ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΑΝΟΙΚΤΑ ακαδημαϊκά μαθήματα ΠΠ

Εισαγωγή στη Βιοπληροφορική

Ενότητα 5: Αλγόριθμοι Συσταδοποίησης και Κατηγοριοποίησης Βιολογικών Δεδομένων

Μακρής Χρήστος, Τσακαλίδης Αθανάσιος, Περδικούρη Αικατερίνη

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Η/Υ και Πληροφορικής

Σκοποί ενότητας

- Σκοπός της ενότητας είναι η παρουσίαση των αλγορίθμων συσταδοποίησης και κατηγοριοποίησης βιολογικών δεδομένων

Βασική Βιβλιογραφική Πηγή στην οποία βασίζονται οι διαφάνειες

- DATA MINING, Margaret H. Dunham. ΕΚΔΟΣΕΙΣ ΝΕΩΝ ΤΕΧΝΟΛΟΓΙΩΝ ΙΔΙΩΤΙΚΗ ΚΕΦΑΛΑΙΟΥΧΙΚΗ ΕΤΑΙΡΕΙΑ Έκδοση: 1η/2004
- Μ. Χαλκίδη, Μ. Βαζιργιάννης, Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό (Εκδόσεις: Τυπωθήτω), 2005

Περιεχόμενα ενότητας

- Ορισμός προβλήματος
- Μετρικές ομοιότητας
- Κατηγορίες αλγορίθμων κατηγοριοποίησης
- Εφαρμογές σε προβλήματα μοριακής βιολογίας

Αλγόριθμοι Συσταδοποίησης και Κατηγοριοποίησης Βιολογικών Δεδομένων

Κατηγοριοποίηση Δεδομένων: Ορισμός προβλήματος

- Έστω $X=\{x_1,\ldots,x_n\}$ το σύνολο των δεδομένων όπου $x_i=(x_{i1},\ldots,x_{id})$ ένα διάνυσμα μήκους *d*
- Στόχος της διαδικασίας κατηγοριοποίησης είναι να αντιστοιχίσει το σύνολο των δεδομένων σε ένα πεπερασμένο σύνολο *k* ομάδων, όχι απαραίτητα αμοιβαίως αποκλειόμενες:

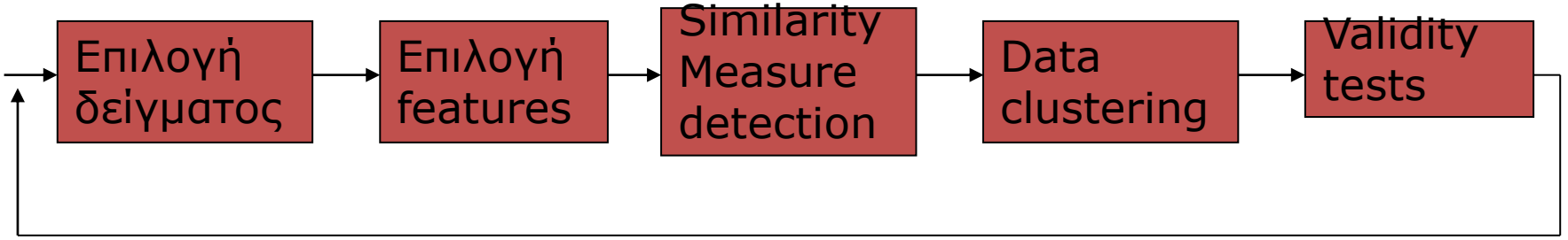
  -- ομάδες γνωστές => επιβλεπόμενη μάθηση (κατηγοριοποίηση)

  -- ομάδές μη γνωστές => μη επιβλεπόμενη μάθηση (clustering, ομαδοποίηση, συσταδοποίηση)

Εφαρμογές

- Μείωση Δεδομένων
- Παραγωγή/’Ελεγχος υπόθεσης
- Πρόβλεψη.
- Summarization
- Compression
- Efficiently finding nearest neighbours

Φάσεις Συσταδοποίησης



Κριτήρια καταλληλότητας μετρικών ομοιότητας

- Συμμετρία: *d(x,y)=d(y,x)>=0*
- Ανισότητα τριγώνου: *d(x,y)<=d(x,z)+d(y,z)*
- Διαφοροποίηση των ανόμοιων σημείων: d(x,y)≠0, x≠y
- Ταυτοποίηση των όμοιων σημείων: d(x,x’)=0, if x=x’.

Μετρικές ομοιότητας

- Συνιστώσες συσχέτισης: (απόσταση αντικειμένων j και k)

$$r_{jk} = \frac{\sum (x_{ij} - x'_j)(x_{ik} - x'_k)}{\sqrt{\sum (x_{ij} - x'_j)^2 (x_{ik} - x'_k)^2}}$$

(χρησιμοποιείται κυρίως για διακριτές μεταβλητές, όπου x'_j και x'_k οι μέσες τιμές των j και k)

- Μετρικές απόστασης:

$$d_{ij} = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}$$

πιο γενικά η L_p μετρική:

$$d_{ij} = \left(\sum_{k=1}^{d} | x_{ik} - x_{jk} |^p\right)^{1/p}$$

συνήθως απαραίτητη η κανονικοποίηση $x'_{ik}=(x_{ik}-x_k)/s_k$
(x_k μέση τιμή, s_k απόκλιση)

- Συνιστώσες σχέσης: χρησιμοποιείται με δυαδικές μεταβλητές: συνιστώσα Jaccard(A,B)=|A∩B|/|A∪B|

- Πιθανοτικές συνιστώσες ομοιότητας

Μέθοδοι Συσταδοποίησης

Οι μέθοδοι μπορούν να κατηγοριοποιηθούν με βάση:

- Τον τύπο δεδομένων που εισάγονται στον αλγόριθμο.
- Τη μέθοδο που καθορίζει την συσταδοποίηση του συνόλου των δεδομένων.
- Τη θεωρία και τις θεμελιώδεις έννοιες στις οποίες είναι βασισμένες οι τεχνικές ανάλυσης συστάδας.

Κατηγοριοποίηση με βάση τύπο δεδομένων

- Συσταδοποίηση αριθμητικών δεδομένων (πιο γενικά χρήση L_p μετρικής

$$\text{Ευκλείδια Απόσταση} = \sqrt{\sum (x_i - y_i)^2}$$

$$\text{City-block Απόσταση} = \sum |(x_i - y_i)|$$

- Κατηγορική Συσταδοποίηση

- Κειμενική Συσταδοποίηση

Κατηγοριοποίηση με βάση Μέδοδο Συσταδοποίησης

- Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)
- Συσταδοποίηση Διαμέρισης (Partitioning Clustering)
- Ασαφής συσταδοποίηση (Fuzzy Clustering)

- Συσταδοποίηση βασισμένη στα δίκτυα Kohonen (Kohonen Net Clustering)
- Συσταδοποίηση βασισμένη στην πυκνότητα (Density-based Clustering)
- Συσταδοποίηση βασισμένη σε πλέγμα (Grid-based Clustering)
- Συσταδοποίηση υποχώρων (Subspace Clustering).

Παράμετροι Ομαδοποίησης

$$C = \frac{\sum_{i=1}^{N} t_i}{N} \qquad R = \sqrt{\frac{\sum_{i=1}^{N} (t_i - C)^2}{N}}$$

$$D = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} (t_i - t_i)}{N(N-1)}}$$

Ιεραρχικοί Αλγόριθμοι

- Συσσωρευτικοί (Agglomerative)
- Διαιρετικοί (Divisive)

Ιεραρχικοί Αλγόριθμοι Ομαδοποίησης(2)

- Κριτήρια σύνδεσης:
  - –Μονή σύνδεση(single linkage): μικρότερη απόσταση ανάμεσα σε ζεύγη συστάδων.
  - –Μέση σύνδεση(average linkage): μέση απόσταση ανάμεσα σε ζεύγη συστάδων.
  - –Πλήρης σύνδεση(complete linkage): μέγιστη απόσταση ανάμεσα σε ζεύγη συστάδων.
  - –Απόσταση μεταξύ centroids

  Ο single link αλγόριθμος έχει ως μειονέκτημα το καλούμενο chaining effect.

Διαμεριστικοί Αλγόριθμοι

- Δημιουργούν ομάδες σε ένα βήμα
- Απαιτείται (κατά κανόνα) γνώση του μεγέθους της ομάδας
- Συνήθως διαχειρίζεται στατικά σύνολα
- Μερικοί ιεραρχικοί αλγόριθμοι μπορούν να μετατραπούν σε διαμεριστικούς, παράδειγμα ο MST

Αλγόριθμος Τετραγωνικού σφάλματος

- Αρχικό σύνολο από clusters centers επιλεγμένο τυχαία.
- Ανάθεσε τα αντικείμενα στα πιο κοντινά cluster centers
- Επαναϋπολόγισε το κέντρο του κάθε cluster
- Υπολόγισε το τετραγωνικό σφάλμα:

$$s_i = \sum_{j=1}^{m} \left\| t_{ji} - C_i \right\|^2 \qquad S = \sum_{i=1}^{k} s_i$$

Επανέλαβε την ανωτέρω διαδικασία μέχρις ότου η διαφορά μεταξύ δύο διαδοχικών τετραγωνικών σφαλμάτων, να είναι πιο μικρή από ένα όριο.

k-Means Αλγόριθμος

K-Means

- Αρχικό σύνολο από clusters centers επιλεγμένο τυχαία.
- Ανάθεσε τα αντικείμενα στα πιο κοντινά cluster centers
- Επαναϋπολόγισε το κέντρο του κάθε cluster
- Επαναληπτικά, αντικείμενα μετακινούνται ανάμεσα σε σύνολο από clusters έως ότου εντοπιστεί το επιθυμητό σύνολο.
- Ουσιαστικά ο αλγόριθμος επιχειρεί να ελαχιστοποιήσει τη μέση τετραγωνική απόσταση των δεδομένων από τα πλησιέστερα κέντρα των συστάδων και δίνεται από τον τύπο (άρα μπορεί να θεωρηθεί αλγόριθμος τετραγωνικού σφάλματος, αν και τα κριτήρια σύγκλισης ποικίλλουν)

Ο Bisecting K-Means Αλγόριθμος

1. Initialize the list to contain the cluster with all points
2. **Repeat**
3. Remove a cluster from the list of clusters
4. {perform several “trial” bissections}
5. **for** i =1 to number of trials **do**
6. bisect the selected cluster using basic k-means
7. **end for**
8. Select the two clusters with the lowest total SSE
9. Add these clusters to the list of clusters.
10. Until the list of clusters contains K clusters

PAM (Partitioning Around Medoids)

- ***Partitioning Around Medoids (PAM) (K-Medoids)***
- Αντιμετωπίζει ικανοποιητικά τους outliers.
- Η διάταξη εισόδου δεν επηρεάζει τα αποτελέσματα.
- Δεν κλιμακώνεται ικανοποιητικά.
- Κάθε cluster αντιπροσωπεύεται με ένα μόνο αντικείμενο που καλείται ***medoid.***
- Το αρχικό σύνολο k medoids επιλέγεται τυχαία.

PAM

1. Τυχαία επιλογή Κ αντιπροσώπων για τις συστάδες.

2. Υπολογισμός του συνολικού κόστους $\mathbf{TC_{ih}}$ για όλα τα ζεύγη των αντικειμένων O_j, O_h όπου το O_i είναι το τρέχον επιλεγμένο αντικείμενο και το O_h είναι ένα μη επιλεγμένο αντικείμενο.

3. Επιλέγουμε το ζεύγος O_i, O_h το οποίο αντιστοιχεί στο $\min_{Oi,\ Oh} TC_{ih}$. Εάν το συνολικό κόστος είναι αρνητικό αντικαθιστούμε το O_i με το O_h και επιστρέφουμε στο βήμα 2.

4. Διαφορετικά, για κάθε μη επιλεγμένο αντικείμενο, βρίσκουμε το αντικείμενο αντιπρόσωπο που προσεγγίζει περισσότερο. Τότε ο αλγόριθμος σταματά.

Αλγόριθμος CLARA (Clustering LARge Applications)

1. Για i = 1... 5, επαναλαμβάνουμε τα ακόλουθα βήματα:

2. Επιλέγουμε ένα δείγμα 40 + 2k αντικειμένων με τυχαίο τρόπο από το σύνολο των δεδομένων και καλούμε τον αλγόριθμο PAM για να βρούμε τους k αντιπροσώπους για τις συστάδες.

3. Για κάθε αντικείμενο Oj στο σύνολο δεδομένων, καθορίζουμε πιο από τα k medoids προσεγγίζει περισσότερο το Oj.

4. Υπολογίζουμε την συνολική ανομοιότητα για την συσταδοποίηση που λαμβάνεται από το προηγούμενο βήμα. Εάν αυτή η τιμή είναι μικρότερη από το τρέχον ελάχιστο, χρησιμοποιούμε αυτή την τιμή του ελαχίστου σαν τρέχον ελάχιστο και διατηρούμε τα k medoids που βρήκαμε στο βήμα 2 σαν το καλύτερο σύνολο των medoids που έχουμε μέχρι στιγμής.

5. Επιστρέφουμε στο βήμα 1 και ξεκινάμε με την επόμενη επανάληψη.

Αλγόριθμος CLARANS (Clustering Large Applications based on Randomized Search)

1. Αρχικοποίηση των παραμέτρων numlocal (αριθμός τοπικών *βέλτιστων που θα αναζητηθούν*) και *maxneighbor (μέγιστος αριθμός γειτόνων που μπορούν να εξεταστούν*). Αρχικοποιούμε το i σε 1 και θέτουμε ως ελάχιστο κόστος mincost έναν μεγάλο αριθμό.

2. Καθορισμός της μεταβλητής current (τρέχον κόμβος προς *εξέταση*) ώστε να αναφέρεται σε έναν αρχικό κόμβο $G_{n,k}$.

3. Θέτουμε το j ίσο με 1.

4. Θεωρούμε έναν τυχαίο γείτονα S του τρέχοντος και υπολογίζουμε το κόστος αντικατάστασης του τρέχοντος κόμβου από τον γειτονικό κόμβο.

5. Εάν ο S έχει μικρότερο κόστος, θέτουμε ως τρέχον κόμβο (current) τον S και επιστρέφουμε στο βήμα 3.

6. Διαφορετικά, αυξάνουμε το j κατά 1. Εάν $j \leq$ maxneighbor, επιστρέφουμε στο *βήμα 4*.

7. Διαφορετικά, όταν το *j > maxneighbor,* συγκρίνουμε το κόστος του τρέχοντος κόμβου *current* με το ελάχιστο κόστος *mincost*. Εάν το πρώτο είναι μικρότερο από το mincost, θέτουμε ως mincost το κόστος του current και ορίζουμε ως καλύτερο κόμβο (*bestnode*) τον current.

8. Αυξάνουμε το ***i*** κατά 1. Εάν *i > numlocal,* εξάγουμε τον καλύτερο κόμβο και η διαδικασία σταματά. Διαφορετικά, επιστρέφουμε στο βήμα 2.

Πιθανοτικές Μέθοδοι Ομαδοποίησης

Τα δεδομένα είναι ένα δείγμα από ένα μεικτό μοντέλο διαφορετικών κατανομών.

Τα δεδομένα παράγονται:

1. επιλέγοντας τυχαία ένα μοντέλο j με πιθανότητα τ_j, $j=1:k$
2. Σημειώνοντας ένα στοιχείο από την αντίστοιχη κατανομή.

Η περιοχή γύρω από τον μέσο όρο κάθε κατανομής αποτελεί μία φυσική συστάδα με γνωστή μέση τιμή και διασπορά. Θεωρώντας ότι κάθε στοιχείο ανήκει σε μία μόνο συστάδα, υπολογίζουμε το $P(x \text{ ανήκει σε } C_j)$

Ομαδοποίηση βάση πυκνότητας

Οι βασικές παράμετροι είναι:

(α) η ύπαρξη ενός στοιχείου πυρήνα

(β) η απόσταση ενός στοιχείου από ένα στοιχείο πυρήνα

(γ) το πλήθος των στοιχείων που είναι κοντά σε ένα στοιχείο πυρήνα, η γειτνίαση ενός στοιχείου x: ε-neighborhood, δηλαδή το σύνολο των στοιχείων με απόσταση μικρότερη του ε

(δ) η συνεκτικότητα δύο στοιχείων που μπορούν να προσεγγιστούν από ένα κοινό στοιχείο πυρήνα (κλασσικό παράδειγμα αλγορίθμου: DBSCAN)

Γραφοθεωρητικές Μέθοδοι

- Υπολογίζεται ένας γράφος εγγύτητας
- Διαγράφεται οποιαδήποτε ακμή του γράφου είναι μεγαλύτερη από τις γειτονικές της
- Το δάσος που προκύπτει αποτελεί το σύνολο των διαμορφούμενων συστάδων

❑HCS

-- αναζητά τις ισχυρές συνιστώσες, υπολογίζοντας την ελάχιστη τομή, αν ικανοποιεί κάποιο κριτήριο (λ.χ. αριθμός ακμών>πλήθος κορυφών), ο γράφος G είναι μία συστάδα, αλλιώς διασπάται σε δύο συνιστώσες και η διαδικασία επαναλμβάνεται.

❑CLICK (CLuster Identification via Connectivity Kernels)

-- ανάλογης λογικής, τα δεδομένα αρχικά κανονικοποιούνται, και υπάρχει και βήμα αντιμετώπισης outliers και συγχώνευση.

❑CLIFF (Clustering via Iterative Feature Filtering)

❑CAST (Cluster Affinity Search Technique)

Χρηστιμοποιεί ένα πίνακα ομοιότητας S και ένα κατώφλι t. Ένα στοιχείο έχει υψηλή ομοιότητα αν $\alpha(x)>=t|C_{open}|$, όπου α(x) η σχέση που έχει το x με τα υπόλοιπα στοιχεία της συστάδας, στη συνέχεια ο αλγόριθμος προσθέτει στοιχεία υψηλής πυκνότητας και αφαιρεί στοιχεία χαμηλής πυκνότητας έως ότου σταθεροποιηθεί μία συστάδα.

Συσταδοποίηση για σύνολα με Κατηγορικά Δεδομένα (1)

- Η χρήση της ευκλείδιας απόστασης, και η χρήση διαμεριστικού αλγορίθμου προβληματική, καθώς υπολογίζοντας κέντρα το κέντρο όλο και περισσότερο απλώνεται σε περισσότερα πεδία

- Η χρήση Jaccard coefficient όπου η ομοιότητα ανάμεσα από δύο συναλλαγές T_1 και T_2 είναι $\left|\frac{T_1 \cap T_2}{T_1 \cup T_2}\right|$

  έχει το πρόβλημα ότι δεν ελέγχει την συνολική ποιότητα του cluster αλλά ελέγχει μόνο τοπικά.

Συσταδοποίηση για σύνολα με Κατηγορικά Δεδομένα (2)

ROCK (RObust Clustering Algorithm for Categorical Attribute)

Εισάγει δύο νέες έννοιες:

- *Γείτονες*. Οι γείτονες ενός σημείου είναι εκείνα τα σημεία τα οποία παρουσιάζουν σημαντική ομοιότητα με αυτό. Θεωρούμε την **sim (p_i, p_j)** ως την συνάρτηση ομοιότητας με βάση την οποία εκτιμούμε την εγγύτητα μεταξύ δύο σημείων και η οποία κυμαίνεται μεταξύ του **0** και **1**. Η συνάρτηση μπορεί να είναι ένα οποιαδήποτε καλά ορισμένο μέτρο απόστασης ή ακόμα και μία μη μετρική συνάρτηση (**π.χ. μία συνάρτηση ομοιότητας που παρέχεται από ειδικούς στο πεδίο που ανήκουν τα στοιχεία που συγκρίνουμε**). Δεδομένης λοιπόν μίας συνάρτησης ομοιότητας και ενός ορίου θ (θ $\in$[0,1]), ένα ζεύγος σημείων p_i, p_j είναι γείτονες εάν ισχύει η ακόλουθη ανισότητα:

$$\mathrm{sim}(p_i, p_j) \geq \theta$$

- *Δεσμοί*. Ο δεσμός link(p_i, p_j) ορίζεται ως ο αριθμός των κοινών γειτόνων μεταξύ των στοιχείων pi, pj.

Συνάρτηση Κριτήριο

Η ακόλουθη συνάρτηση κριτήριο θα πρέπει να μεγιστοποιείται για k συστάδες:

$$E_1 = \sum_{i=1}^{k} n_i \sum_{p_q, p_r \in C_i} \frac{link(p_q, p_r)}{n_i^{1+2f(\theta)}}$$

f(θ) μία παράμετρος η οποία ελέγχει τους γείτονες ενός κόμβου, $n^{f(\theta)}$ ο μέσος όρος γειτόνων ενός κόμβου (κάθε στοιχείο συνεισφέρει σε ένα ζεύγος γειτόνων του).

Μέτρα Ποιότητας

Μπορούμε να ορίσουμε το μέτρο ποιότητας $g(C_i,C_j)$ ως εξής:

$$\sum_{i=1}^{k} \sum_{p_q, p_r \in C_i} \text{link}(p_q, p_r)$$

$$g(C_i, C_j) = \frac{\text{link}[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}}$$

Συσταδοποίηση Kohonen Net

- Τα νευρωνικά δίκτυα Kohonen παρέχουν έναν τρόπο κατηγοριοποίησης των δεδομένων μέσω αυτό-οργανωμένων (**self-organizing**) δικτύων τεχνητών νευρώνων. Δύο βασικές έννοιες που κυριαρχούν στα δίκτυα Kohonen και είναι σημαντικό να κατανοήσουμε είναι, η *ανταγωνιστική μάθηση* και η *αυτό-οργάνωση*.
- Ο όρος *ανταγωνιστική μάθηση* αφορά στην εύρεση ενός νευρώνα ο οποίος προσεγγίζει περισσότερο το πρότυπο εισόδου. Το δίκτυο στη συνέχεια τροποποιεί αυτό τον νευρώνα και τους γειτονικούς του *(***ανταγωνιστική μάθηση με αυτόοργάνωση***)* έτσι ώστε να μοιάζουν περισσότερο με το πρότυπο.

Το επίπεδο ενεργοποίησης είναι:

$$\text{Activation level}_j = \sqrt{\sum_{i=1}^{n} (W_{ij} - X_i)^2}$$

Συσταδοποίηση Kohonen Net

- Τα νευρωνικά δίκτυα Kohonen παρέχουν έναν τρόπο κατηγοριοποίησης των δεδομένων μέσω αυτό-οργανωμένων (**self-organizing**) δικτύων τεχνητών νευρώνων. Δύο βασικές έννοιες που κυριαρχούν στα δίκτυα Kohonen και είναι σημαντικό να κατανοήσουμε είναι, η *ανταγωνιστική μάθηση* και η *αυτό-οργάνωση*.
- Ο όρος *ανταγωνιστική μάθηση* αφορά στην εύρεση ενός νευρώνα ο οποίος προσεγγίζει περισσότερο το πρότυπο εισόδου. Το δίκτυο στη συνέχεια τροποποιεί αυτό τον νευρώνα και τους γειτονικούς του *(***ανταγωνιστική μάθηση με αυτόοργάνωση***)* έτσι ώστε να μοιάζουν περισσότερο με το πρότυπο.

Το επίπεδο ενεργοποίησης είναι:

$$\text{Activation level}_j = \sqrt{\sum_{j=1}^{n} (W_{ij} - X_i)^2}$$

Αλγόριθμος Kohonen

Τα βασικά βήματα του Kohonen αλγορίθμου είναι τα εξής:

✓**Βήμα 1ο** : Για κάθε νευρώνα στο επίπεδο Kohonen λαμβάνεται ένα πλήρες αντίγραφο ενός προτύπου εισόδου.

✓**Βήμα 2ο** : Βρίσκουμε το νευρώνα που είναι ο «νικητής». Ο νικητής είναι αυτός με το μικρότερο επίπεδο ενεργοποίησης:

$$\mathrm{AL_j} = \sqrt{\sum_{j=1}^{n} (W_{ij} - X_j)^2}$$

Βήμα 3ο : Για κάθε νευρώνα που είναι «νικητής» καθώς και για τους φυσικούς γειτονικούς του κόμβους, χρησιμοποιείται ο ακόλουθος κανόνας εκπαίδευσης για την τροποποίηση των βαρών:

$$W_{ij}(t+1) = W_{ij}(t) + \alpha(t) * gamma(t) * [X_i - W_{ij}(t)]$$

$$gamma(t) = \exp\{-0.5 * [r_{ij} / sigma(t)]^2\}$$

όπου ***α*** είναι ο ρυθμός μάθησης ο οποίος μειώνεται με το χρόνο (**αρχίζει από την τιμή 1 και μειώνεται σταδιακά μέχρι την τιμή 0**), rij είναι η απόσταση μεταξύ του νικητή και του κόμβου που πρόκειται να ενημερωθεί και sigma είναι η ακτίνα γειτονίας η οποία μειώνεται με το χρόνο.

✓**Βήμα 4ο** : Επανάληψη των βημάτων 1-3 για κάθε νέο πρότυπο εισόδου.

✓**Βήμα 5ο** : Επανάληψη βήματος 4 έως ότου όλα τα πρότυπα εισόδου εξεταστούν (**αυτό καθορίζει την τιμή του 1**).

✓**Βήμα 6ο** : Επανάληψη βήματος 5 για ένα καθορισμένο αριθμό φορών

Μηχανές Υποστήριξης Διανύσματος

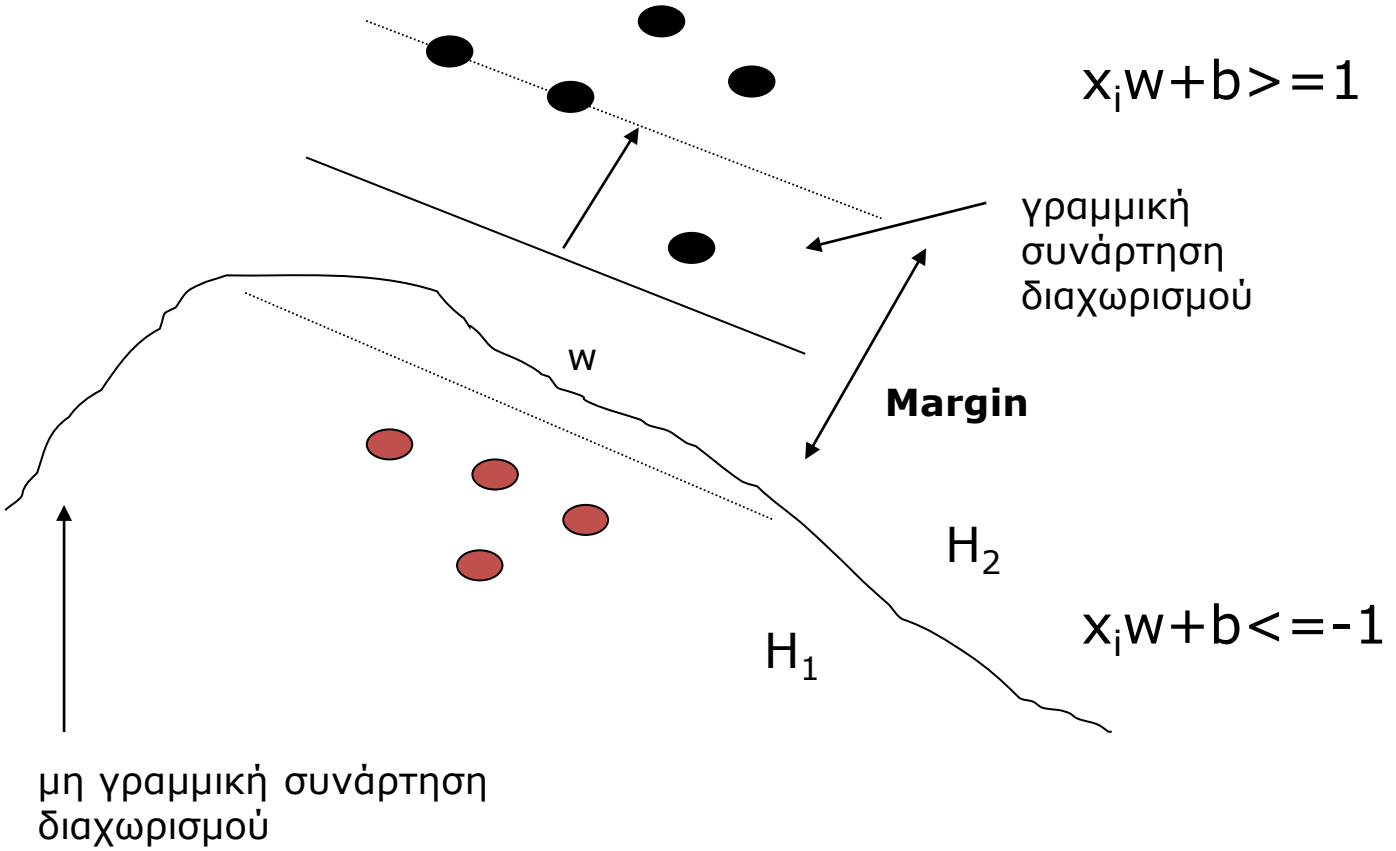
- Βασικές ιδέες:

  -- καθορισμός ενός βέλτιστου υπερ-επιπέδου που διαχωρίζει τα δεδομένα

  -- εφαρμογή και σε μη γραμμικά διαχωρίσιμα δεδομένα (μέσω κατάλληλου μ/χ).

  -- αντιστοίχιση των δεδομένων εισόδου σε νέο χώρο διαστάσεων που μπορούμε να διαχειριστούμε ευκολότερα.

Παράδειγμα Μηχανής Υποστήριξης Διανύσματος



Εφαρμογές σε προβλήματα μοριακής βιολογίας

- Ομαδοποίηση εκφράσεων γονιδίων – clustering gene expression profiles (με SOMs)
- Αναγνώριση περιοχών πρόσδεσης του DNA – identification of binding sites
- Κατηγοριοποίηση στοιχισμένων ακολουθιών

Τέλος Ενότητας

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αθηνών**» 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.

Σημειώματα

Σημείωμα Ιστορικού Εκδόσεων Έργου

Το παρόν έργο αποτελεί την έκδοση 1.0.

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

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.