Μάθηση Κατανομών Πιθανότητας και Ομαδοποίηση

Κώστας Διαμαντάρας
Τμήμα Πληροφορικής
ΤΕΙ Θεσσαλονίκης

Μάθηση κατανομής πιθανότητας

- Σε όλη την ανάλυση μέχρι τώρα έγινε σιωπηρά η παραδοχή ότι γνωρίζουμε την κατανομή πιθανότητας $P(\boldsymbol{x}|C_i)$ για κάθε κλάση C_i. Πρακτικά όμως, συχνά, αυτό δεν ισχύει.
- Συνήθως διαθέτουμε μόνο K δείγματα $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K$ από τα οποία πρέπει να «εκτιμήσουμε» (βλ. να «μάθουμε») τη συνάρτηση $P(\boldsymbol{x}|C_i)$. Αντιμετωπίζουμε, δηλαδή, το γενικότερο πρόβλημα της εκτίμησης μιας κατανομής πιθανότητας.
- Αν το $\boldsymbol{x}$ παίρνει μόνο διακριτές τιμές τότε η εκτίμησή μας είναι σχετικά εύκολη:

$$P(\boldsymbol{x} = \boldsymbol{a}|C_i) = \frac{K_a}{K}$$

  - K_a = το πλήθος των δειγμάτων όπου $\boldsymbol{x} = \boldsymbol{a}$,
  - K = το συνολικό πλήθος των δειγμάτων
- Μεγαλύτερο ενδιαφέρον έχει το πρόβλημα όταν το $\boldsymbol{x}$ μπορεί να πάρει συνεχείς τιμές.

Η μέθοδος του ιστογράμματος

- Ας δούμε κατ' αρχήν την απλή περίπτωση όπου επιθυμούμε να εκτιμήσουμε την κατανομή μιας απλής μεταβλητής *x* (όχι διάνυσμα ***x***).
- **Μέθοδος του ιστογράμματος (histogram):** Χωρίζουμε το διάστημα [–*M*,*M*] όπου παίρνει τιμές το *x* σε *B* ίσα τμήματα που καλούνται κελιά. Το μέγεθος του κάθε κελιού είναι $V=2M/B$. Μετράμε πόσα δείγματα K_i πέφτουν σε κάθε κελί.
- Η ακολουθία $K_1, \ldots, K_B$ λέγεται *ιστόγραμμα*. Η εκτίμησή μας για την πυκνότητα πιθανότητας μέσα στο κελί *i* είναι σταθερή και ίση με

$$\hat{p}_i = \frac{K_i}{KV}$$

- Πλεονεκτήματα της μεθόδου:
  - Απλότητα
  - Δεν απαιτείται να υποθέσουμε ότι η κατανομή ανήκει σε κάποια οικογένεια κατανομών (πχ. τις Γκαουσσιανές, ή άλλες)

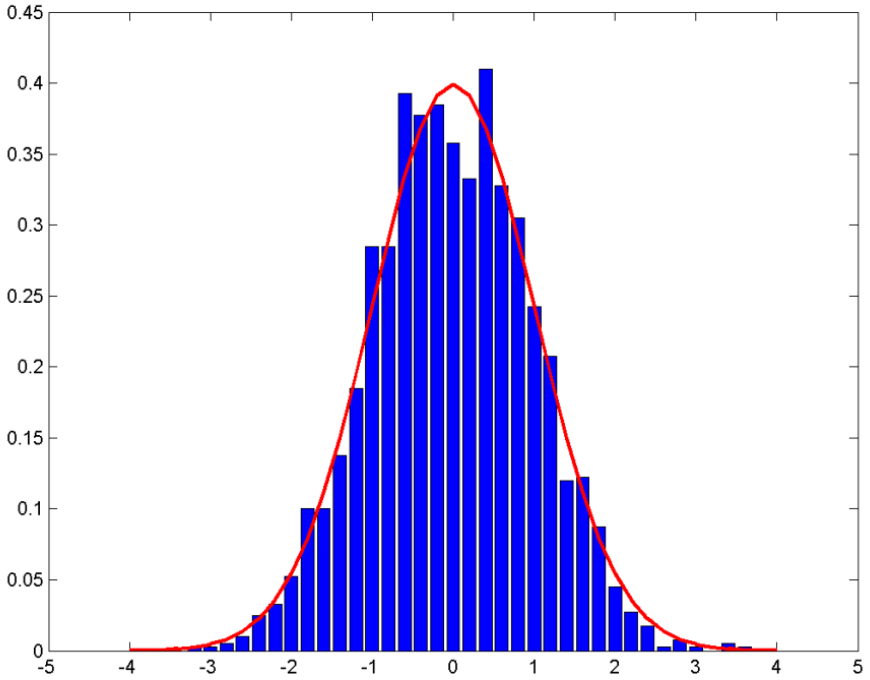
Η μέθοδος του ιστογράμματος

- Μειονεκτήματα
  - Μη συνεχείς τιμές. Όχι ομαλή μετάβαση από κελί σε κελί
  - Κακή ποιότητα εκτίμησης
  - Εξάρτηση από το πλάτος των κελιών. Δες παραδείγματα

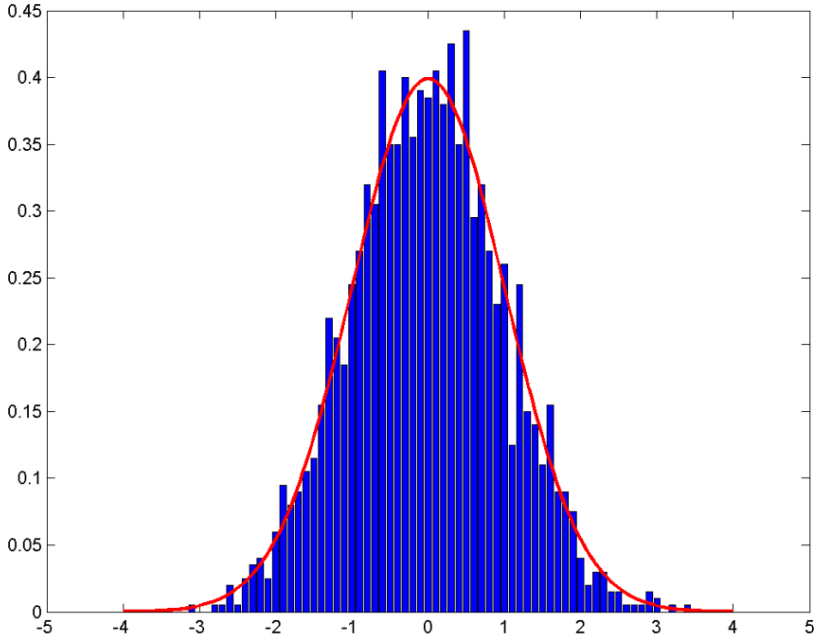
Ιστόγραμμα κανονικής κατανομής από K=2000 δείγματα.

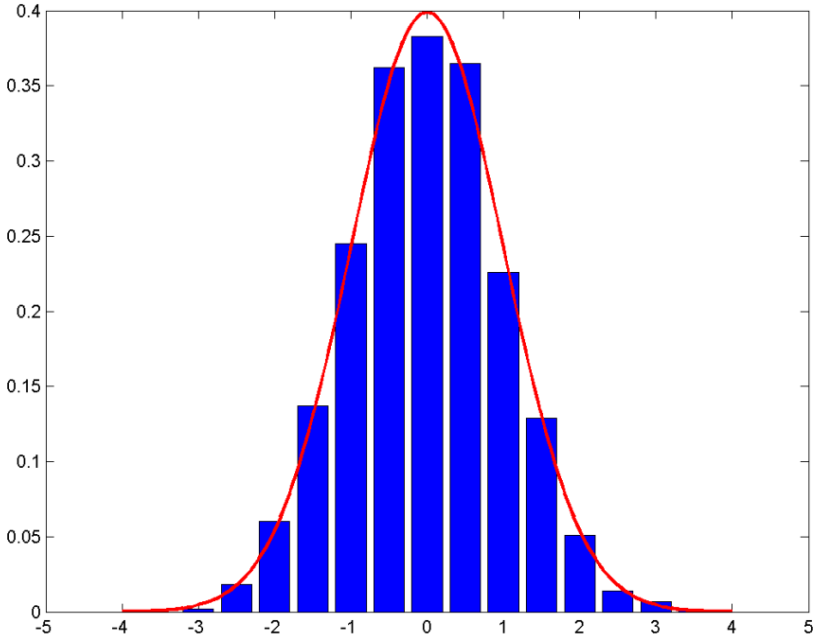
= ιστόγραμμα

= κανονική κατανομή



Η μέθοδος του ιστογράμματος





Η μέθοδος των παραθύρων

- Συλλέγουμε K δείγματα $x_1, x_2, \ldots, x_K$
- Για κάθε δείγμα x_i τοποθετούμε μια συνάρτηση που καλείται «παράθυρο», πχ. την Γκαουσσιανή, κεντραρισμένη στο σημείο αυτό και με πλάτος h (αυθαίρετο), δηλαδή

$$w_i(x) = \frac{1}{h\sqrt{2\pi}} \exp\left\{-\frac{(x - x_i)^2}{2h^2}\right\}$$

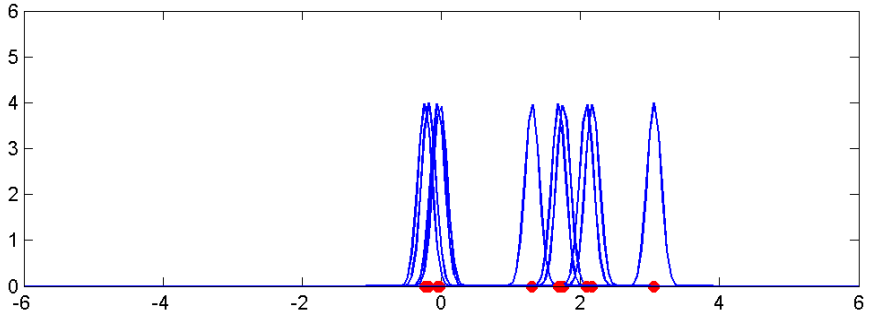
- Συνολικά η εκτίμησή μας για την κατανομή πιθανότητας είναι

$$\hat{p}(x) = \frac{1}{K}\sum_{i=1}^{K} w_i(x) = \frac{1}{Kh\sqrt{2\pi}}\sum_{i=1}^{K} \exp\left\{-\frac{(x - x_i)^2}{2h^2}\right\}$$

- Αν τα δείγματα είναι διανύσματα $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K$ τότε απλώς

$$w_i(\boldsymbol{x}) = \frac{1}{h\sqrt{2\pi}} \exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{x}_i\|^2}{2h^2}\right\}$$

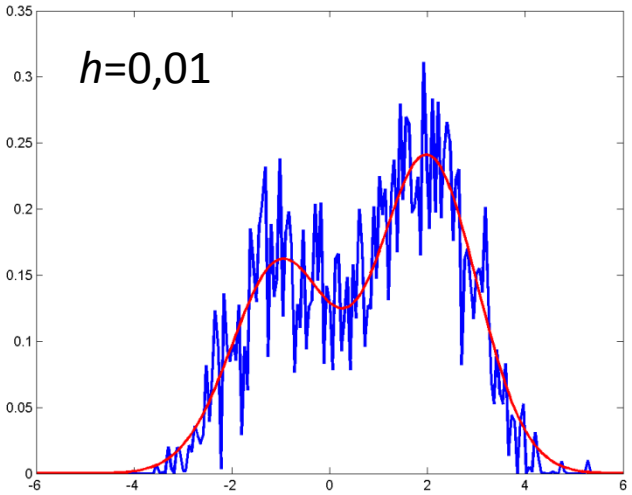
Η μέθοδος των παραθύρων

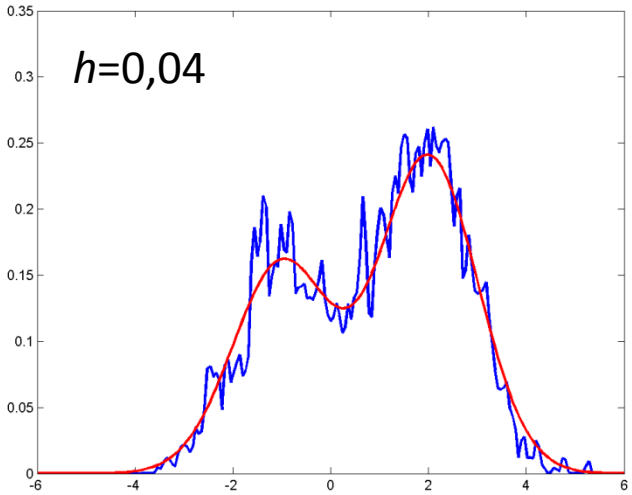


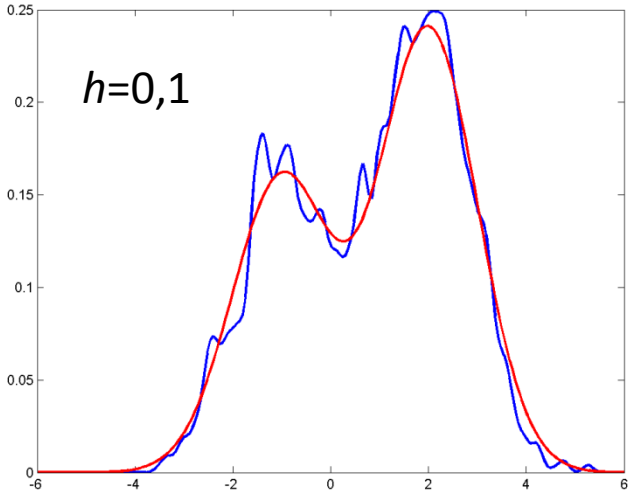
Για κάθε δείγμα ● $x(k)$ βάζουμε ένα Γκαουσσιανό παράθυρο ⌒ κεντραρισμένο στο δείγμα αυτό.

Στο τέλος αφού βάλουμε όλα τα παράθυρα για όλα τα δείγματα, η εκτίμησή μας είναι ο μέσος όρος, δηλαδή αθροίζουμε όλα τα παράθυρα και διαιρούμε δια K (το πλήθος των δειγμάτων και των παραθύρων επίσης). Το πλάτος των παραθύρων h είναι αυθαίρετο και το ορίζει ο χρήστης. Παίζει σημαντικό ρόλο στην ποιότητα της εκτίμησης και την ομαλότητα της καμπύλης εκτίμησης. Δες παραδείγματα

Η μέθοδος των παραθύρων







= *Εκτίμηση*

= *Πραγματική κατανομή*

Όσο μεγαλώνει το *h* τόσο πιο ομαλή είναι η καμπύλη εκτίμησης.

Μειονέκτημα: Δυσκολία εντοπισμού του βέλτιστου *h*.

Αν η κατανομή είναι Γκαουσσιανή...

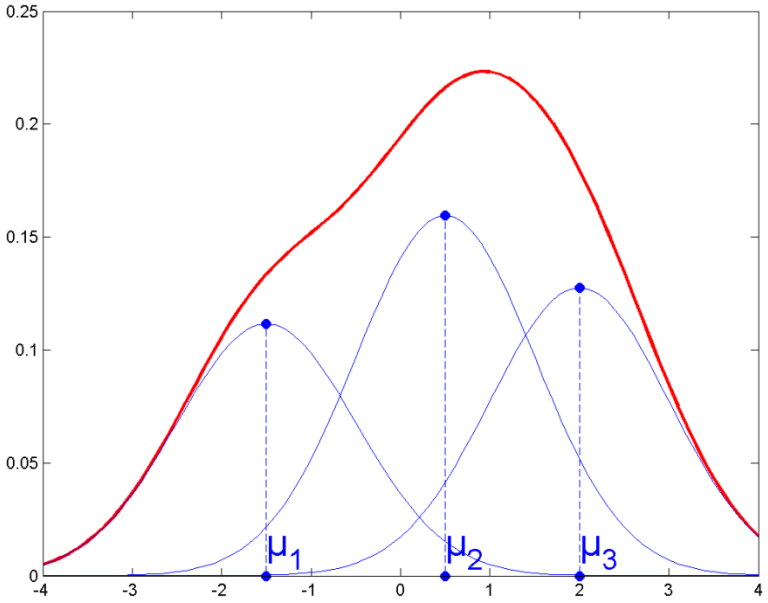
- Αν γνωρίζουμε εκ των προτέρων ότι η κατανομή είναι Γκαουσσιανή τότε τα πράγματα είναι αρκετά απλά: μπορούμε να εκτιμήσουμε όλες τις παραμέτρους της κατανομής, δηλαδή τη μέση τιμή και τη διασπορά, μέσα από τα δείγματα.
- Αν x: αριθμός:
  - Κατανομή: $$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$
  - Εκτίμηση μέσης τιμής: $$\hat{\mu} = \frac{1}{K}\sum_{i=1}^{K} x(k)$$
  - Εκτίμηση διασποράς: $$\hat{\sigma}^2 = \frac{1}{K}\sum_{i=1}^{K} (x(k) - \hat{\mu})^2$$

Αν η κατανομή είναι Γκαουσσιανή...

- Αν $\boldsymbol{x}$: διάνυσμα διάστασης n:
  - Κατανομή: $$p(\boldsymbol{x}) = \frac{1}{(2\pi \cdot \det(\boldsymbol{\Sigma}))} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$
  - Εκτίμηση μέσης τιμής: $$\hat{\boldsymbol{\mu}} = \frac{1}{K}\sum_{i=1}^{K} \boldsymbol{x}(k)$$
  - Εκτίμηση πίνακα διασποράς: $$\hat{\boldsymbol{\Sigma}} = \frac{1}{K}\sum_{i=1}^{K} (\boldsymbol{x}(k)-\hat{\boldsymbol{\mu}})(\boldsymbol{x}(k)-\hat{\boldsymbol{\mu}})^T$$
- Βασικό μειονέκτημα: συνήθως δεν γνωρίζουμε εκ των προτέρων το είδος της κατανομής. Αν το γνωρίζαμε, τότε υπάρχουν αντίστοιχες φόρμουλες εκτίμησης των παραμέτρων για όλες τις γνωστές κατανομές (Gauss, Poisson, Bernoulli, Εκθετική, Rayleigh, Βήτα, Γάμμα, Maxwell, Δυωνυμική, κλπ)

Αν η κατανομή είναι μίγμα Γκαουσσιανών...

- 1 Διάσταση:
- Η κατανομή είναι το άθροισμα *m* Γκαουσσιανών κατανομών.
- Κάθε κατανομή έχει το δικό της κέντρο μ_i και τη δική της διασπορά σ_i^2.
- Κάθε κατανομή από τις m περιγράφει μια ομάδα (cluster) προτύπων G_i



$$p(x) = \sum_{i=1}^{m} P(G_i) \underbrace{\frac{1}{\sigma_i \sqrt{2\pi}} \exp\left\{ -\frac{(x-\mu_i)^2}{2\sigma_i^2} \right\}}_{P(x|G_i)}$$

Ο αλγόριθμος Expectation Maximization (EM) για 1 διάσταση

Αρχικοποίηση: $\hat{P}(G_1) = \cdots = \hat{P}(G_m) = \frac{1}{m}$, $\hat{\mu}_i = τυχαία$, $\hat{\sigma}_1 = \cdots = \hat{\sigma}_m = 1$ $(\pi\chi.)$

Για κάθε επανάληψη *e*=1,…,MAX {

Για κάθε πρότυπο *k*=1,…,*K* {

Για κάθε ομάδα *i*=1,…,*m* {

$$P(G_i|x_k) = \frac{\frac{1}{\hat{\sigma}\sqrt{2\pi}}\exp\left\{-\frac{(x_k-\hat{\mu}_i)^2}{2\hat{\sigma}_i^2}\right\}\hat{P}(G_i)}{\sum_{j=1}^{m}\frac{1}{\hat{\sigma}\sqrt{2\pi}}\exp\left\{-\frac{(x_k-\hat{\mu}_j)^2}{2\hat{\sigma}_j^2}\right\}\hat{P}(G_j)}$$

}

}

Για κάθε ομάδα *i*=1,…,*m* {

$$\hat{\mu}_i = \frac{\sum_{k=1}^{K} P(G_i|x_k)x_k}{\sum_{k=1}^{K} P(G_i|x_k)}$$

$$\hat{\sigma}_i^2 = \frac{1}{2}\frac{\sum_{k=1}^{K} P(G_i|x_k)(x_k-\hat{\mu}_i)^2}{\sum_{k=1}^{K} P(G_i|x_k)}$$

$$\hat{P}(G_i) = \frac{1}{K}\sum_{k=1}^{K} P(G_i|x_k)$$

}

}

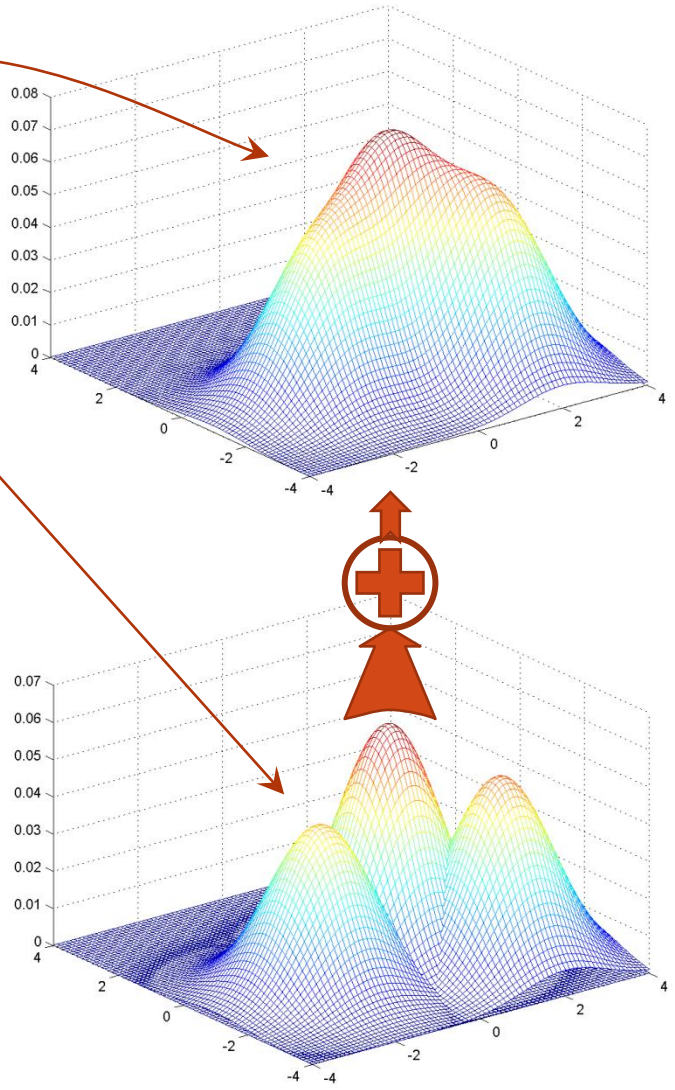
Ο αλγόριθμος ΕΜ

- Επαναληπτικός αλγόριθμος. Εκτιμά (μαθαίνει) όλο και καλύτερα τις παραμέτρους $P(G_i)$, μ_i, και σ_i^2 όσο προχωρούν οι επαναλήψεις.
- Δυστυχώς δεν υπάρχει εγγυημένη σύγκλιση στις σωστές παραμέτρους. Μερικές φορές ο αλγόριθμος κάνει λάθος.
- Προϋποθέτει ότι γνωρίζουμε το πλήθος m των Ομάδων.
- Η αρχικοποίηση παίζει σημαντικό ρόλο

Αν η κατανομή είναι μίγμα Γκαουσσιανών...

- 2 Διαστάσεις:
- Η κατανομή είναι το άθροισμα *m* Γκαουσσιανών κατανομών.
- Κάθε κατανομή έχει το δικό της κέντρο $\boldsymbol{\mu}_i$ και πίνακα διασποράς $\boldsymbol{\Sigma} = \sigma_i^2\mathbf{I}$.
- Κάθε κατανομή από τις *m* περιγράφει μια ομάδα προτύπων G_i (cluster)

$$p(\boldsymbol{x}) = \sum_{i=1}^{m} P(G_i) \underbrace{\frac{1}{2\pi\sigma_i^2} \exp\left\{-\frac{\|x-\mu_i\|^2}{2\sigma_i^2}\right\}}_{P(\boldsymbol{x}|G_i)}$$



Ο αλγόριθμος ΕΜ για 2 διαστάσεις

Αρχικοποίηση: $\hat{P}(G_1) = \cdots = \hat{P}(G_m) = \frac{1}{m}$, $\hat{\boldsymbol{\mu}}_i = τυχαία$, $\hat{\sigma}_1 = \cdots = \hat{\sigma}_m = 1$ $(\pi\chi.)$

Για κάθε επανάληψη *e*=1,…,MAX {

Για κάθε πρότυπο *k*=1,…,*K* {

Για κάθε ομάδα *i*=1,…,*m* {

$$P(G_i|\boldsymbol{x}_k) = \frac{\frac{1}{2\pi\hat{\sigma}_i^2}\exp\left\{-\frac{\|\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_i\|^2}{2\hat{\sigma}_i^2}\right\}\hat{P}(G_i)}{\sum_{j=1}^{m}\frac{1}{2\pi\hat{\sigma}_j^2}\exp\left\{-\frac{\|\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_j\|^2}{2\hat{\sigma}_j^2}\right\}\hat{P}(G_j)}$$

}

}

Για κάθε ομάδα *i*=1,…,*m* {

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{K} P(G_i|\boldsymbol{x}_k)\boldsymbol{x}_k}{\sum_{k=1}^{K} P(G_i|\boldsymbol{x}_k)}$$

$$\hat{\sigma}_i^2 = \frac{1}{2}\frac{\sum_{k=1}^{K} P(G_i|\boldsymbol{x}_k)\|\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_i\|^2}{\sum_{k=1}^{K} P(G_i|\boldsymbol{x}_k)}$$

$$\hat{P}(G_i) = \frac{1}{K}\sum_{k=1}^{K} P(G_i|\boldsymbol{x}_k)$$

}

}

Ο αλγόριθμος ΕΜ για 2 διαστάσεις

Αρχικοποίηση: $\hat{P}(G_1) = \cdots = \hat{P}(G_m) = \frac{1}{m}$, $\hat{\boldsymbol{\mu}}_i = τυχαία, \hat{\sigma}_1 = \cdots = \hat{\sigma}_m = τυχαία$

Για κάθε πρότυπο k=1,…,K {

$$P(G_i|\boldsymbol{x}_k) = \frac{\frac{1}{2\pi\hat{\sigma}_i^2}\exp\left\{-\frac{\|\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_i\|^2}{2\hat{\sigma}_i^2}\right\}\hat{P}(G_i)}{\sum_{j=1}^{m}\frac{1}{2\pi\hat{\sigma}_j^2}\exp\left\{-\frac{\|\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_j\|^2}{2\hat{\sigma}_j^2}\right\}\hat{P}(G_j)}$$

}

Για κάθε ομάδα i=1,…,m {

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{k=1}^{K} P(G_i|\boldsymbol{x}_k)\boldsymbol{x}_k}{\sum_{k=1}^{K} P(G_i|\boldsymbol{x}_k)}$$

$$\hat{\sigma}_i^2 = \frac{1}{2}\frac{\sum_{k=1}^{K} P(G_i|\boldsymbol{x}_k)\|\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_i\|^2}{\sum_{k=1}^{K} P(G_i|\boldsymbol{x}_k)}$$

$$\hat{P}(G_i) = \frac{1}{K}\sum_{k=1}^{K} P(G_i|\boldsymbol{x}_k)$$

}

Άσκηση-3

- Υλοποιήστε τον αλγόριθμο ΕΜ για μια διάσταση σε MATALB
- Χρησιμοποιήστε το ακόλουθο <link> σετ δεδομένων και τρέξτε τον αλγόριθμο
- Προαιρετικά: Υλοποιήστε τον αλγόριθμο ΕΜ για δύο διαστάσεις σε MATLAB
- Χρησιμοποιήστε το ακόλουθο <link> σετ δεδομένων και τρέξτε τον αλγόριθμο

Γιατί ειναι σημαντικά τα μίγματα Γκαουσσιανών;

- Αποδεικνύεται ότι με αθροίσματα Γκαουσσιανών μπορούμε να προσεγγίσουμε οποιαδήποτε συνάρτηση κατανομής με όση ακρίβεια θέλουμε!
- Αρκεί να έχουμε πολλές Γκαουσσιανές...
- Πόσες Γκαουσσιανές ομάδες απαιτούνται ; Ποια είναι η τιμή του *m* ; Η απάντηση δεν είναι εύκολη (και βασικά άγνωστη!)
- Μειονεκτήματα:
  - Ο αλγόριθμος ΕΜ απαιτεί τη γνώση του *m*. Πάντως αν χρησιμοποιήσουμε αρκετά μεγάλο *m* έχουμε καλή πιθανότητα να προσεγγίσουμε οποιαδήποτε συνάρτηση κατανομής.
  - Ο αλγόριθμος ΕΜ μπορεί να μην βρεί την καλύτερη δυνατή λύση αλλά κάτι κοντά σε αυτή...
  - Θέμα τιμών αρχικοποίησης. Ποιες τιμές να δώσω;;;

Ομαδοποίηση (Clustering)

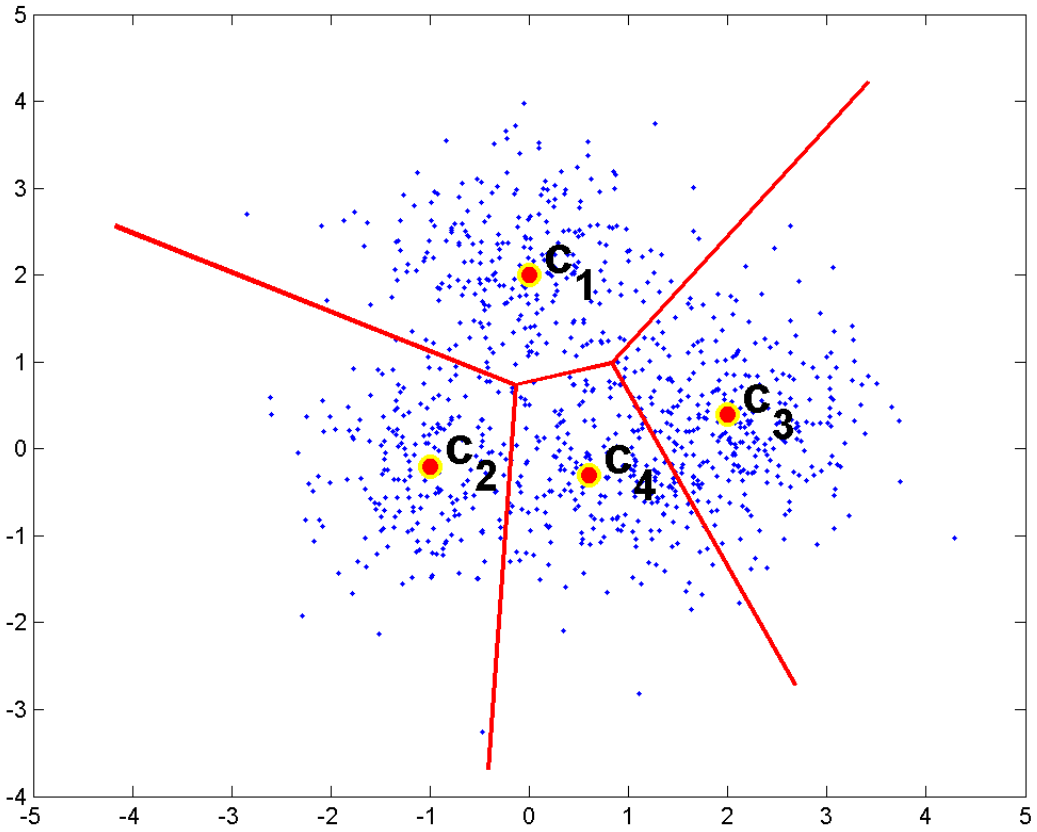
- Στη γενική περίπτωση μια κατανομή αποτελείται από μια ή περισσότερες μικρότερες ομάδες κατανομών οι οποίες μπορεί να είναι Γκαουσσιανές, αλλά μπορεί και όχι.
- Το πρόβλημα της **ομαδοποίησης** (**clustering**) αφορά στον εντοπισμό των ομάδων (κέντρα, διασπορές, κλπ) αλλά και στην απόφαση για κάθε πρότυπο σε ποια ομάδα ανήκει.
- Αν οι κατανομές των ομάδων είναι γνωστές, δηλαδή όλες οι ομάδες ακολουθούν, πχ. την κατανομή Poisson, τότε μπορούμε να εφαρμόσουμε τον αλγόριθμο ΕΜ (εφαρμόζεται για οποιαδήποτε κατανομή ομάδων, όχι μόνο για Γκαουσσιανές ομάδες, αλλά δεν θα μπούμε σε λεπτομέρειες...)
- Τι γίνεται όμως αν δεν γνωρίζουμε τις κατανομές των ομάδων;

Ο αλγόριθμος *K*-μέσων (*K*-means)

- Ο αλγόριθμος αυτός δεν υποθέτει κάποια συγκεκριμένη συνάρτηση κατανομής.
- Θεωρεί γνωστό το πλήθος m των ομάδων (clusters) .
- Θεωρεί ότι κάθε ομάδα G_i αντιπροσωπεύεται από ένα κεντρικό σημείο $\boldsymbol{c}_i$ (κέντρο)

1. Κάθε πρότυπο $\boldsymbol{x}_k$ ανήκει στην ομάδα G_i της οποίας το κέντρο βρίσκεται πιο κοντά στο $\boldsymbol{x}_k$
2. Το κέντρο $\boldsymbol{c}_i$ είναι ο μέσος όρων των προτύπων που ανήκουν στην ομάδα G_i

Ο αλγόριθμος των *K*-μέσων



Ο αλγόριθμος *K*-μέσων

Δίνονται: Τα πρότυπα $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_P$, το πλήθος κέντρων *K*

Έξοδος: *K* κέντρα $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K$

Αρχικοποίησε τα $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K$ σε τυχαίες τιμές.

Επανέλαβε {

Για κάθε πρότυπο $i = 1, \ldots, P$ {

Βρες το κοντινότερο κέντρο $\boldsymbol{c}_j$ στο $\boldsymbol{x}_i$

$label(i) \leftarrow j$ }

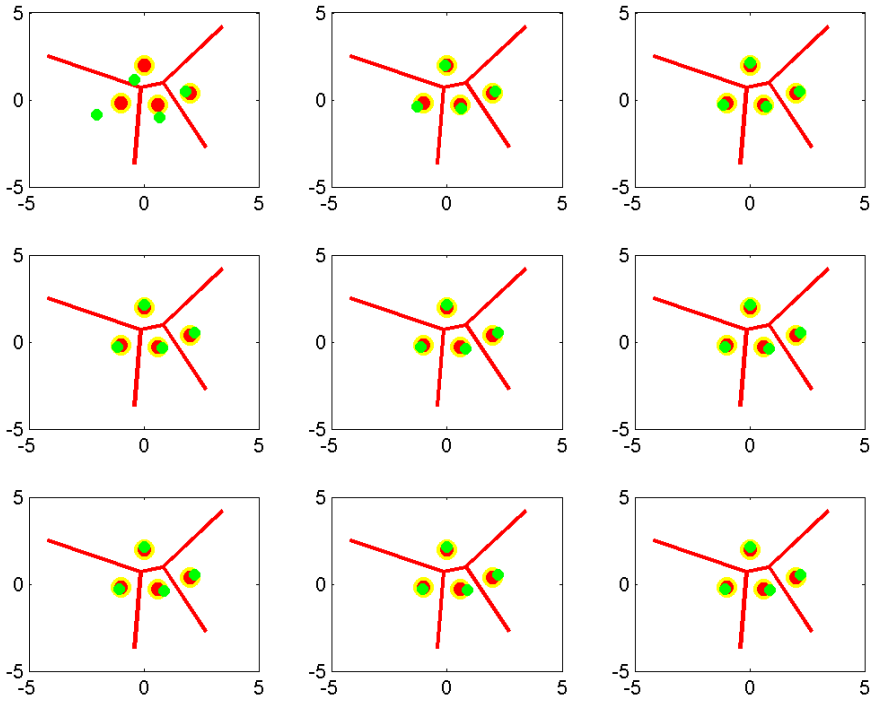
Για κάθε κέντρο $k = 1, \ldots, K$ {

$\boldsymbol{c}_k \leftarrow$ μέσος όρος των $\boldsymbol{x}_i$ τα οποία ανήκουν στην κλάση *k*, δηλαδή για όσα *i*: $label(i) = k$ }

} *Μέχρι* να μην υπάρξει καμία αλλαγή στα κέντρα $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K$

Σύγκλιση του αλγορίθμου

- Δεν υπάρχει εγγυημένη σύγκλιση αλλά συνήθως τα αποτελέσματα είναι καλά έως πολύ καλά.



Άσκηση-4

- Υλοποιήστε τον αλγόριθμο k-means σε MATLAB
- Χρησιμοποιήστε το ακόλουθ σετ δεδομένων <link>