Αναγνώριση Προτύπων

Ομαδοποίηση – Μοντέλα μειγμάτων

Ανδρέας Λ. Συμεωνίδης

Αν. Καθηγητής

Τμήμα Ηλεκτρολόγων Μηχ/κών
&
Μηχ/κών Υπολογιστών, Α.Π.Θ.

Email: asymeon@eng.auth.gr

issel
Intelligent Systems
& Software Engineering
Labgroup

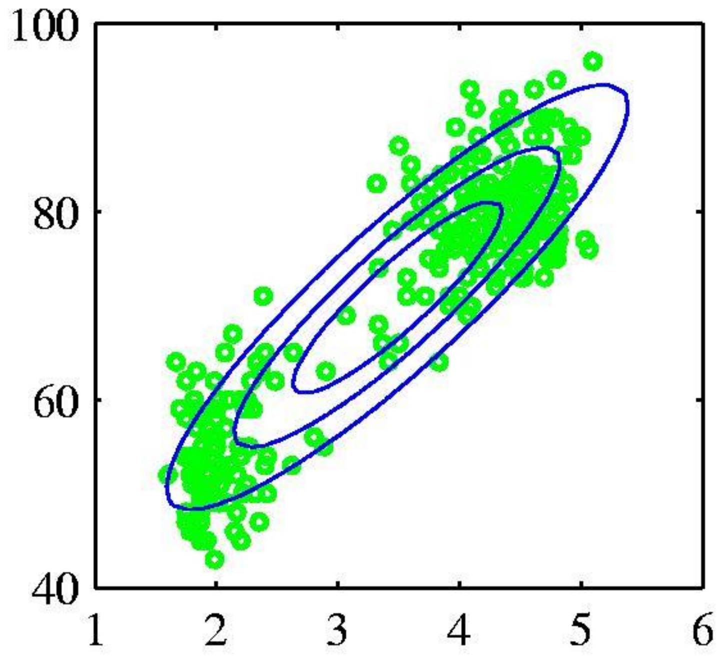
ARISTOTLE
UNIVERSITY OF
THESSALONIKI

Διάρθρωση διάλεξης

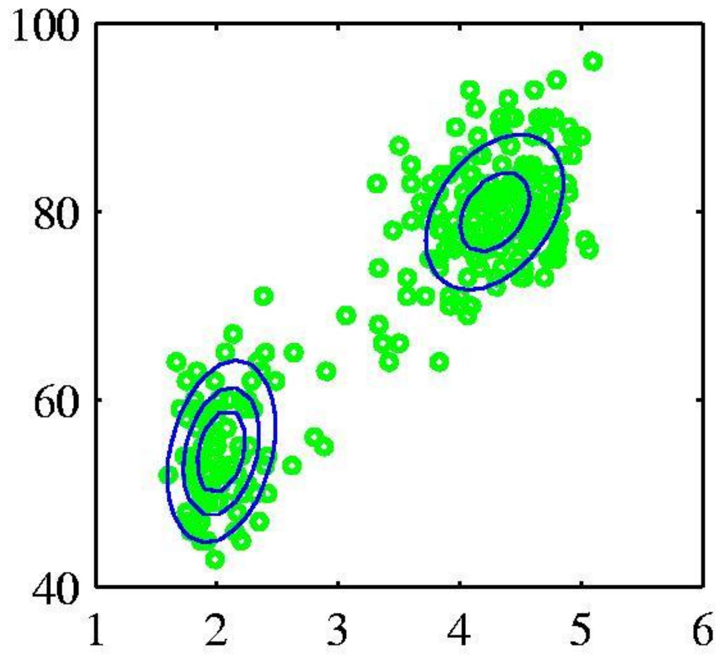
- Μείγμα γκαουσιανών
  - Γενική περιγραφή
  - Θεωρητική ανάλυση
  - Χαρακτηριστικά παραδείγματα
- Ο αλγόριθμος Expectation-Maximization
  - Γενικές αρχές
  - Θεωρητική ανάλυση
  - Χαρακτηριστικά παραδείγματα
  - Πλεονεκτήματα και μειονεκτήματα

Μείγμα Gaussians (1)

Έστω το παρακάτω σετ



Μια Gaussian



Μείγμα δυο Gaussians

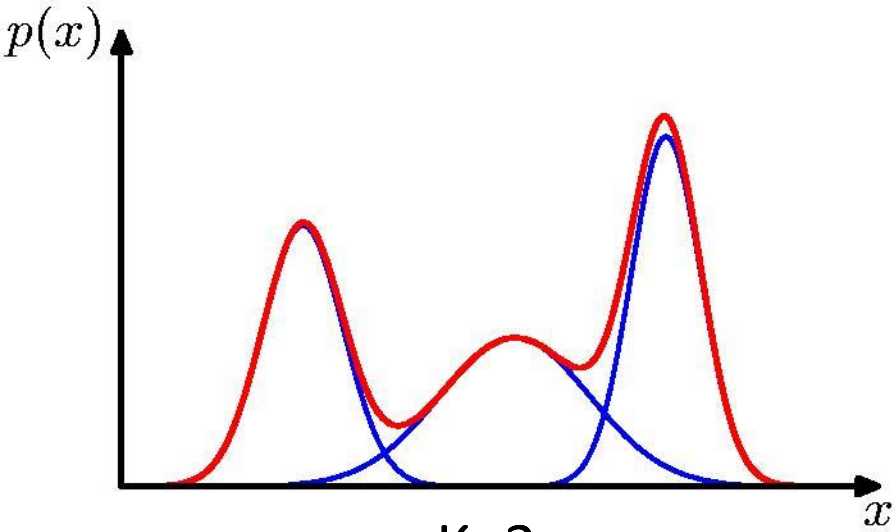
Μείγμα Gaussians (2)

- Συνδυάστε απλά μοντέλα σε ένα σύνθετο:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Τμήμα}}$$

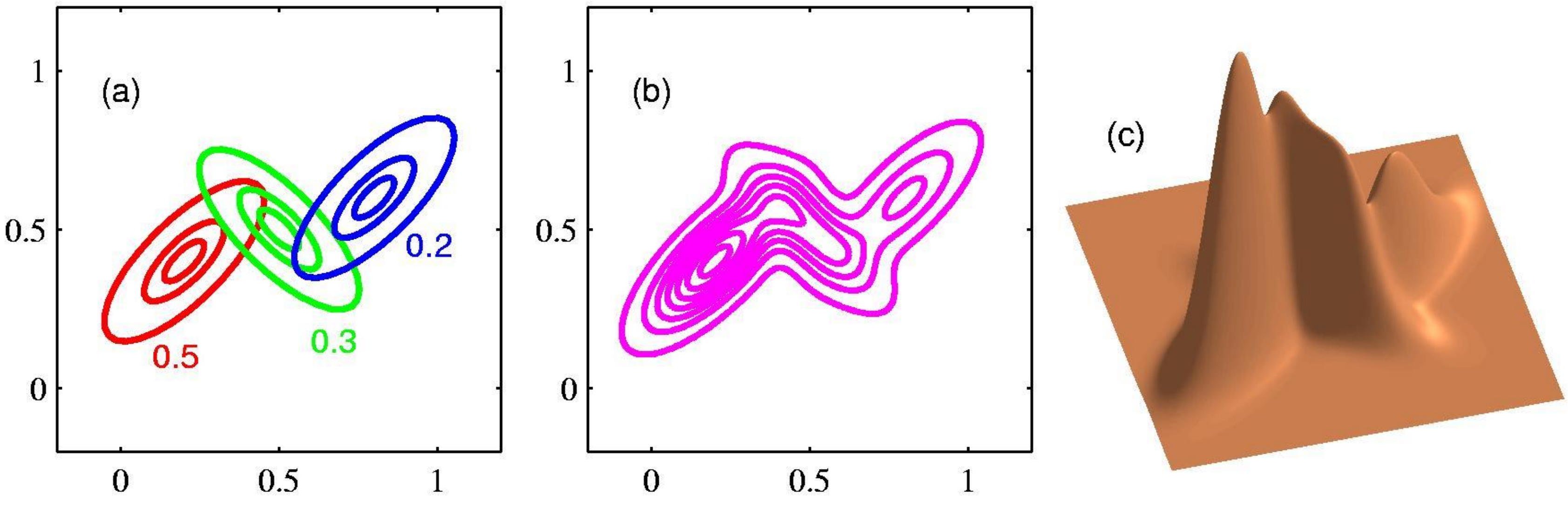
Συντελεστής μείξης

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$



K=3

Μείγμα Gaussians (3)



Μείγμα Gaussians (4)

- Καθορισμός των παραμέτρων *μ*, *Σ*, και *π* χρησιμοποιώντας μέγιστη λογαριθμική πιθανότητα (maximum log likelihood)

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \underbrace{\ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}}_{\substack{\text{Λογάριθμος του αθροίσματος} \\ \text{Δεν υπάρχει μέγιστο κλειστής φόρμας}}}$$

- Λύση: χρήση τυπικών, επαναληπτικών μεθόδων αριθμητικής βελτιστοποίησης ή χρήση του *expectation maximization* - EM αλγορίθμου.

Συχνά απλοποιούμε το μοντέλο

- Τι θυσιάζουμε;
  - Πιστότητα μοντελοποίησης
  - Πιστότητα αλληλεπιδράσεων
  - Ακρίβεια
- Πώς δικαιολογείται αυτή η ύπαρξη απλούστευσης;
  - Αυξημένη “προφανής” τυχαιότητα στον κόσμο
  - Μεγαλύτερες διακυμάνσεις
- Τι κερδίζουμε;
  - Έναν λιγότερο εκφραστικό χώρο υποθέσεων
  - Γρηγορότερη εκμάθηση/σύγκλιση
- Θέλει προσοχή!

Πολλές φορές η απλούστευση δεν επηρεάζει την επίδοση

- Πιθανοί λόγοι:
  - Συνήθως αποκλείει τις λεπτομέρειες
    - Συνήθως το “γραμμικό” είναι το κυρίαρχο χαρακτηριστικό ή αυτό με τη μεγαλύτερη επιρροή
  - Απλουστευμένο μοντέλο $\rightarrow$ λιγότερες παράμετροι
    - Σετ δεδομένων ίδιου μεγέθους $\rightarrow$ η πληροφορία & η εμπιστοσύνη επικεντρώνεται σε εκείνες τις παραμέτρους που μοντελοποιούν τις κυρίαρχες επιδράσεις
  - Θέλουμε να επιλέξουμε τα χαρακτηριστικά που παρέχουν “απόδειξη”
    - Διαισθητικά (ή με δοκιμή και σφάλμα), αποφεύγουμε τα χαρακτηριστικά που έχουν πολύ ισχυρές αλληλεπιδράσεις

Πίσω στον Bayes / Παραμετροποιήσιμα μοντέλα

Λανθάνουσες μεταβλητές, Ελλιπείς τιμές και ο αλγόριθμος EM

- Τι συμβαίνει αν δεν μπορούμε να παρατηρήσουμε ορισμένες μεταβλητές;
- Αυτές είναι γνωστές ως λανθάνουσες μεταβλητές, οι οποίες κάνουν (μέχρι και) το σετ εκπαίδευσης μη πλήρες
- Συνήθως, οι μεταβλητές αυτές (οι μη παρατηρούμενες) είναι και οι πιο ενδιαφέρουσες:
  - η ειλικρίνεια κάποιου που κάνει μια αίτηση, η πρόθεση αποπληρωμής ενός δανείου, η ύπαρξη φθοράς στους μύες της καρδιάς…
- Δε χρειαζόμαστε υποχρεωτικά μοντέλο ταξινόμησης. Απλά ένα μοντέλο που σέβεται την ύπαρξη λανθανουσών μεταβλητών
- Γνωρίζουμε τις σχέσεις ανάμεσα στις μεταβλητές (έστω ένα πιθανοτικό δίκτυο)
- Παρατηρούμε τις τιμές για κάποιες από τις μεταβλητές

Ο αλγόριθμος EM (Expectation Maximization)

- Τι μπορούμε να κάνουμε;
- Θέλουμε τις τιμές των παραμέτρων (CPT entries) που κάνουν τα δεδομένα πιο πιθανά
- Χρειαζόμαστε τιμές παραμέτρων για τις λανθάνουσες τυχαίες μεταβλητές.
- **Ο αλγόριθμος EM:**
  - Κάνουμε τις καλύτερες προβλέψεις μας για τα δεδομένα που λείπουν
  - Στη συνέχεια επάγουμε τις παραμέτρους που λείπουν με βάση τις προβλέψεις μας
  - Οι τιμές των παραμέτρων είναι πιθανότατα λανθασμένες, γι’ αυτό επαναλαμβάνουμε τη διαδικασία
  - Σταματάμε όταν υπάρχει σύγκλιση

Ο αλγόριθμος ΕΜ (Expectation Maximization)

- Ο ΕΜ είναι μια ομάδα αλγορίθμων η οποία χρησιμοποιείται για την εκτίμηση μιας κατανομής πιθανότητας δεδομένης της ύπαρξης ελλειπουσών τιμών.
- Για τη χρήση του απαιτείται μια υπόθεση για την υποκείμενη κατανομή πιθανότητας.
- Ο αλγόριθμος μπορεί να είναι πολύ ευαίσθητος σε σχέση με την υπόθεση αυτή και το σημείο εκκίνησης (δηλ. την αρχική εκτίμηση των παραμέτρων).
- Είναι ένας hill-climbing αλγόριθμος και συγκλίνει σε κάποιο τοπικό μέγιστο της συνάρτησης πιθανοφάνειας (likelihood).

Η συνάρτηση πιθανοφάνειας

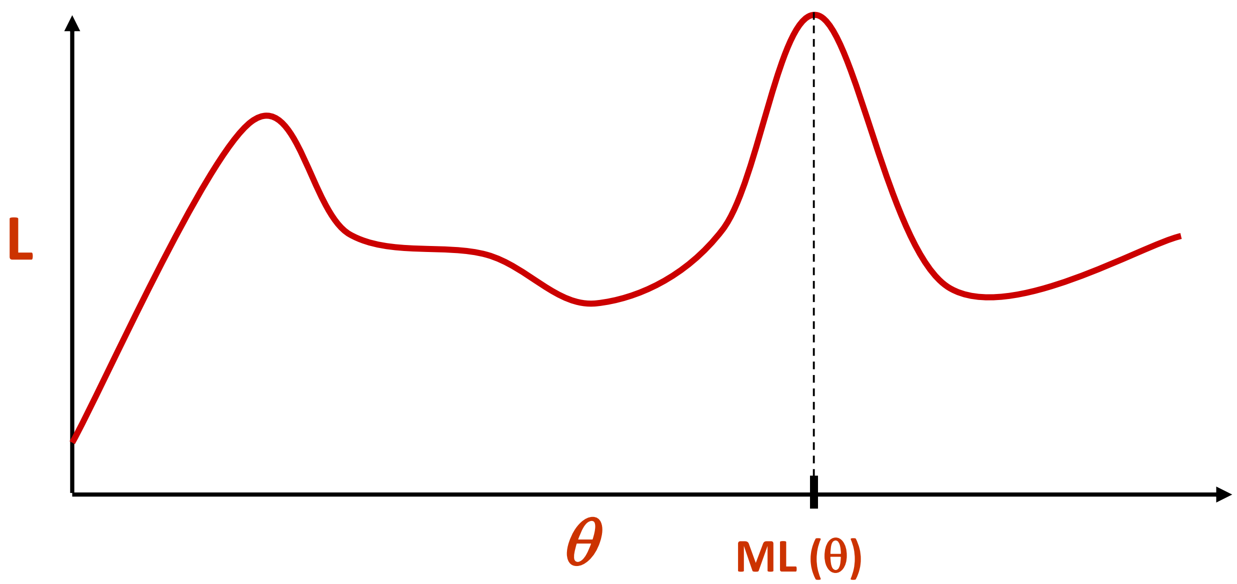
- Σετ δεδομένων εκπαίδευσης *D*
- Μια παραμετρική οικογένεια μοντέλων *w/* παραμέτρων ϑ
- Επιθυμούμε ϑ που μεγιστοποιεί το $Pr(\vartheta \mid D)$

Κατά Bayes: $Pr(\vartheta \mid D) = Pr(D \mid \vartheta) * Pr(\vartheta) / Pr(D)$

- Το $Pr(D \mid \vartheta)$, όταν το διαχειριζόμαστε ως μια συνάρτηση του ϑ, ονομάζεται συνάρτηση πιθανοφάνειας: $L(\vartheta;D) = Pr(D \mid \vartheta)$ (κάποιες φορές $L(\vartheta,D)$ ή $L(\vartheta|D)$)
- Η συνάρτηση πιθανοφάνειας ΔΕΝ είναι μια κατανομή πιθανότητας ως προς ϑ και το ολοκλήρωμα/άθροισμα δε χρειάζεται να είναι = 1.0
- Η μέγιστη πιθανοφάνεια (Maximum Likelihood – ML) ϑ, είναι αυτή που μεγιστοποιεί τη συνάρτηση πιθανοφάνειας

Η συνάρτηση πιθανοφάνειας

- Επιθυμούμε τη μέγιστη πιθανοφάνεια ϑ, ακόμη και αν το D δεν είναι ολόκληρο
- Η συνάρτηση πιθανοφάνειας μπορεί να μην είναι ομαλή ή μονότονη
- Οι αναλυτικές λύσεις είναι συνήθως δύσκολες ή αδύνατες

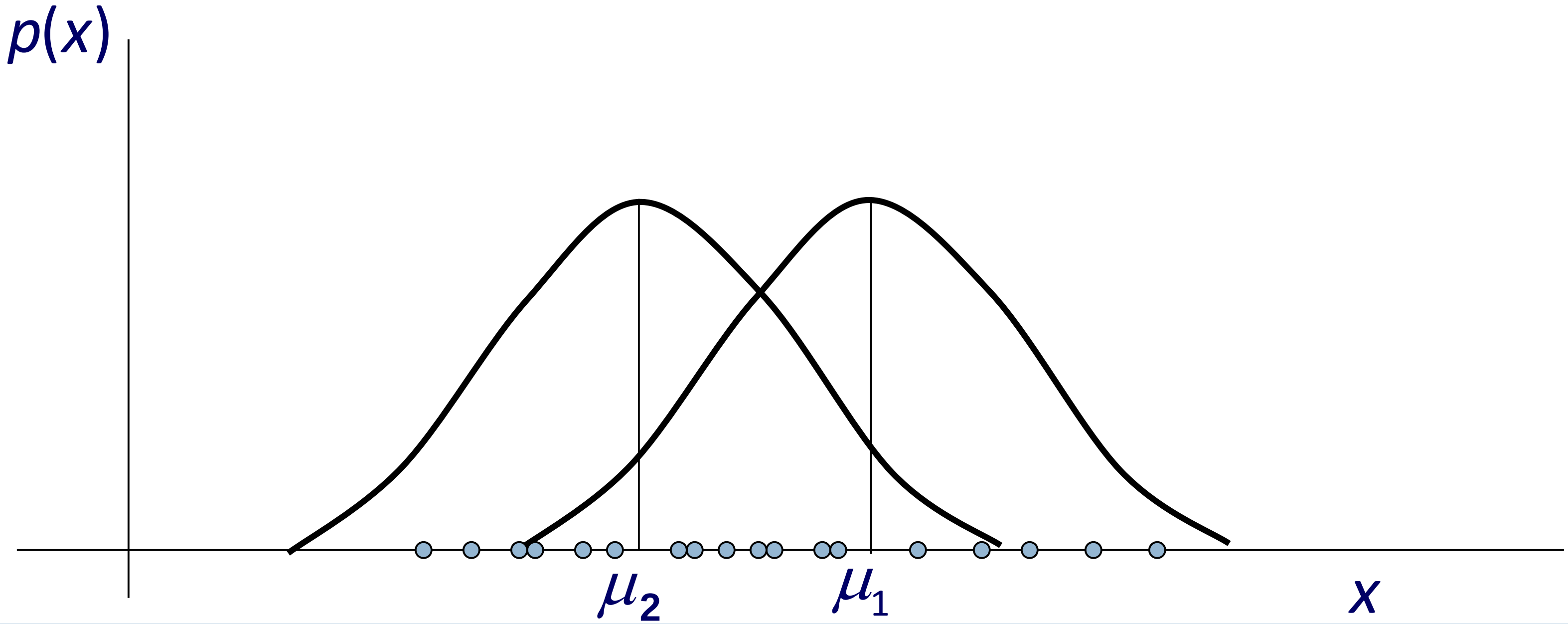


Διαισθητικό παράδειγμα

- Θεωρήστε ότι υπάρχουν *K* διεργασίες που παράγουν στοχαστικά δεδομένα
- Κάθε διεργασία παράγει δεδομένα κανονικής κατανομής, παρότι οι κανονικές κατανομές είναι διαφορετικές
- Μόνο οι μέσες τιμές διαφέρουν. Η διακύμανση είναι η ίδια (έστω σ=2).
- Δε γνωρίζουμε ποια διεργασία έχει παράξει ποια δεδομένα
- Για κάθε σημείο, ποια διεργασία το έχει παράξει είναι λανθάνουσα μεταβλητή γι’ αυτό.

Ομαδοποίηση

- Παρατηρούμε τα γαλάζια σημεία
- Ανεξάρτητα δείγματα από 2 κανονικές κατανομές, με διαφορετική μέση τιμή αλλά ίδια τυπική απόκλιση
- Υπολογίστε τη μέση τιμή μ_i, i=1,2 και την τυπική απόκλιση σ από τα δεδομένα



Μέγιστη Πιθανοφάνεια Πλήρης Παρατηρησιμότητα

- Εάν γνωρίζαμε ποια διεργασία έχει παράξει ποια σημεία, η εύρεση των πιο πιθανών παραμέτρων θα ήταν εύκολη.

$$\mathbf{p}(\mathbf{x} \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathbf{exp}[-\frac{1}{2\sigma^2}(\mathbf{x} - \mu)^2]$$

- Παίρνουμε πολλά σημεία, *D = {x₁,…,xₘ}*
- Η μεγιστοποίηση της λογαριθμικής πιθανοφάνειας (log-likelihood) είναι ανάλογη με την μείωση:

$$\ln(L(\theta \mid D)) = \ln(P(D \mid \mu)) = \sum_i -\frac{1}{2\sigma^2}(x_i - \mu)^2$$

- Με υπολογισμό της παραγώγου ως προς P βρίσκουμε ότι το ελάχιστο σημείο, δηλ. η πιο πιθανή μέση τιμή είναι:

$$\mu_{ML} = \mathbf{argmin}_{\mu} \sum_i (x_i - \mu)^2 \qquad \mu = \frac{1}{m}\sum_i x_i$$

Ένα μείγμα κατανομών

- Δε γνωρίζουμε πώς δημιουργήθηκε κάθε σημείο x_i
- Δε γνωρίζουμε τη γενική παραμετρική μορφή της πιθανότητας με την οποία το παρατηρούμενο σημείο x_i δημιουργήθηκε από την κατανομή *j: (μ_j,σ)*

$$P_{ij} = P(\mu_j \mid x_i) = \frac{P(x_i \mid \mu_j)P(\mu_j)}{P(x_i)} = \frac{\frac{1}{k}P(x = x_i \mid \mu = \mu_j)}{\sum_{n=1}^{k} \frac{1}{k}P(x = x_i \mid \mu = \mu_n)} =$$

$$= \frac{\exp[-\frac{1}{2\sigma^2}(x_i - \mu_j)^2]}{\sum_{n=1}^{k} \exp[-\frac{1}{2\sigma^2}(x_i - \mu_n)^2]}$$

Ένα μείγμα κατανομών

- Έστω $X=\{x_1,\ldots x_m\}$ τα παρατηρούμενα δεδομένα (m ανεξάρτητα δείγματα)
- Για κάθε σημείο x_i, ορίστε k δυαδικές κρυφές μεταβλητές, $z_{i1},z_{i2},\ldots,z_{ik}$ έτσι ώστε $z_{ij}=1$ *αν και μόνο αν* το x_i έχει ληφθεί από την *j-th* κατανομή.
- Έστω Y το πλήρες σετ δεδομένων.
- *E[Y]* είναι η προσδοκία (expectation) του Y. Θυμηθείτε ότι το *E* είναι γραμμικός τελεστής

$$E[z_{ij}]=1\bullet P(x_i \text{ was sampled from } \mu_j)+ 0\bullet P(x_i \text{ was not sampled from } \mu_j)= P_{ij}$$

$$\mathbf{E[Y]} = \sum_{y_i} y_i P(Y = y_i) \qquad \mathbf{E[X+Y]} = \mathbf{E[X]} + \mathbf{E[Y]}$$

Παράδειγμα: ο K-Means...

- **Expectation**:

$$p(y_i | \theta) = p(x_i, z_{i1}, \ldots, z_{ik} | \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\sum_j z_{ij}(x_i - \mu_j)^2\right]$$

- Υπολογισμός της πιθανοφάνειας δεδομένων των παρατηρημένων σημείων $D = \{x_1, \ldots, x_m\}$

$$\ln(P(Y | \theta)) = \sum_{i=1}^{m} -\frac{1}{2\sigma^2}\sum_j z_{ij}(x_i - \mu_j)^2$$

- και της υπόθεσης ϑ' (απορρίπτεται ο σταθερός παράγοντας)

$$E[\ln(P(Y | \theta'))] = E\left[\sum_{i=1}^{m} -\frac{1}{2\sigma^2}\sum_j z_{ij}(x_i - \mu_j)^2\right] =$$

$$= \sum_{i=1}^{m} -\frac{1}{2\sigma^2}\sum_j E[z_{ij}](x_i - \mu_j)^2$$

Παράδειγμα: ο K-Means…

- **Maximization**: Μεγιστοποίηση του

$$Q(\theta \mid \theta') = \sum_{i=1}^{m} - \frac{1}{2\sigma^2} \sum_j E[z_{ij}](x_i - \mu_j)^2$$

- σε σχέση με το:

$$\frac{dQ}{d\mu_j} = C\sum_{i=1}^{m} E[z_{ij}](x_i - \mu_j)\mu_j = 0$$

- συνεπάγεται:

$$\mu_j = \frac{\sum_{i=1}^{m} E[z_{ij}]x_i}{\sum_{i=1}^{m} E[z_{ij}]}$$

Ο K-Means συνολικά...

- Δεδομένου ενός σετ D = {x_1,…,x_m} από σημεία, μάντεψε τις αρχικές παραμέτρους $\sigma, \mu_1, \mu_2, \ldots, \mu_k$
- Υπολόγισε (για όλα τα i,j):

Expectation:
ορισμός των κρυφών τιμών

$$p_{ij} = E[z_{ij}] = \frac{\exp[-\frac{1}{2\sigma^2}(x_i - \mu_j)^2]}{\sum_{n=1}^{k} \exp[-\frac{1}{2\sigma^2}(x_i - \mu_n)^2]}$$

- Υπολόγισε τις παραμέτρους ML:

Maximization:
ανανέωση των παραμέτρων

$$\mu_j = \frac{\sum_{i=1}^{m} E[z_{ij}]x_i}{\sum_{i=1}^{m} E[z_{ij}]}$$

- Επανέλαβε μέχρι τη σύγκλιση
- Παρατηρήστε ότι ο αλγόριθμος θα βρει τον βέλτιστο K-means με τη λογική της ελαχιστοποίησης του SSE.

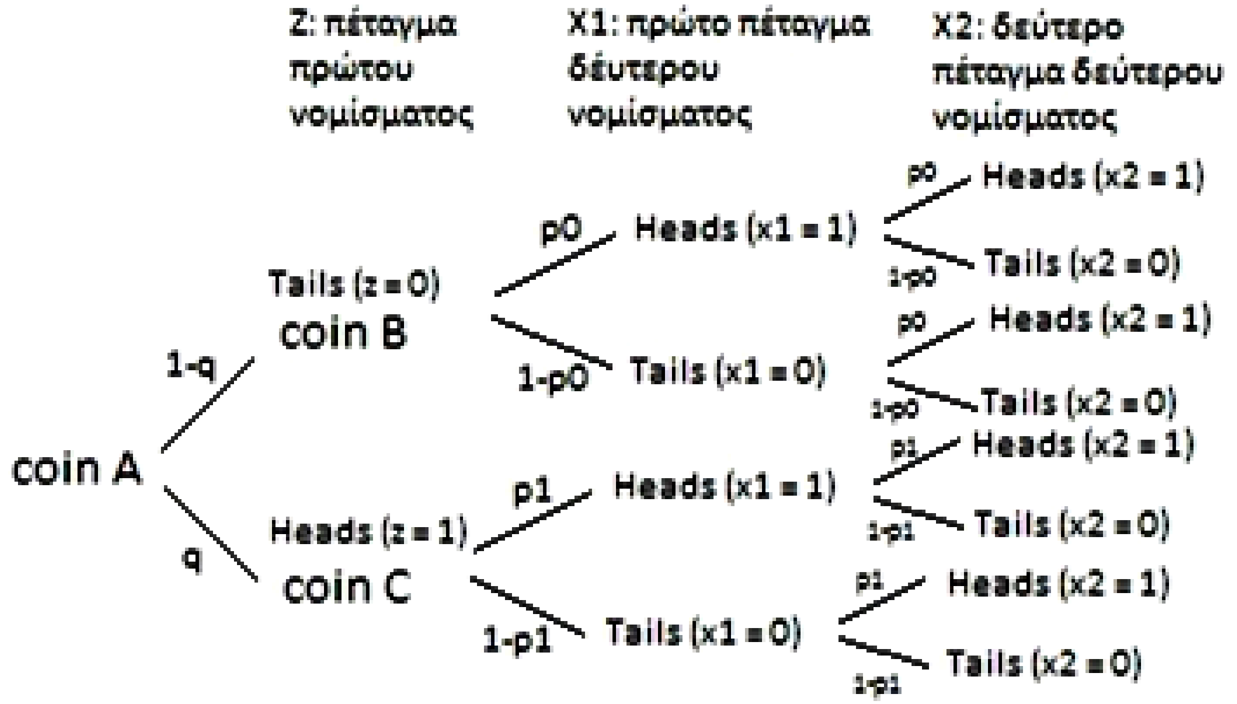
Συνοψίζοντας τον ΕΜ…

- Ο ΕΜ είναι μια γενική διαδικασία μάθησης όταν υπάρχουν μη παρατηρούμενες μεταβλητές.
- Ο ΕΜ είναι ένας επαναληπτικός αλγόριθμος, οποίος αποδεικνύεται ότι συγκλίνει:
  - Σε τοπικό μέγιστο της συνάρτησης πιθανοφάνειας
  - Δεδομένου ότι έχει γίνει μια αρχική υπόθεση μιας οικογένειας από κατανομές πιθανότητας
- Εάν οι αρχικές υποθέσεις είναι σχετικά ορθές, τότε ο ΕΜ αποδεικνύεται πολύ χρήσιμος στην πράξη.
- Ως παραδείγματα, είδαμε το στρίψιμο τριών κερμάτων και τον K-means αλγόριθμο.

Άσκηση EM…

- Θεωρήστε ότι έχουμε μία τυχαία μεταβλητή Z που αναπαριστά το αποτέλεσμα της ρίψης ενός κέρματος.
- Η μεταβλητή αυτήν παίρνει την τιμή 0 εάν το αποτέλεσμα του κέρματος είναι γράμματα και την τιμή 1 σε διαφορετική περίπτωση.
- Σε περίπτωση που το αποτέλεσμα του κέρματος είναι γράμματα (Z=0) πραγματοποιούμε 2 ακόμα ρίψεις με ένα δεύτερο διαφορετικό κέρμα το οποίο εμφανίζει πιθανότητα p_0 να φέρει κορώνα.
- Εάν το αποτέλεσμα του πρώτου κέρματος είναι κορώνα (Z=1), πραγματοποιούμε τις 2 αυτές ρίψεις χρησιμοποιώντας ένα τρίτο κέρμα που εμφανίζει αντίστοιχη πιθανότητα εμφάνισης κορώνας p_1.
- Η πιθανότητα το πρώτο κέρμα να φέρει κορώνα είναι ίση με q. Στην παρακάτω εικόνα απεικονίζεται η διαδικασία του πειράματος με τις αντίστοιχες πιθανότητες.

Άσκηση EM (συν.)…



Άσκηση EM (συν.)…

Α. Εάν θεωρήσουμε ότι p_0=1/4, p_1=3/4 και q=1/2, να υπολογίσετε σύμφωνα με το θεώρημα Bayes την πιθανότητα η μεταβλητή Z να είχε την τιμή 1, δεδομένου ότι οι τιμές των x_1, x_2 που παρατηρήθηκαν κατά την εκτέλεση του πειράματος ήταν αντίστοιχα 1 και 0.

Άσκηση EM (συν.)…

- β) Ύστερα από 10 επαναλήψεις του πειράματος που περιγράφηκε στην εκφώνηση παρατηρούμε τις εξής τριάδες (Z, x_1, x_2).
- (? 1 0), (? 0 0), (? 1 1), (? 0 1), (? 1 1), (? 0 0), (? 1 0), (? 0 1), (? 1 0), (? 0 0)
- Γνωρίζοντας ότι p_0=1/4, p_1=3/4 και αγνοώντας τελείως την τιμή του q αλλά και της μεταβλητής Z (αποτέλεσμα πρώτου κέρματος), να εφαρμοστεί ο αλγόριθμος EM (Expectation Maximization) για 2 επαναλήψεις προκειμένου να προσδιοριστούν η μη παρατηρίσιμη μεταβλητή (Z) σε κάθε περίπτωση καθώς και η παράμετρος q. Να περιγράψετε τα βήματα του αλγορίθμου και να συμπληρώσετε τους παρακάτω πίνακες. Θεωρήστε ως αρχική τιμή της παραμέτρου q στην πρώτη επανάληψη του αλγορίθμου την τιμή 0.1 (q^1=0.1).

Άσκηση EM (συν.)...

1η Επανάληψη ΕΜ				2η Επανάληψη ΕΜ		
X_1	X_2	Ζ θεωρώντας ότι $q^1=0.1$		X_1	X_2	Ζ θεωρώντας ότι $q^2=$
1	0			1	0	
0	0			0	0	
1	1			1	1	
0	1			0	1	
1	1			1	1	
0	0			0	0	
1	0			1	0	
0	1			0	1	
1	0			1	0	
0	0			0	0	

Πηγές

- Introduction to Data Mining, Tan, Steinbach, Kumar.
- Pattern recognition, Theodoridis & Koutroumbas
- Pattern recognition, Bishop.