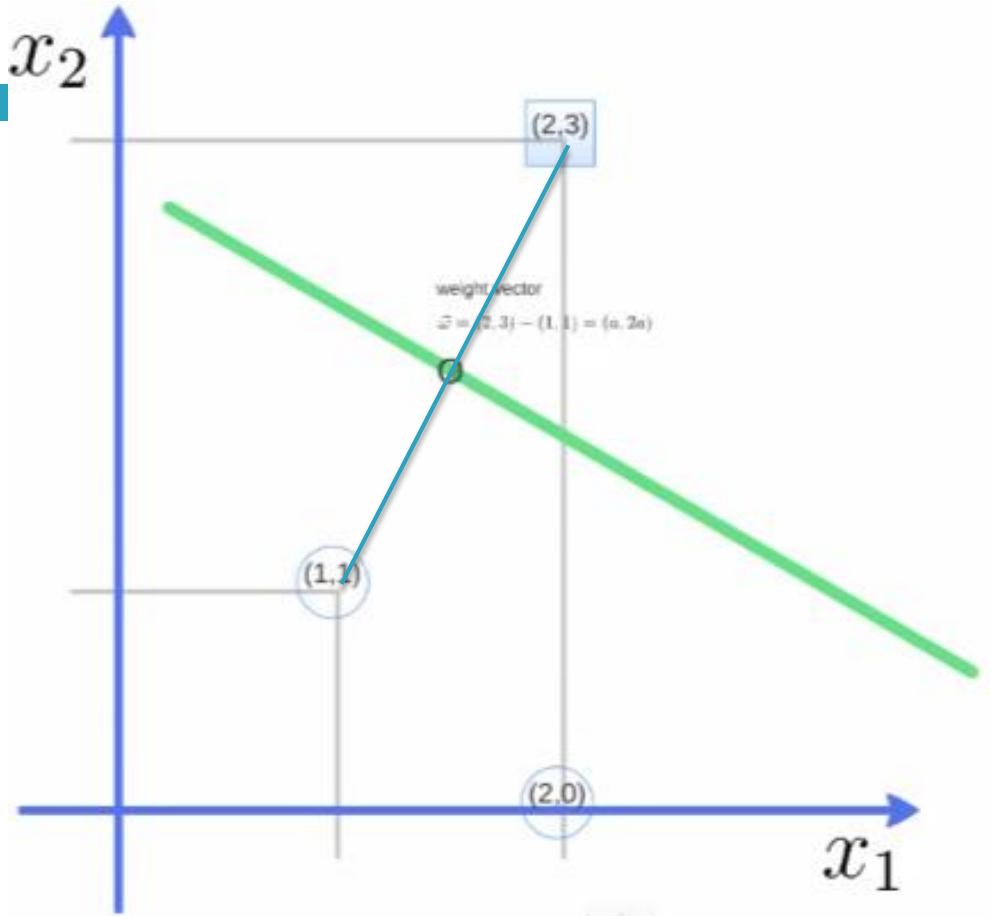
$$\frac{1}{\|\vec{w}\|} + \frac{1}{\|\vec{w}\|} = \frac{2}{\|\vec{w}\|}$$

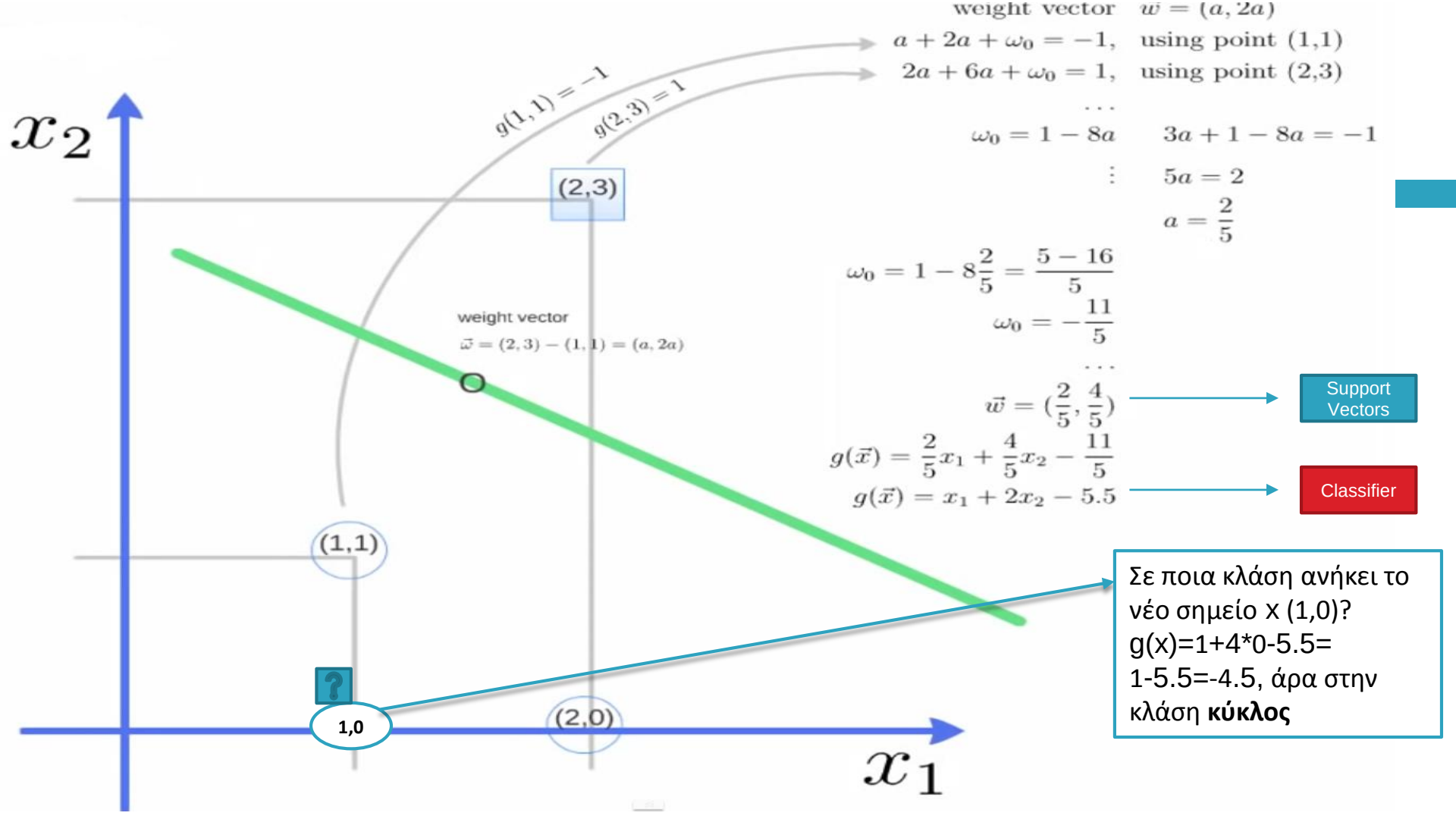
ελαχιστοποιώντας αυτόν τον όρο, μεγιστοποιούμε το διαχωρισμό!

Είναι πρόβλημα βελτιστοποίησης περιορισμών - constrained optimization problem
Επιλύεται με αριθμητικές μεθόδους (π.χ., quadratic programming)

Παράδειγμα Γραμμικά Διαχωρίσιμων Δεδομένων

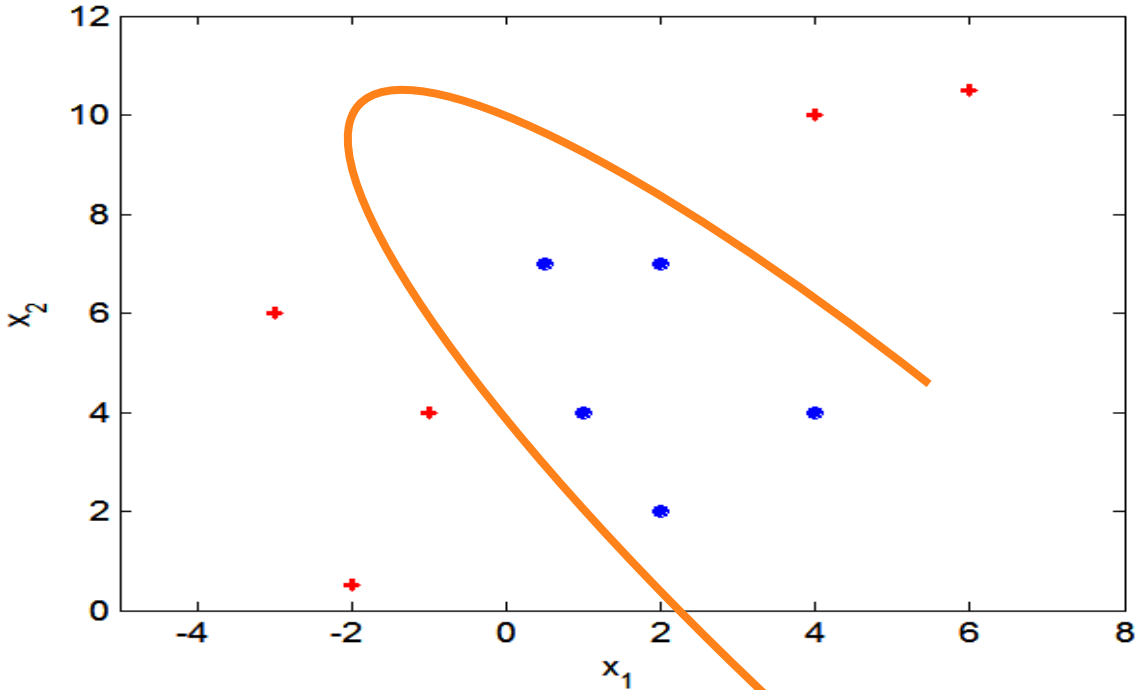
- w=(2,3)-(1,1)=(a,2a)
  - a=σταθερά




weight vector $\vec{w} = (a, 2a)$
$a + 2a + \omega_0 = -1$, using point (1,1)
$2a + 6a + \omega_0 = 1$, using point (2,3)
...
$\omega_0 = 1 - 8a$
$3a + 1 - 8a = -1$
$5a = 2$
$a = \frac{2}{5}$
$\omega_0 = 1 - 8\frac{2}{5} = \frac{5-16}{5}$
$\omega_0 = -\frac{11}{5}$
...
$\vec{w} = (\frac{2}{5}, \frac{4}{5})$
$g(\vec{x}) = \frac{2}{5}x_1 + \frac{4}{5}x_2 - \frac{11}{5}$
$g(\vec{x}) = x_1 + 2x_2 - 5.5$
Support Vectors
Classifier
$g(1,1) = -1$
$g(2,3) = 1$
x_2
x_1
(2,3)
(1,1)
(2,0)
1,0
weight vector
$\vec{w} = (2,3) - (1,1) = (a, 2a)$
Σε ποια κλάση ανήκει το νέο σημείο x (1,0)? g(x)=1+4*0-5.5= 1-5.5=-4.5, άρα στην κλάση **κύκλος**

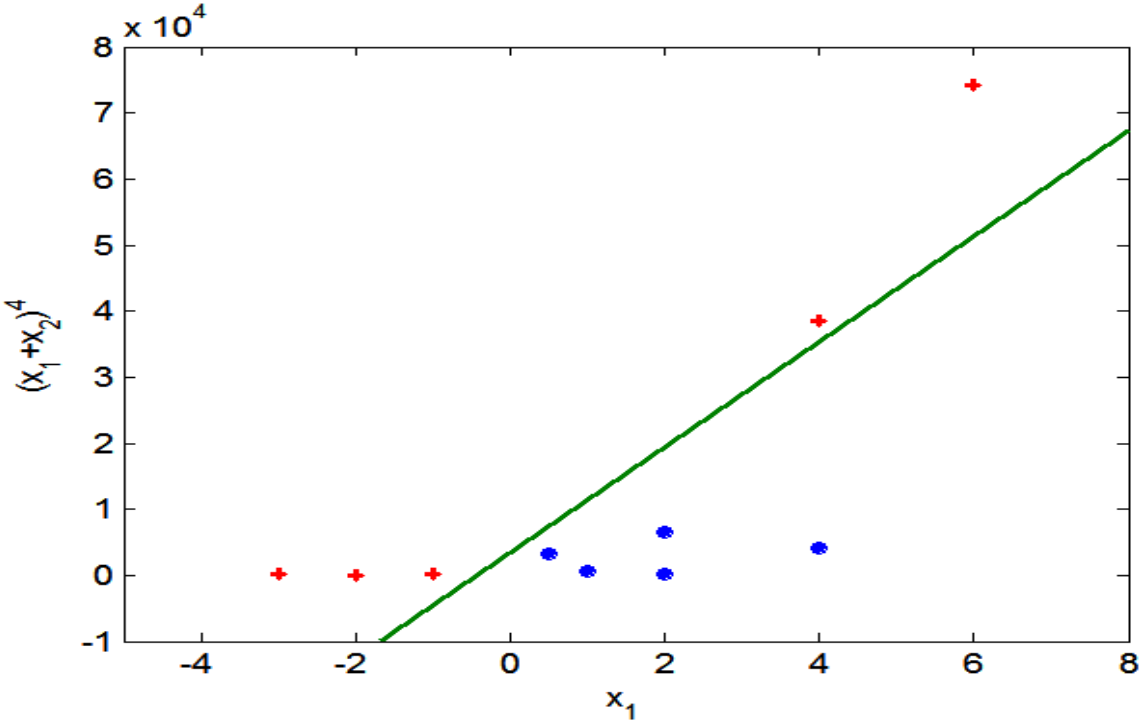
Μη-γραμμικά Support Vector Machines

- Τι γίνεται όταν το όριο απόφασης δεν είναι γραμμικό;



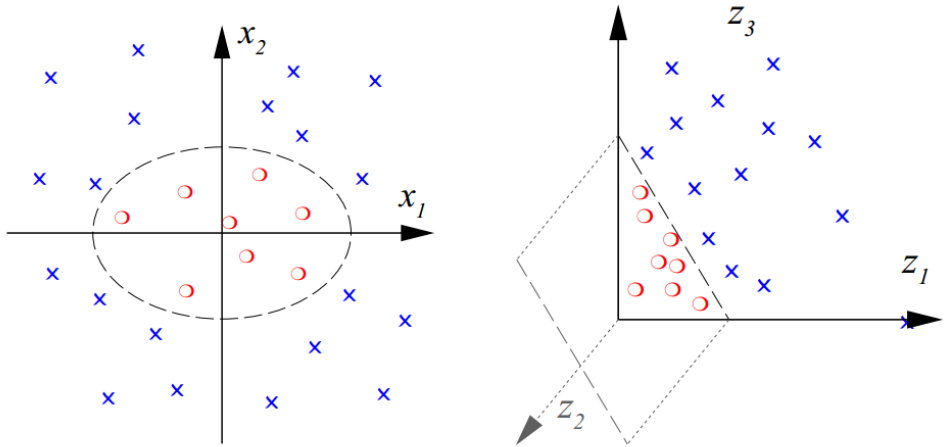
Μη-γραμμικά Support Vector Machines

- Μετασχηματισμός των δεδομένων σε υψηλότερη διάσταση (Kernel Trick)



Kernel Trick-δεύτερο παράδειγμα

$$\Phi : R^2 \rightarrow R^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{(2)}x_1x_2, x_2^2)$$



Παράθυρα Parzen

- Βασική ιδέα:
  - Τα δεδομένα πηγάζουν από διαφορετικές κατανομές πιθανότητας
    - Λογικά τα δεδομένα της κάθε κλάσης θα έχουν την ίδια κατανομή πιθανότητας
    - Έτσι, αν βρούμε τις κατανομές αυτές για κάθε κλάση, μπορούμε να ταξινομήσουμε οποιοδήποτε νέο παράδειγμα
      - Πως;
      - Αντιπαραβάλλοντας το με τις ήδη υπάρχουσες κατανομές και βλέποντας με ποια μοιάζει περισσότερο

Παράθυρα Parzen

- Κάθε δεδομένο εκπαίδευσης θεωρείται πως ακολουθεί μια Gaussian κατανομή με κέντρο το ίδιο το σημείο.
- Επομένως, η συνολική πιθανότητα (για 1 διάσταση) είναι:

$$p(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - x)^2}{2\sigma^2}\right)$$

Παράθυρα Parzen

- Παράδειγμα
- Έστω 5 σημεία
  - x1=2
  - x2=2.5
  - x3=3
  - x4=1
  - x5=6
- Να βρείτε την Parzen pdf στο x=3 για σ=1.

Παράθυρα Parzen

- Λύση:
  - Εφαρμογή του γενικού τύπου για κάθε σημείο

$$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_1-x)^2}{2}\right) \qquad \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_2-x)^2}{2}\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(2-3)^2}{2}\right) = 0.2420 \qquad = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(2.5-3)^2}{2}\right) = 0.3521$$

Παράθυρα Parzen

- Παρομοίως....

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_3 - x)^2}{2}\right) = 0.3989$$

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_4 - x)^2}{2}\right) = 0.0540$$

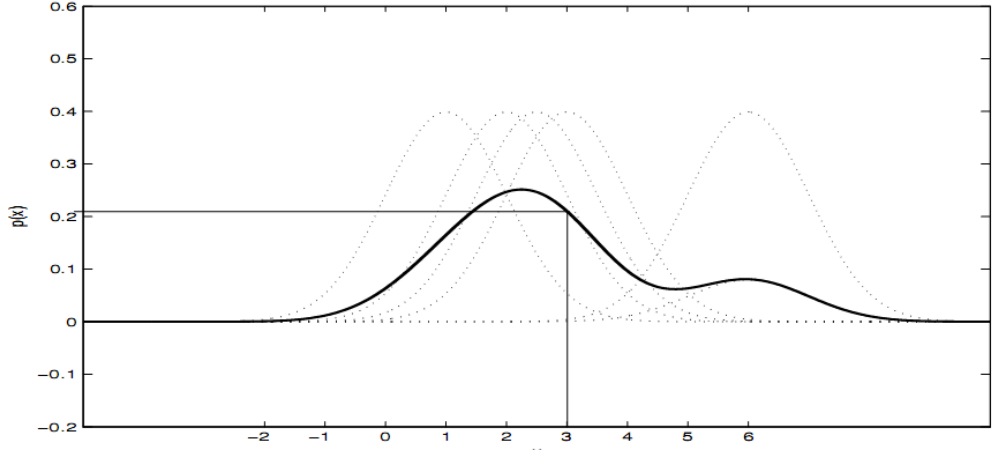
$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_5 - x)^2}{2}\right) = 0.0044$$

Παράθυρα Parzen

- Τελικά:
  - $p(x = 3) = (0.2420 + 0.3521 + 0.3989$

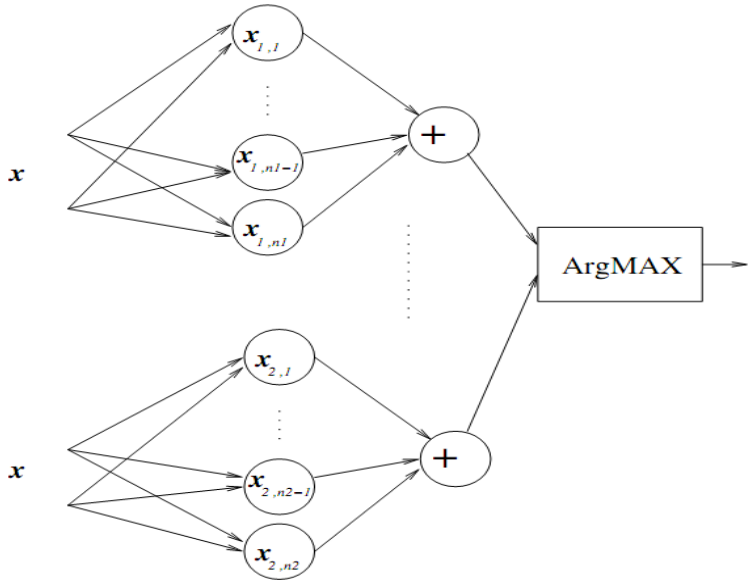
    $+0.0540 + 0.0044)/5 \quad = 0.2103$

  - Το συνολικό παράθυρο Parzen είναι:



Γενίκευση για ταξινόμηση-PNN

- PNN=Probabilistic Neural Network, επέκταση των Parzen windows για multiclass προβλήματα
- Γενική δομή:



PNN-Πράδειγμα

- Έστω δυο κλάσεις, η class1 και η class2
- Class1:
  - x1,1=2
  - x1,2=2.5
  - x1,3=3
  - x1,4=1
  - x1,5=6
- Class 2:
  - x2,1=6
  - x2,2=6.5
  - x2,3=7
- Τα παράθυρα Parzen για σ=1, για κάθε κλάση είναι:

PNN-Παράδειγμα

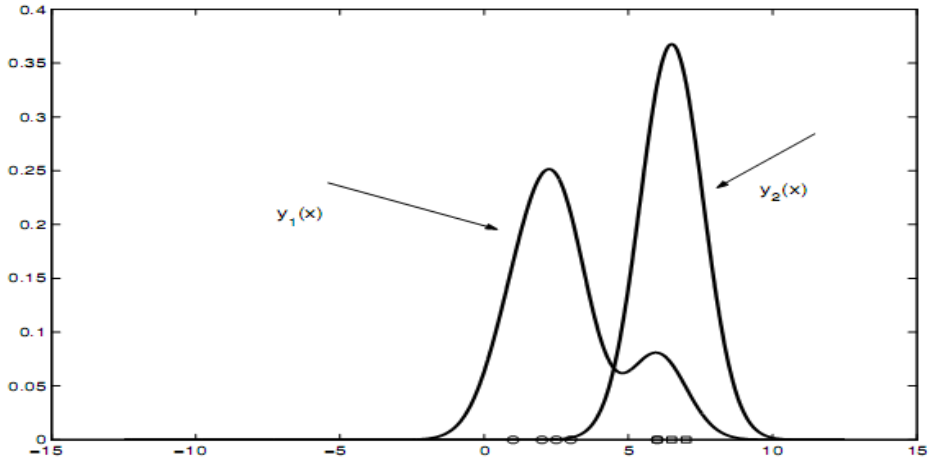
$$y_1(x) = \frac{1}{5}\sum_{i=1}^{5}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_{1,i}-x)^2}{2}\right)$$

$$y_2(x) = \frac{1}{3}\sum_{i=1}^{3}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_{2,i}-x)^2}{2}\right)$$

- Ο ταξινομητής Parzen συγκρίνει τις τιμές y1(x) και y2(x). Αν y1(x) > y2(x) τοτε το x ανήκει στην κλάση class1

PNN-παράδειγμα

- Από τις προηγούμενες διαφάνειες
  - $y_1(3) = 0.2103$



- Για την class2 έχουμε:
  - $$y_2(3) = \frac{1}{3\sqrt{2\pi}}\left\{\exp\left(-\frac{(6-3)^2}{2}\right)\right.$$
    $$+\exp\left(-\frac{(6.5-3)^2}{2}\right)$$
    $$\left.+\exp\left(-\frac{(7-3)^2}{2}\right)\right\}$$
    $$= 0.0011 < 0.2103 = y_1(x)$$

Παράδειγμα για >1D διανύσματα

- Έστω τα ακόλουθα διανύσματα για τις 2 κλάσεις:
$x_{1,1}$	1	0
$x_{1,2}$	0	1
$x_{1,3}$	1	1

    class 1

- Ένα νέο διάνυσμα x=[0.5 0.5] τι ετικέτα θα πάρει;

$x_{2,1}$	-1	0
$x_{2,2}$	0	-1

class 2

PNN-Παράδειγμα

- Για την κλάση class1

$$y_1(\mathbf{x}) = \frac{1}{3}\{\exp\left(-\frac{(1-0.5)^2+(0-0.5)^2}{2}\right) + \exp\left(-\frac{(0-0.5)^2+(1-0.5)^2}{2}\right) + \exp\left(-\frac{(1-0.5)^2+(1-0.5)^2}{2}\right)\}$$

$$= 0.7788$$

- Για την κλάση class 2

$$y_2(\mathbf{x}) = \frac{1}{2}\{\exp\left(-\frac{(-1-0.5)^2+(0-0.5)^2}{2}\right) + \exp\left(-\frac{(0-0.5)^2+(-1-0.5)^2}{2}\right)$$

$$= 0.4724$$

Άρα το x=[0.5 0.5] ταξινομείται ως class 1

Δίκτυα Ακτινικής Βάσης (RBF)

- RBF(Radial Basis Function)

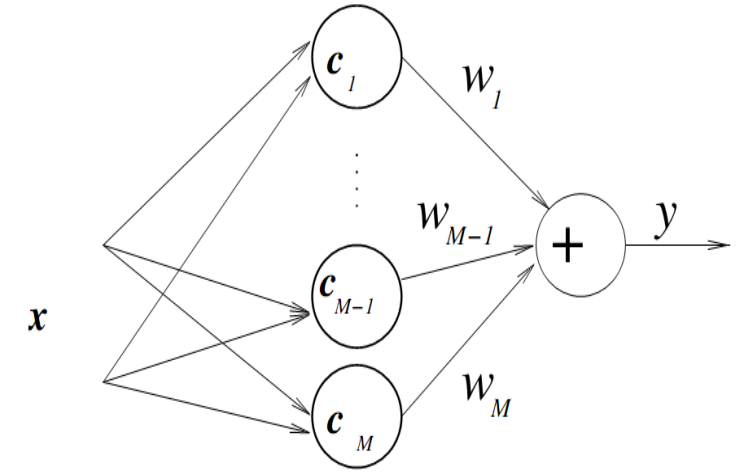
κέντρα

$$y(\mathbf{x}) = \sum_{i=1}^{M} w_i \exp\left(-\frac{(\|\mathbf{x} - \mathbf{c}_i\|)^2}{2\sigma^2}\right)$$

έξοδος

βάρη

είσοδος

- Αρχιτεκτονική:
  - 

d_m=μέγιστη απόσταση μεταξύ των κέντρων

$$\sigma_j = \frac{d_m}{\sqrt{2M}}$$

M=αριθμός κέντρων

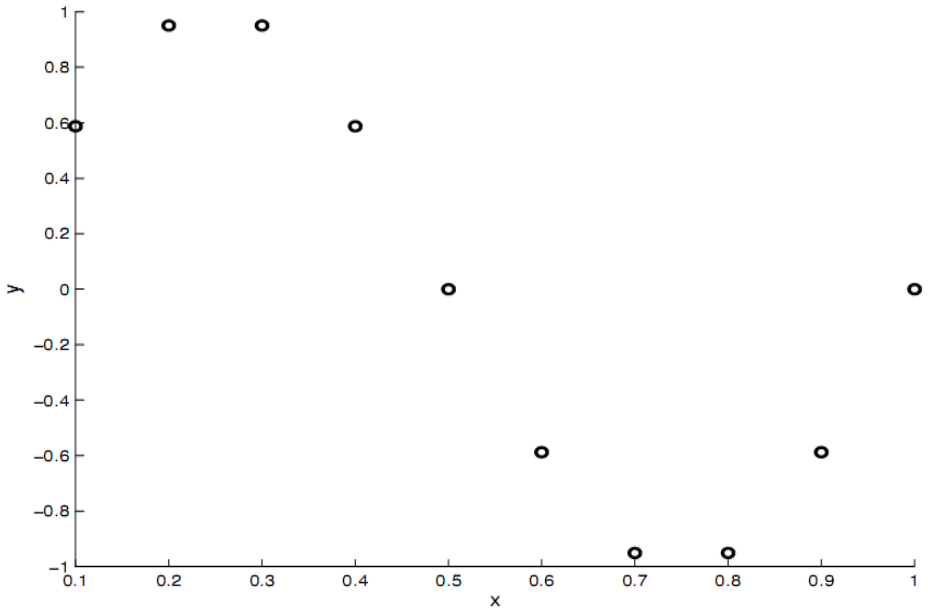
RBF-curve fitting

- Έστω ένα σύνολο 10 παραδειγμάτων, όπως φαίνεται παρακάτω. Τα παραδείγματα έχουν δημιουργηθεί με βάση τη συνάρτηση: t=sin(2πx)

i	1	2	3	4	5
x_i	0.1	0.2	0.3	0.4	0.5
t_i	0.5878	0.9511	0.9511	0.5878	0.0000

i	6	7	8	9	10
x_i	0.6	0.7	0.8	0.9	1
t_i	-0.5878	-0.9511	-0.9511	-0.5878	0.0000

RBF-Curve Fitting



- Έστω (θα δούμε αργότερα πως) ότι έχουμε 4 κέντρα ci.
  - c1=0.2 $\exp\left(-\frac{(x-0.2)^2}{2}\right)$
  - c2=0.4 $\exp\left(-\frac{(x-0.4)^2}{2}\right)$
  - c3=0.6 $\exp\left(-\frac{(x-0.6)^2}{2}\right)$
  - c4=0.8 $\exp\left(-\frac{(x-0.8)^2}{2}\right)$
- Έστω επίσης σ=1.

RBF-Curve Fitting

- Για το 10 παραδείγματα εισόδου φτιάχνουμε τον πίνακα Φ, ώστε:

$$\Phi = \begin{pmatrix} \phi_{1,1} & \phi_{1,2} & \phi_{1,3} & \phi_{1,4} \\ \phi_{2,1} & \phi_{2,2} & \phi_{2,3} & \phi_{3,4} \\ \vdots & \vdots & \dots & \vdots \\ \phi_{9,1} & \phi_{9,2} & \phi_{9,3} & \phi_{9,4} \\ \phi_{10,1} & \phi_{10,2} & \phi_{10,3} & \phi_{10,4} \end{pmatrix}$$

- με

$$\phi_{i,1} = \exp\left(-\frac{(x_i - 0.2)^2}{2}\right), \quad i = 1, 2, 3..., 10$$

$$\phi_{i,2} = \exp\left(-\frac{(x_i - 0.4)^2}{2}\right), \quad i = 1, 2, 3..., 10$$

$$\phi_{i,3} = \exp\left(-\frac{(x_i - 0.6)^2}{2}\right), \quad i = 1, 2, 3..., 10$$

$$\phi_{i,4} = \exp\left(-\frac{(x_i - 0.8)^2}{2}\right), \quad i = 1, 2, 3..., 10$$

RBF-Curve Fitting

- Με βάση την αρχιτεκτονική των RBF, προκύπτουν 10 εξισώσεις:

$$
\begin{cases}
\phi_{1,1}w_1 + \phi_{1,2}w_2 + \phi_{1,3}w_3 + \phi_{1,4}w_4 = t_1 \\
\phi_{2,1}w_1 + \phi_{2,2}w_2 + \phi_{2,3}w_3 + \phi_{2,4}w_4 = t_2 \\
\phi_{3,1}w_1 + \phi_{3,2}w_2 + \phi_{3,3}w_3 + \phi_{3,4}w_4 = t_3 \\
\qquad\qquad\qquad\qquad \cdots\cdots\cdots \\
\phi_{10,1}w_1 + \phi_{10,2}w_2 + \phi_{10,3}w_3 + \phi_{10,3}w_4 = t_{10}
\end{cases}
$$

RBF-Curve Fitting

Λύση:

$$\mathbf{w} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$$

- Οι οποίες μετατρέπονται σε:

$$\underbrace{\begin{pmatrix} \phi_{1,1} & \phi_{1,2} & \phi_{1,3} & \phi_{1,4} \\ \phi_{2,1} & \phi_{2,2} & \phi_{2,3} & \phi_{3,4} \\ \vdots & \vdots & \dots & \vdots \\ \phi_{9,1} & \phi_{9,2} & \phi_{9,3} & \phi_{9,4} \\ \phi_{10,1} & \phi_{10,2} & \phi_{10,3} & \phi_{10,4} \end{pmatrix}}_{\Phi} \underbrace{\begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix}}_{\mathbf{w}} = \underbrace{\begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_{10} \end{pmatrix}}_{\mathbf{t}}$$

RBF-Curve Fitting

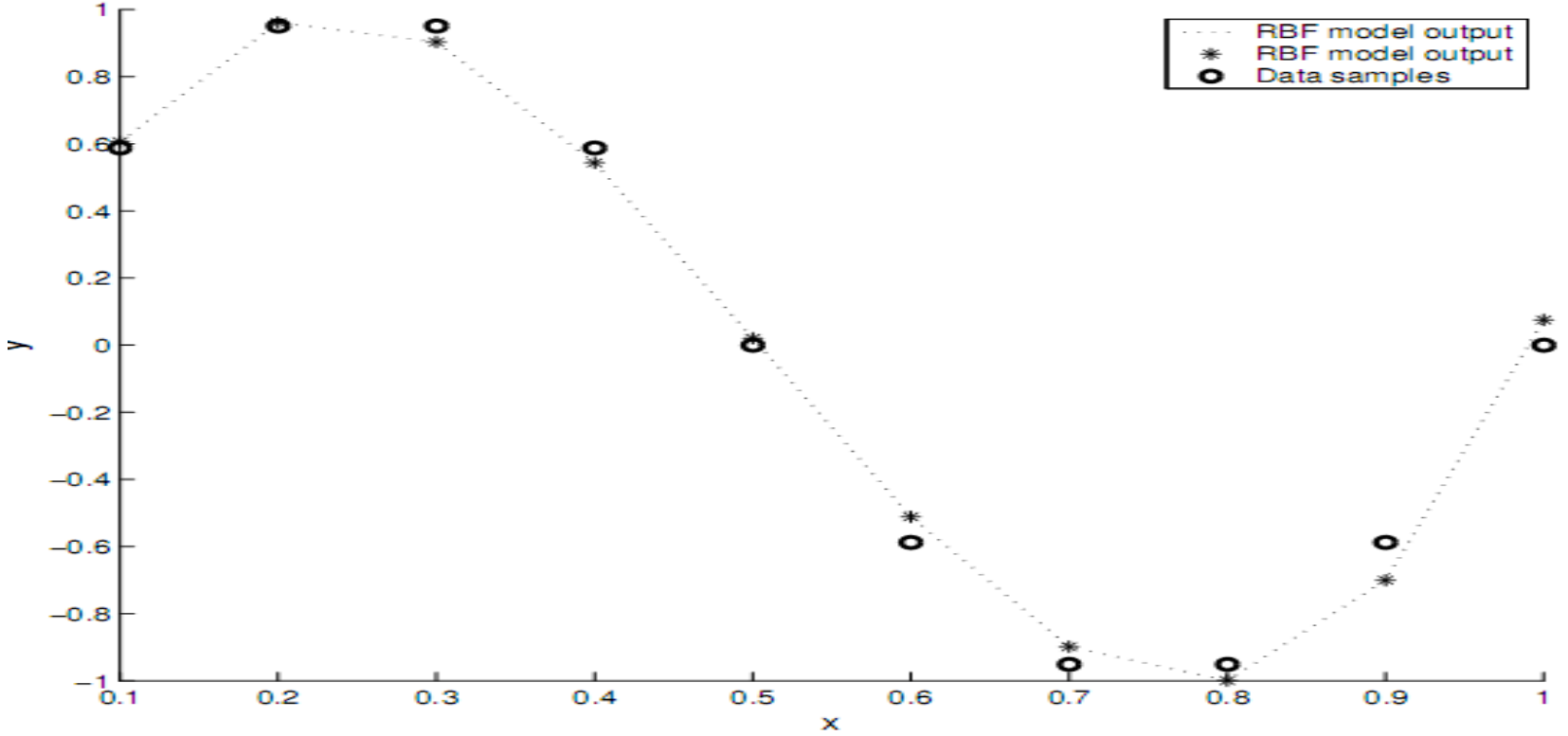
- Στο παράδειγμα μας

$$\mathrm{w} = [-3083.3, 8903.8, -8892.6, 3071.6]^T$$

- Επομένως για κάθε x ο RBF ταξινομητής προβλέπει το y ως

$$y(x) = \sum_{i=1}^{4} w_i \exp\left(-\frac{(x - c_i)^2}{2}\right)$$

RBF-Curve Fitting



RBF-Ταξινομητής

- Πολύ καλός όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα
- Π.χ. Το πρόβλημα XOR

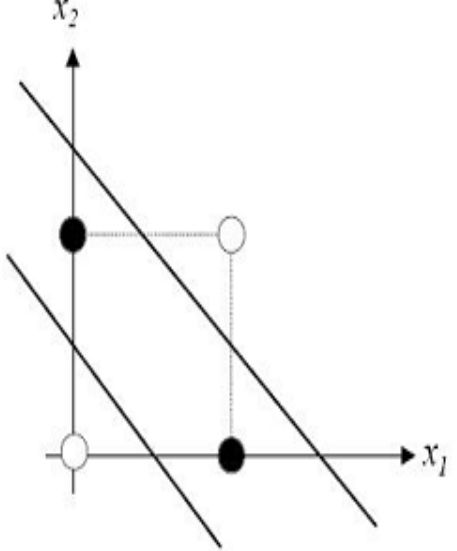
- Παραπλήσια λειτουργία με το curve fitting
  - επιλογή κέντρων
  - Εύρεση του πίνακα Φ
  - Επίλυση του συστήματος για W
  - Ταξινόμηση

XOR

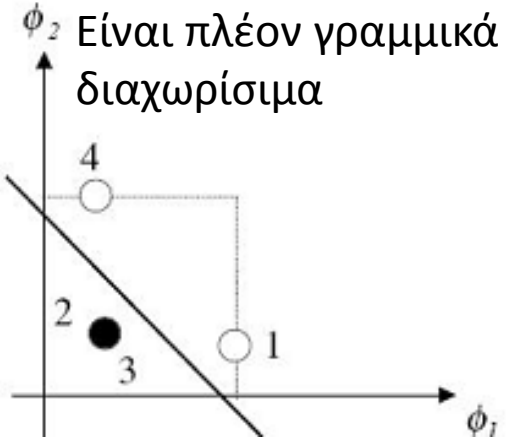
$$\phi = \exp\left(-\frac{(\|\mathbf{x}-\mathbf{c}_i\|)^2}{2\sigma^2}\right)$$

- 4 σημεία και 2 κλάσεις, επιλέγουμε
- M=2
- c1=<0,0>
- c2=<1,1>
- d_{max}=sqrt(2)
- $\sigma_j = \frac{d_m}{\sqrt{2M}}$

p	x_1	x_2	t
1	0	0	0
2	0	1	1
3	1	0	1
4	1	1	0



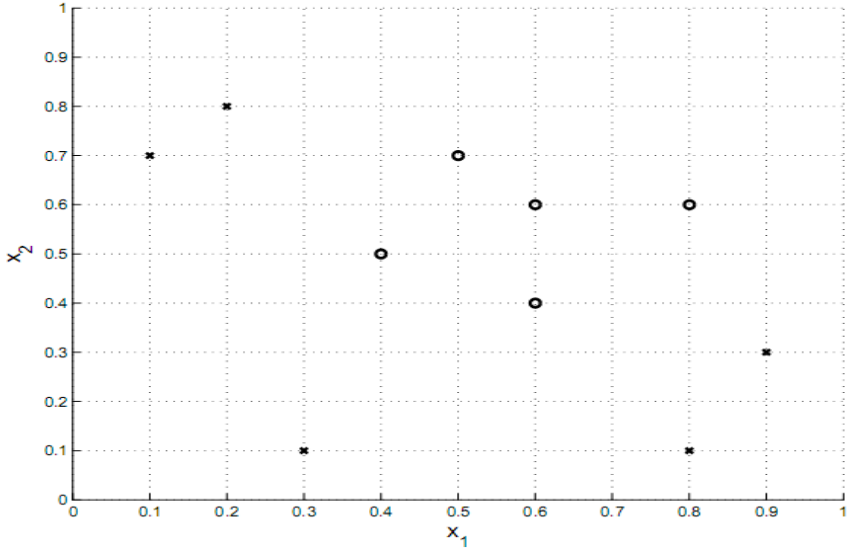
p	x_1	x_2	ϕ_1	ϕ_2
1	0	0	1.0000	0.1353
2	0	1	0.3678	0.3678
3	1	0	0.3678	0.3678
4	1	1	0.1353	1.0000



Είναι πλέον γραμμικά διαχωρίσιμα

RBF-παράδειγμα

- Έστω 10 παραδείγματα 2 διαστάσεων, 2 κλάσεων (-1,+1)



i	1	2	3	4	5
$x_{1,i}$	0.5	0.4	0.6	0.6	0.8
$x_{2,i}$	0.7	0.5	0.6	0.4	0.6
t_i	-1	-1	-1	-1	-1

i	6	7	8	9	10
$x_{1,i}$	0.2	0.1	0.9	0.8	0.3
$x_{2,i}$	0.8	0.7	0.3	0.1	0.1
t_i	1	1	1	1	1

RBF-Παράδειγμα

- Έστω 4 κέντρα
  - c1=[0.5,0.7]
  - c2=[0.6,0.4]
  - c3=[0.2,0.8]
  - c4=[0.9,0.3]

- Άρα έχουμε 4 basis functions:

$$\phi_1(\mathbf{x}) = \exp\left(-\frac{(x_1-0.5)^2+(x_2-0.7)^2}{2}\right)$$

$$\phi_2(\mathbf{x}) = \exp\left(-\frac{(x_1-0.6)^2+(x_2-0.4)^2}{2}\right)$$

$$\phi_3(\mathbf{x}) = \exp\left(-\frac{(x_1-0.2)^2+(x_2-0.8)^2}{2}\right)$$

$$\phi_4(\mathbf{x}) = \exp\left(-\frac{(x_1-0.9)^2+(x_2-0.3)^2}{2}\right)$$

RBF-Παράδειγμα

$$\phi_{i,1} = \exp\left(-\frac{(x_{1,i}-0.2)^2+(x_{2,i}-0.8)^2}{2}\right),$$

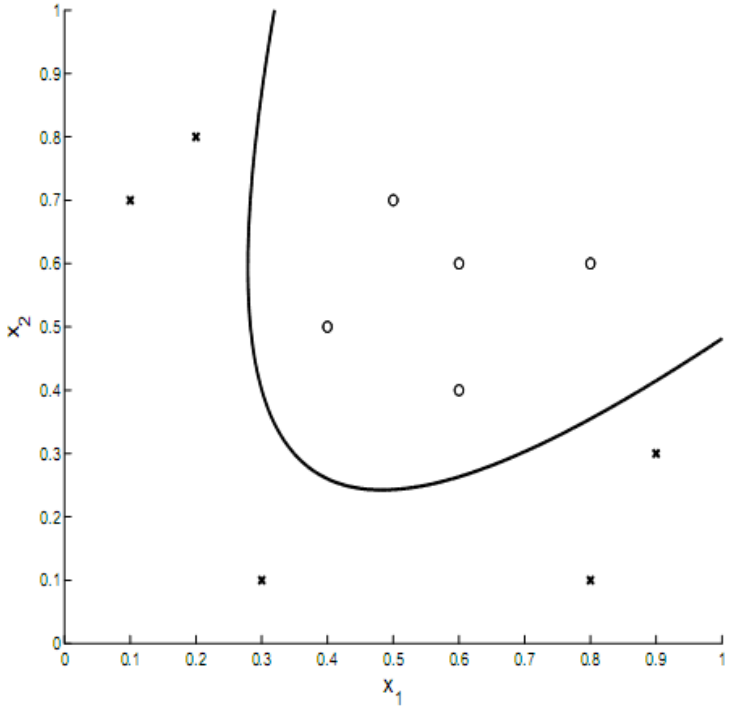
$$\phi_{i,2} = \exp\left(-\frac{(x_{1,i}-0.5)^2+(x_{2,i}-0.7)^2}{2}\right),$$

$$\phi_{i,3} = \exp\left(-\frac{(x_{1,i}-0.6)^2+(x_{2,i}-0.4)^2}{2}\right),$$

$$\phi_{i,4} = \exp\left(-\frac{(x_{1,i}-0.2)^2+(x_{2,i}-0.8)^2}{2}\right),$$

$$\mathbf{w} = [70.5912, 37.4476, -63.3062, -52.7027]^T$$

$$g(\mathbf{x}) = \sum_{i=1}^{4} w_i \phi_i(\mathbf{x})$$

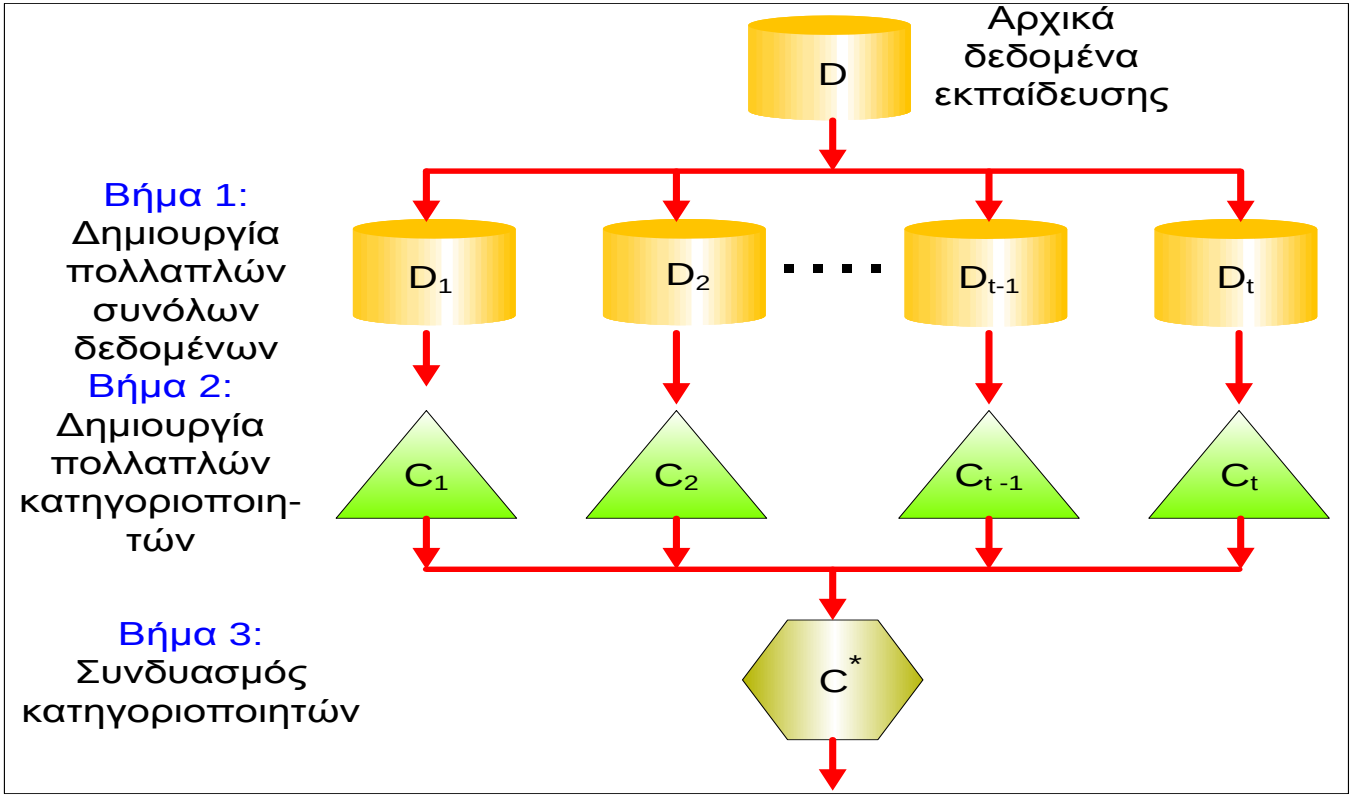


Συγκεντρωτικές μέθοδοι

- Κατασκευή ενός συνόλου κατηγοριοποιητών από τα σώματα δεδομένων

- Η πρόβλεψη της κλάσης ενός νέου παραδείγματος γίνεται αθροίζοντας τις προβλέψεις των διαφόρων κατηγοριοποιητών

Γενική ιδέα

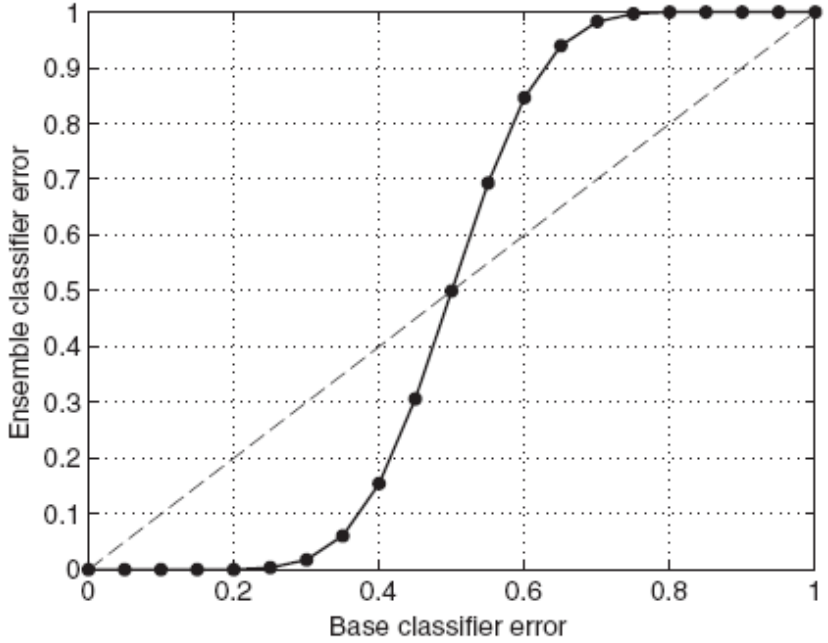

D
Αρχικά δεδομένα εκπαίδευσης
Βήμα 1: Δημιουργία πολλαπλών συνόλων δεδομένων
D1
D2
Dt-1
Dt
Βήμα 2: Δημιουργία πολλαπλών κατηγοριοποιητών
C1
C2
Ct -1
Ct
Βήμα 3: Συνδυασμός κατηγοριοποιητών
C*

Για ποιο λόγο λειτουργεί;

- Έστω ότι υπάρχουν 25 κατηγοριοποιητές
  - Κάθε ένας με σφάλμα (error rate), $\varepsilon = 0.35$
  - Υποθέτουμε ότι είναι ανεξάρτητοι
  - Πιθανότητα ενός συγκεντρωτικού κατηγοριοποιητή να κάνει εσφαλμένη πρόβλεψη:

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$$

Σύγκριση σφάλματος



Παραδείγματα

- Πως φτιάχνουμε ένα συγκεντρωτικό κατηγοριοποιητή;
- Πειράζοντας τα Δεδομένα
  - Bagging
  - Boosting
- Πειράζοντας τις Ιδιότητες
  - Random Forests

Bagging

- Δειγματοληψία με αντικατάσταση

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

- Κατασκευή ενός κατηγοριοποιητή σε κάθε δείγμα εκκίνησης (bootstrap)

- Κάθε δείγμα έχει πιθανότητα $(1 – 1/n)^n$ να επιλεχθεί

Παράδειγμα

Round	x=0.1	x=0.2	x=0.3	x=0.4	x=0.5	x=0.6	x=0.7	x=0.8	x=0.9	x=1.0
1	1	1	1	-1	-1	-1	-1	-1	-1	-1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	-1	-1	-1	-1	-1	-1	-1
4	1	1	1	-1	-1	-1	-1	-1	-1	-1
5	1	1	1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	1	1	1
7	-1	-1	-1	-1	-1	-1	-1	1	1	1
8	-1	-1	-1	-1	-1	-1	-1	1	1	1
9	-1	-1	-1	-1	-1	-1	-1	1	1	1
10	1	1	1	1	1	1	1	1	1	1
Sum	2	2	2	-6	-6	-6	-6	2	2	2
Sign	1	1	1	-1	-1	-1	-1	1	1	1
True Class	1	1	1	-1	-1	-1	-1	1	1	1

Bagging Round 1:

x	0.1	0.2	0.2	0.3	0.4	0.4	0.5	0.6	0.9	0.9
y	1	1	1	1	-1	-1	-1	-1	1	1

x <= 0.35 ==> y = 1
x > 0.35 ==> y = -1

Bagging Round 2:

x	0.1	0.2	0.3	0.4	0.5	0.8	0.9	1	1	1
y	1	1	1	-1	-1	1	1	1	1	1

x <= 0.65 ==> y = 1
x > 0.65 ==> y =

-1

Bagging Round 3:

x	0.1	0.2	0.3	0.4	0.4	0.5	0.7	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

x <= 0.35 ==> y = 1
x > 0.35 ==> y = -1

Bagging Round 4:

x	0.1	0.1	0.2	0.4	0.4	0.5	0.5	0.7	0.8	0.9
y	1	1	1	-1	-1	-1	-1	-1	1	1

x <= 0.3 ==> y = 1
x > 0.3 ==> y = -1

Bagging Round 5:

x	0.1	0.1	0.2	0.5	0.6	0.6	0.6	1	1	1
y	1	1	1	-1	-1	-1	-1	1	1	1

x <= 0.35 ==> y = 1
x > 0.35 ==> y = -1

Bagging Round 6:

x	0.2	0.4	0.5	0.6	0.7	0.7	0.7	0.8	0.9	1
y	1	-1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 7:

x	0.1	0.4	0.4	0.6	0.7	0.8	0.9	0.9	0.9	1
y	1	-1	-1	-1	-1	1	1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 8:

x	0.1	0.2	0.5	0.5	0.5	0.7	0.7	0.8	0.9	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 9:

x	0.1	0.3	0.4	0.4	0.6	0.7	0.7	0.8	1	1
y	1	1	-1	-1	-1	-1	-1	1	1	1

x <= 0.75 ==> y = -1
x > 0.75 ==> y = 1

Bagging Round 10:

x	0.1	0.1	0.1	0.1	0.3	0.3	0.8	0.8	0.9	0.9
y	1	1	1	1	1	1	1	1	1	1

x <= 0.05 ==> y = -1
x > 0.05 ==> y = 1

Boosting

- Μια επαναληπτική διαδικασία που προσαρμοστικά αλλάζει την κατανομή των δεδομένων εκπαίδευσης επικεντρώνοντας σε παραδείγματα που προηγουμένως έχουν ταξινομηθεί λανθασμένα
  - Αρχικά, αναθέτουμε ίσα βάρη σε όλες τις N εγγραφές
  - Αντίθετα με το Bagging, τα βάρη μπορεί να αλλάζουν στο τέλος κάθε γύρου

Boosting

- Οι εγγραφές που κατηγοριοποιούνται λανθασμένα **αυξάνουν** τα βάρη τους
- Οι εγγραφές που κατηγοριοποιούνται ορθά **μειώνουν** τα βάρη τους

Original Data	1	2	3	4	5	6	7	8	9	10
Boosting (Round 1)	7	3	2	8	7	9	4	10	6	3
Boosting (Round 2)	5	4	9	4	2	5	1	7	4	2
Boosting (Round 3)	4	4	8	10	4	5	4	6	3	4

- Το παράδειγμα 4 είναι δύσκολο να κατηγοριοποιηθεί
- Το βάρος του αυξάνεται, επομένως είναι πιο πιθανό να επιλεγεί στον επόμενο γύρο

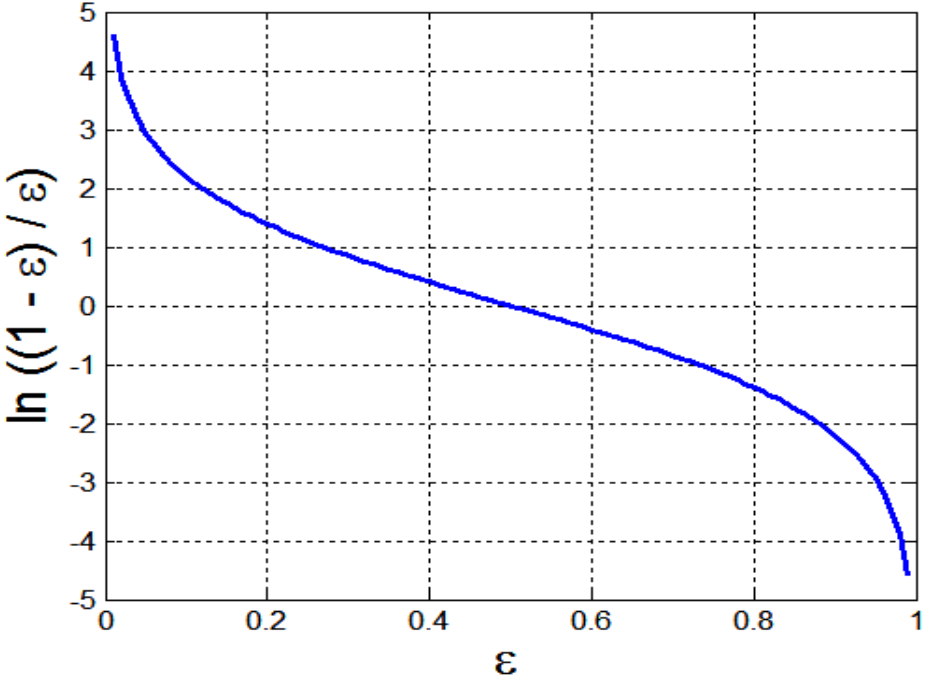
Παράδειγμα: Adaboost

- Κατηγοριοποιητές βάσης: C_1, C_2, …, C_T
- Error rate:

$$\varepsilon_i = \frac{1}{N}\sum_{j=1}^{N} w_j \delta\left(C_i(x_j) \neq y_j\right)$$

- Σημαντικότητα ενός κατηγοριοποιητή:

$$\alpha_i = \frac{1}{2}\ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$$



Παράδειγμα: Adaboost

- Αναβάθμιση βαρών:

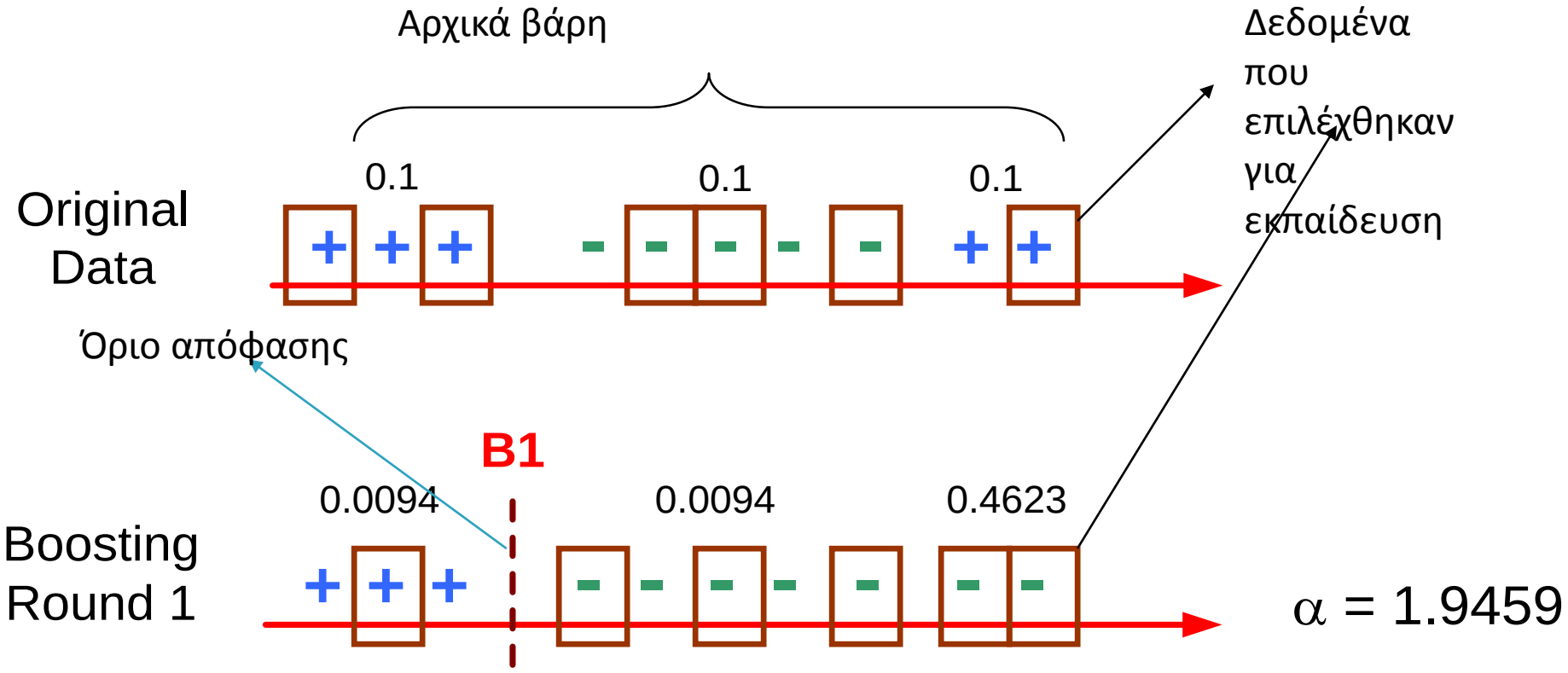
$$w_i^{(j+1)} = \frac{w_i^{(j)}}{Z_j} \begin{cases} \exp^{-\alpha_j} & \text{αν } C_j(x_i) = y_i \\ \exp^{\alpha_j} & \text{αν } C_j(x_i) \neq y_i \end{cases}$$

όπου Z_j παράγοντας κανονικοποίησης

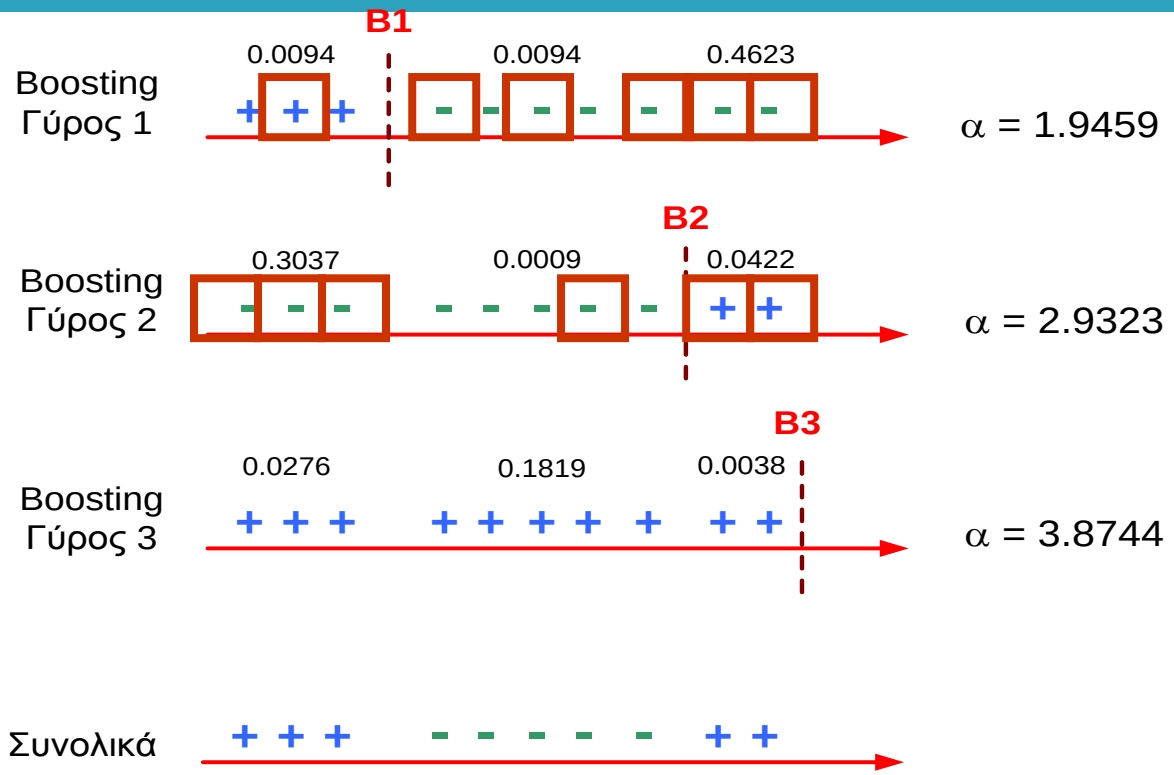
- Αν κάποιος ενδιάμεσος γύρος παράγει σφάλμα μεγαλύτερο του 50%, τα βάρη ξαναγίνονται ίσα με 1/n και επαναλαμβάνεται η δειγματοληψία
- Κατηγοριοποίηση:

$$C^*(x) = \underset{y}{\operatorname{argmax}} \sum_{j=1}^{T} \alpha_j \delta\left(C_j(x) = y\right)$$

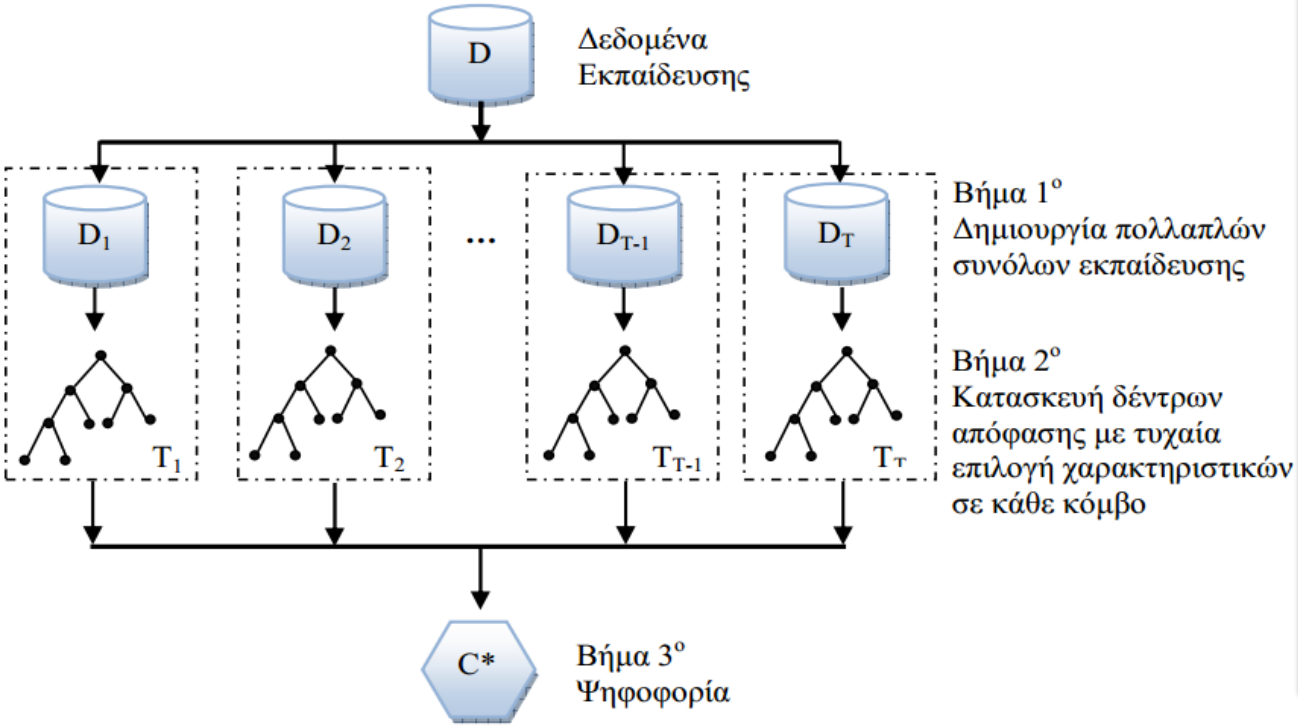
AdaBoost


Αρχικά βάρη
Δεδομένα που επιλέχθηκαν για εκπαίδευση
Original Data
0.1
0.1
0.1
Όριο απόφασης
B1
Boosting Round 1
0.0094
0.0094
0.4623
α = 1.9459

AdaBoost


B1
0.0094
0.0094
0.4623
Boosting
Γύρος 1
α = 1.9459
B2
0.3037
0.0009
0.0422
Boosting
Γύρος 2
α = 2.9323
B3
0.0276
0.1819
0.0038
Boosting
Γύρος 3
α = 3.8744
Συνολικά

Random Forests



- Μπορούν να εκπαιδευτούν σε σύνολα δεδομένων υψηλής διάστασης όπως είναι τα κείμενα και οι εικόνες, χωρίς να εμφανίσουν σημαντικό βαθμό overfitting
- Παρουσιάζουν ανεκτικότητα ως προς το όρυβο και αριθμητικών σφαλμάτων στα δεδομένα εκπαίδευσης (π.χ. απόκρυψη μέρους του αντικειμένου, ελλιπή δεδομένα).
- Για την επαγωγή κάθε δέντρου περίπου το 1/3 των παραδειγμάτων δεν επιλέγεται για εκπαίδευση. Αυτά τα παραδείγματα καλούνται Out-of-Bag παραδείγματα και μπορούν να χρησιμοποιηθούν για την εκτίμηση της πιθανότητας σφάλματος, εξαλείφοντας την ανάγκη ύπαρξης ενός συνόλου ελέγχου ή εφαρμογής της τεχνικής cross-validation

Γενετικοί αλγόριθμοι

- Θεωρία της εξέλιξης (C. Darwin, 1858)
- Εξέλιξη = Διαδικασία που οδηγεί στην αύξηση της ικανότητας ενός πληθυσμού να επιβιώνει και να αναπαράγεται σε ένα δεδομένο περιβάλλον (Εξελικτική προσαρμογή)

- Τοπίο προσαρμογής (S. Wright, 1932)
- Οι κορυφές αντιστοιχούν στη βέλτιστη προσαρμογή των ειδών

- Προσομοίωση της διαδικασίας εξέλιξης

Ορολογία των Γενετικών Αλγορίθμων

- Δανεισμένη από το χώρο της φυσικής Γενετικής.
- Αναφέρονται σε ***άτομα*** ή ***γενότυπα*** μέσα σε ένα πληθυσμό. Πολύ συχνά αυτά τα άτομα καλούνται επίσης ***χρωμοσώματα***.
- Τα **χρωμοσώματα** αποτελούνται από διάφορα στοιχεία που ονομάζονται ***γονίδια.***
- Κάθε γονίδιο επηρεάζει την κληρονομικότητα ενός ή περισσότερων χαρακτηριστικών.

Πως δουλεύουν;

1. Διατηρούν έναν πληθυσμό κωδικοποιημένων πιθανών λύσεων
2. Εξελίσσουν τον πληθυσμό εφαρμόζοντας σε αυτόν διάφορες γενετικές διαδικασίες:
   - Διαδικασίες επιλογής,
   - Διαδικασίες αναπαραγωγής,
   - Διαδικασίες μετάλλαξης.
3. Δημιουργούν νέο πληθυσμό που αντικαθιστά τον προηγούμενο.
4. Επαναλαμβάνουν τη διαδικασία έως ότου «βρουν λύση».

Πως Δουλεύουν;
Γενετικοί Τελεστές

- **Επιλογή**: επιλέγει με κάποιο τρόπο τα «καταλληλότερα» μέλη του πληθυσμού και τα «περνάει» στο νέο πληθυσμό.
- **Διασταύρωση**: συνδυάζει τα στοιχεία δύο χρωμοσωμάτων γονέων για να δημιουργήσει δύο νέους απογόνους ανταλλάσσοντας αντίστοιχα κομμάτια από τους γονείς.
- **Μετάλλαξη**: αλλάζει αυθαίρετα ένα ή περισσότερα γονίδια ενός συγκεκριμένου χρωμοσώματος.

Πως Δουλεύουν;

- Ένας Γ.Α. πρέπει να αποτελείται από τα παρακάτω πέντε τμήματα:
  - **Γενετική αναπαράσταση**
  - **Τρόπο δημιουργίας ενός αρχικού πληθυσμού**
  - **Αντικειμενική συνάρτηση αξιολόγησης**
  - **Γενετικούς τελεστές**
  - **Τιμές για τις διάφορες παραμέτρους**

Ένας απλός Γενετικός Αλγόριθμος

1. Κωδικοποίηση (Coding)
2. Αρχικοποίηση (Initialization)
3. Αποκωδικοποίηση (Decoding)
4. Υπολογισμός ικανότητας ή αξιολόγηση (Fitness calculation ή evaluation)
5. Επιλογή (Selection)
6. Αναπαραγωγή (Reproduction)
7. Διασταύρωση (Crossover ή mating)
8. Μετάλλαξη (Mutation)
9. Επανάληψη από το βήμα (2) μέχρι να ικανοποιηθεί το κριτήριο τερματισμού του Γ.Α.

Παράδειγμα

Εύρεση μεγίστου της

$F(x)=x^2$

όπου x είναι ακέραιος στο διάστημα [1, 31].

Παράδειγμα
Κωδικοποίηση

- Η κωδικοποίηση είναι προφανής:
  - Θέλουμε να αναπαραστήσουμε 31 αριθμούς οπότε θα χρησιμοποιήσουμε χρωμοσώματα των 5 γονιδίων (συμβολοσειρές των 5bits) $2^5=32>31$.

Παράδειγμα

Αρχικοποίηση

- Δημιουργία αρχικού πληθυσμού (έστω μεγέθους 4) με τυχαίο τρόπο:

$A_1 = 0\ 1\ 1\ 0\ 1 = 13_{10}$

$A_2 = 1\ 1\ 0\ 0\ 0 = 24_{10}$

$A_3 = 0\ 1\ 0\ 0\ 0 = 8_{10}$

$A_4 = 1\ 0\ 0\ 1\ 1 = 19_{10}$

Παράδειγμα

Αξιολόγηση

$F(A_1) = 13^2 = 169$

$F(A_2) = 24^2 = 576$

$F(A_3) = 8^2 = 64$

$F(A_4) = 19^2 = 361$

Συνολική Απόδοση: 1170

Μέση απόδοση: 293

Παράδειγμα
Επιλογή

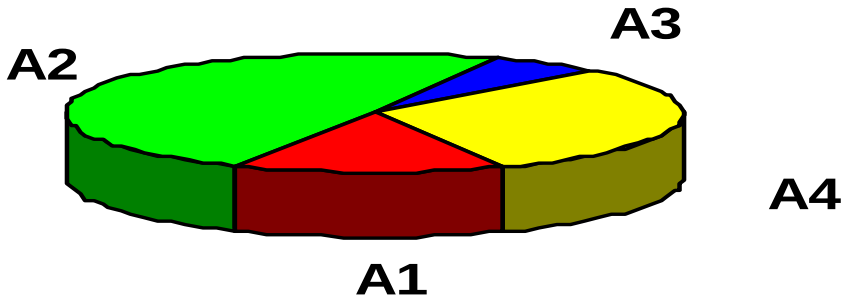
Επιλογή των ατόμων του πληθυσμού που θα «περάσουν» στον επόμενο πληθυσμό. Αυτό μπορεί να γίνει με διάφορους τρόπους όπως για παράδειγμα με τη χρήση μιας εξαναγκασμένης ρουλέτας

Στην εξαναγκασμένη ρουλέτα κάθε μέλος του πληθυσμού έχει πιθανότητα επιλογής ίση με τη σχετική του απόδοση στον τρέχοντα πληθυσμό.

$P(A_1)$ = 0.14=169/1170

$P(A_2)$ = 0.49=…/1170

$P(A_3)$ = 0.06=…/1170



Παράδειγμα

Αναπαραγωγή

- Ο προσωρινός πληθυσμός μετά την εφαρμογή της εξαναγκασμένης ρουλέτας:

A'_1 = 0 1 1 0 1

A'_2 = 1 1 0 0 0

A'_3 = 1 1 0 0 0

A'_4 = 1 0 0 1 1

Παράδειγμα
Διασταύρωση

Επιλογή με τυχαίο τρόπο των ατόμων που θα διασταυρώσουν το γενετικό υλικό τους:

Έστω ότι διασταυρώνονται το A'_1 με το A'_2 με σημείο διασταύρωσης το 4 και το A'_3 με το A'_4 με σημείο διασταύρωσης το 2:

$A'_1 = 0\,1\,1\,0 \mid 1$
$A'_2 = 1\,1\,0\,0 \mid 0$

$A''_1 = 0\,1\,1\,0 \mid 0$
$A''_2 = 1\,1\,0\,0 \mid 1$

$A'_3 = 1\,1 \mid 0\,0\,0$
$A'_4 = 1\,0 \mid 0\,1\,1$

$A''_3 = 1\,1 \mid 0\,1\,1$
$A''_4 = 1\,0 \mid 0\,0\,0$

Παράδειγμα
Μετάλλαξη

Με τυχαίο τρόπο επιλέγονται γονίδια των οποίων η τιμή αντιστρέφεται:

$A''_1 = 0\,1\,1\,0\,0$

$A''_2 = 1\,1\,0\,0\,1$

$A''_3 = 1\,1\,0\,1\,1$

$A''_4 = 1\,0\,0\,0\,0$

$A'''_1 = 0\,1\,1\,0\,0$

$A'''_2 = 1\,1\,0\,0\,1$

$A'''_3 = 1\,1\,0\,1\,1$

$A'''_4 = 1\,0\,0\,1\,0$

Παράδειγμα
Νέος Πληθυσμός

Ο νέος πληθυσμός που προκύπτει είναι:

$\mathbf{A_1 = 0\ 1\ 1\ 0\ 0} = 12_{10}$ =>F(12)=144

$\mathbf{A_2 = 1\ 1\ 0\ 0\ 1} = 25_{10}$ =>F(25)=625

$\mathbf{A_3 = 1\ 1\ 0\ 1\ 1} = 27_{10}$ =>F(27)=729

$\mathbf{A_4 = 1\ 0\ 0\ 1\ 0} = 18_{10}$ =>F(18)=324

Συνολική Απόδοση: 1822

Μέση απόδοση:455.5

Βασικά Χαρακτηριστικά γενετικών

- Δουλεύουν με μια κωδικοποίηση του συνόλου τιμών που μπορούν να λάβουν οι μεταβλητές και όχι με τις ίδιες τις μεταβλητές του προβλήματος
- Κάνουν αναζήτηση σε πολλά σημεία ταυτόχρονα και όχι μόνο σε ένα
- Χρησιμοποιούν μόνο την αντικειμενική συνάρτηση και καμία επιπρόσθετη πληροφορία
- Χρησιμοποιούν πιθανοθεωρητικούς κανόνες μετάβασης και όχι ντετερμινιστικούς