ΣΤΑΤΙΣΤΙΚΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗ

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
Σαμος

2021

Εισαγωγή

- Υπόθεση
  - Τα δεδομένα προέρχονται από την ανάμειξη (mixture) διαφορετικών κατανομών
  - Κάθε συστάδα εκφράζει μια κατανομή
- Ζητούμενο
  - Να εντοπίσουμε τις εν λόγω κατανομές μαζί με τις παραμέτρους τους
- Μεθοδολογία
  - Εντοπισμός παραμέτρων κάθε κατανομής ⇒ Εκτίμηση Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimation – MLE)
  - Εντοπισμός παραμέτρων μοντέλου ανάμειξης (mixture model) ⇒ Αλγόριθμος Expectation-Maximization (EM)

Εισαγωγή

- Ο αλγόριθμος ΕΜ λειτουργεί εκτιμώντας τα δεδομένα που λείπουν(E-step) και έπειτα εκτιμώντας τις παραμέτρους του μοντέλου με την μεγαλύτερη ομοιότητα (M-step). Η προσέγγιση αυτή απαιτεί η συλλογή αντικειμένων και οι ομάδες τους (clusters) να αναπαρίστανται από ένα στατιστικό μοντέλο. Τα δεδομένα θεωρούνται σαν ένα τυχαίο δείγμα από ένα μίγμα πιθανοτικών κατανομών (distributions).

Μοντέλα Μίξης

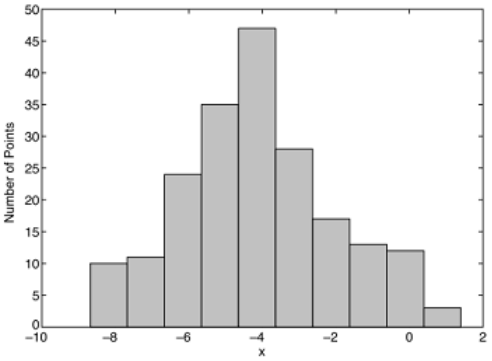
- M πρότυπα εισόδου $\mathcal{X} = \{x_1, x_2, \dots, x_M\}$ που δημιουργούνται από K ανεξάρτητες κατανομές που περιγράφονται από σύνολο παραμέτρων $\Theta = \{\theta_1, \theta_2, \dots, \theta_K\}$
- Πιθανότητα το i-οστό πρότυπο εισόδου να προέρχεται από την j-οστή κατανομή $p(x_i|\theta_j)$
  - Αρκετά συχνά επιλέγεται η πολυμετάβλητη κανονική κατανομή, γιατί δημιουργεί συστάδες ελλειπτικού σχήματος γύρω από τη μέση τιμή της
- Πιθανότητα εμφάνισης του i-οστού προτύπου $p(x_i|\Theta) = \sum_{j=1}^{K} w_j p(x_i|\theta_j)$
  - w_j: συνεισφορά (βάρος) j-οστής κατανομής $(\sum_{j=1}^{K} w_j = 1)$
- Τα πρότυπα δημιουργούνται ανεξάρτητα το ένα από το άλλο $p(\mathcal{X}|\Theta) = \prod_{i=1}^{M} p(x_i|\Theta) = \prod_{i=1}^{M} \sum_{j=1}^{K} w_j p(x_i|\theta_j)$

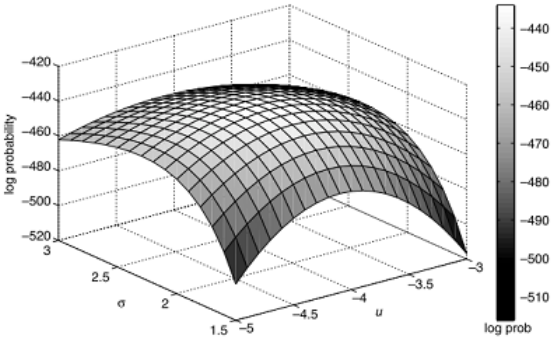
Εκτίμηση μέγιστης πιθανοφάνειας

- $p(\mathcal{X}|\Theta) = \prod_{i=1}^{M} p(x_i|\Theta)$
  - Τα δεδομένα $\mathcal{X}$ είναι γνωστά και αμετάβλητα.
  - Το ζητούμενο είναι οι τιμές των παραμέτρων Θ
- Συνάρτηση πιθανοφάνειας (likelihood) $\mathcal{L}(\Theta|\mathcal{X})$
  - Οι παράμετροι Θ ως συνάρτηση των δεδομένων $\mathcal{X}$
- Εκτίμηση μέγιστης πιθανοφάνειας
  - Βρες τις παραμέτρους εκείνες που μεγιστοποιούν την πιθανότητα εμφάνισης των συγκεκριμένων δεδομένων $\mathcal{X}$: $\hat{\theta} \in \left\{\max_{\theta \in \Theta} \mathcal{L}(\theta|\mathcal{X})\right\}$
  - Στην πράξη χρησιμοποιείται συχνά ο λογάριθμος της πιθανοφάνειας: $l(\Theta|\mathcal{X}) = \ln \mathcal{L}(\Theta|\mathcal{X})$

Εκτίμηση Μέγιστης Πιθανοφάνειας

- $M = 200$ πρότυπα εισόδου μιας διάστασης που προέρχονται από μία ($K = 1$) κανονική κατανομή, άγνωστης μέσης τιμής μ και τυπικής απόκλισης σ
  - $\mathrm{p}(\mathcal{X}|\Theta) = \mathcal{N}(\mu, \sigma) = \prod_{i=1}^{200} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- Εκφράζουμε την κατανομή υπό μορφή πιθανοφάνειας και παίρνουμε τον λογάριθμό της
  - $\ell(\Theta|\mathcal{X}) = -\sum_{i=1}^{200} \frac{(x_i-\mu)^2}{2\sigma^2} - 100 \ln 2\pi - 200 \ln \sigma$
- Λύνουμε τις εξισώσεις $\frac{\partial \ell}{\partial \mu} = 0$ και $\frac{\partial \ell}{\partial \sigma} = 0$
  - Η πιθανοφάνεια μεγιστοποιείται για $\hat{\mu} - 4.1$ και $\hat{\sigma} = 2.1$



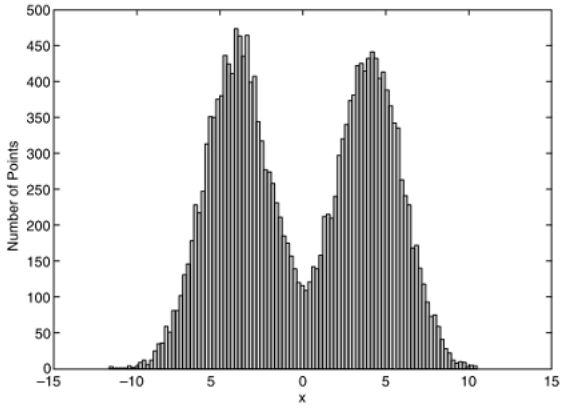


Αλγόριθμος Expectation-Maximization

- Στη γενική περίπτωση των μοντέλων μίξης, δεν γνωρίζουμε από ποια από τις K ανεξάρτητες κατανομές προέρχεται το πρότυπο x_i
- Βήματα αλγορίθμου
  1. Αρχικοποίηση των παραμέτρων του μοντέλου
  2. Επανάληψη
     1. Βήμα Αναμονής (Expectation Step): Υπολογισμός της πιθανότητας το πρότυπο x_i να προέρχεται από την j-οστή κατανομή $p(j|x_i, \Theta)$
     2. Βήμα Μεγιστοποίησης (Maximization Step): Χρησιμοποιώντας τα $p(j|x_i, \Theta)$, εκτίμησε τις παραμέτρους Θ που μεγιστοποιούν την πιθανοφάνεια
  3. Μέχρι τη σύγκληση (ή για ένα καθορισμένο αριθμό βημάτων)
- Ομοιότητα με αλγόριθμο k-μέσων
  - Αποτελεί ειδική περίπτωση του EM για κανονικές κατανομές σφαιρικού σχήματος με ίδιους πίνακες συνδιασποράς αλλά διαφορετικές μέσες τιμές
  - Βήμα Αναμονής ⇒ Ανάθεση κάθε προτύπου σε ένα κέντρο
  - Βήμα Μεγιστοποίησης ⇒ Υπολογισμός των νέων κέντρων

Αλγόριθμος Expectation-Maximization

- $M = 20.000$ πρότυπα, τα οποία υποθέτουμε ότι προέρχονται από 2 κανονικές κατανομές $\mathcal{N}_1(-2,2), \mathcal{N}_2(3,2)$ με ίσα βάρη ($w_1 = w_2 = \frac{1}{2}$)
- Βήμα Αναμονής: Χρήση κανόνα Bayes $p(j|x_i,\Theta) = \frac{w_j p(x_i|\theta_j)}{\sum_{j=1}^{2} w_j p(x_i|\theta_j)}$
- Βήμα Μεγιστοποίησης: $\mu_j = \sum_{i=1}^{20.000} x_i \frac{p(j|x_i,\Theta)}{\sum_{i=1}^{20.000} p(j|x_i,\Theta)}$
  - Στην περίπτωση της κανονικής κατανομής, η εκτίμηση για τη μέγιστη τιμή του μ είναι ένας ζυγισμένος μέσος όρος των δειγμάτων της



Iteration	μ_1	μ_2
0	−2.00	3.00
1	−3.74	4.10
2	−3.94	4.07
3	−3.97	4.04
4	−3.98	4.03
5	−3.98	4.03

$$p(\kappa\alpha\tau\alpha\nu o\mu\acute{\eta}\ \mathrm{j}|\mathrm{x}_i,\theta) = \frac{0.5p(\mathrm{x}_i|\theta_j)}{0.5p(\mathrm{x}_i|\theta_1)+0.5p(\mathrm{x}_i|\theta_2)}$$

Αλγόριθμος Expectation-Maximization

Έστω ότι ένα σημείο x=0 έρχεται στο βήμα 2.

P(x=0|θ1)=0.12 (από τη βασική εξίσωση της κατανομής Gauss) ενώ p(x=0|θ2)=0.06

Άρα p(κατανομή 1|x=0,Θ)=0.12/(0.12+0.06)=0.66

Ενώ p(κατανομή 2|x=0,Θ)=0.06/(0.12+0.06)=0.33

Δηλ το x=0 είναι 2 φορές πιο πιθανό να ανήκει στην κατανομή 1. (με βάση τις αρχικές εκτιμήσεις)

Το ίδιο γίνεται για όλα τα 20000 σημεία

Αλγόριθμος Expectation-Maximization

Maximization:

- Υπολογισμός μ1,μ2 (νέων)

$$\mu 1 = \sum_{i=1}^{20000} x_i \frac{p(\text{κατανομή } 1 \mid x_i, \Theta)}{\sum_{i=1}^{20000} p(\text{κατανομή } 1 \mid x_i, \Theta)}$$

$$\mu 2 = \sum_{i=1}^{20000} x_i \frac{p(\text{κατανομή } 2 \mid x_i, \Theta)}{\sum_{i=1}^{20000} p(\text{κατανομή } 2 \mid x_i, \Theta)}$$

Αλγόριθμος Expectation-Maximization

- Μετά από μερικές επαναλήψεις:

Επανάληψη	μ1	μ2
0	-2.00	3.00
1	-3.74	4.10
2	-3.94	4.07
3	-3.97	4.04
4	-3.98	4.03
5	-3.99	4.02
6	-3.99	4.02

Πλεονεκτήματα

Πιο γενικό μοντέλο από κ -μέσους

Εύρεση συστάδων διαφορετικού μεγέθους και σχήματος

Ευκολότερος χαρακτηρισμός των συστάδων από τις παραμέτρους του μοντέλου

Μειονεκτήματα

Μπορεί να είναι αργός

Μη-πρακτικός για μοντέλα με πολλές παραμέτρους

Δεν λειτουργεί σωστά σε συστάδες με λίγα πρότυπα

Επηρεάζεται από θόρυβο και έκτοπες τιμές