Εξόρυξη Δεδομένων
Δεύτερη Σειρά Ασκήσεων

Ταφλαμπάς Νικόλαος 4500

Άσκηση 1)

A)

Ας θεωρήσουμε σύνολο παρατηρήσεων X = {x1, x2, …, xn}, μεγέθους N, που ακολουθούν την εκθετική κατανομή:

$$L(\lambda, x) = \lambda e^{-\lambda x}$$

Για να βρούμε την παράμετρο λ, αξιοποιούμε την μεθοδολογία Maximum Likelihood Estimation (MLE). Αρχικά, υπολογίζουμε την ολική συνάρτηση πιθανότητα για όλες τις παρατηρήσεις μας:

$$L(\lambda, \chi) = \prod_{i=0}^{N} \lambda e^{-\lambda x_i} \; = \; \lambda^N e^{-\lambda \sum_{i=0}^{N} x_i}$$

Για την απλοποίηση της συναρτήσεις, και την διευκόλυνση μελλοντικών πράξεων, την λογαριθμούμε:

$$LogL(\lambda, \chi) = Nlog(\lambda) - \lambda \sum_{i=0}^{N} x_i$$

Χάρης της φύσης της λογαριθμικής εξίσωσης, το αποτέλεσμα που ψάχνουμε δεν θα αλλοιωθεί. Η παράμετρος λ παίρνει την μέγιστη της τιμή εκεί που η συνάρτηση φτάνει σε ένα τοπικό μέγιστο. Συνεπώς ψάχνουμε το σημείο οπού η παράγωγος της συναρτήσεις είναι 0, και η διπλή παράγωγος είναι αρνητική.

$$\frac{d}{d\lambda} LogL(\lambda, \chi) = 0 \leftrightarrow \frac{N}{\lambda} - \sum_{i=0}^{N} x_i \; = 0 \leftrightarrow$$

$$\lambda = \frac{N}{\sum_{i=0}^{N} x_i}$$

$$\frac{d^2}{d\lambda^2} LogL(\lambda, \chi) = \frac{-N}{\lambda^2} < 0$$

Αφού η 2η παράγωγος είναι πάντα αρνητική, τότε ξέρουμε ότι το λ έχει μέγιστη τιμή ίση με $\frac{N}{\sum_{i=0}^{N} x_i}$

B)

Αρχικά, θα ορίσουμε τις συναρτήσεις με τις οποίες δουλεύουμε.

$$L(x_i) = \prod_{i=0}^{N} (L_1(x_i)\pi_1 + L_2(x_i)\pi_2)$$

Έστω παραπάνω συνάρτηση πιθανοφάνειας για το μεικτό μοντέλο πιθανοτήτων μας, με παραμέτρους $\lambda_1, \lambda_2, \pi_1, \pi_2$. *Υπολογίζουμε την λογαριθμική συνάρτηση πιθανοφάνειας* $LL(x_i) = \ln(L(x_i))$ *για την διευκόλυνση πράξεων όπως και στην προηγούμενη άσκηση.*

Έπειτα, βάση του κανόνα του Bayes μπορώ να βρω την συνάρτηση πιθανότητας κατανομής K, δεδομένου σημείου x_i

$$P(L_k|x_i) = \frac{P(x_i|L_k) * P(L_k)}{\sum_{j=1}^{2} P(x_i|L_j) * P(L_j)} = \frac{P(x_i|L_k) * \pi_k}{\sum_{j=1}^{2} P(x_i|L_j) * \pi_j} = r_{ik}$$

Έχοντας τα παραπάνω, ευκολά μπορούμε να πούμε ότι οι μεταβλητές π_k *για k={1,2} μπορούν να υπολογιστούν αξιοποιώντας την συνεισφορά των σημείων* x_i. *Άρα η πιθανότητα P(*L_k*) ισούται με το άθροισμα της συνεισφοράς των σημείων στην κατανομή K, δια το πλήθος τους.*

$$\pi_k = \frac{\sum_{i=0}^{N} r_{ik}}{N}, \quad k = \{1,2\}$$

Όσο για τις μεταβλητές λ, ο αλγόριθμος EM αξιοποιεί παρόμοια μέθοδο με την MLE στο βήμα M. Λύνουμε ένα Maximization πρόβλημα της συνάρτησης LL ως προς λ:

$$\frac{dLL}{d\lambda_k} = 0 \rightarrow \cdots \rightarrow \sum_{i=0}^{N} P(L_k\,|x_i) = \sum_{i=0}^{N} P(L_k\,|x_i)x_i\lambda_k \rightarrow$$

$$\sum_{i=0}^{N} r_{ik} = \sum_{i=0}^{N} r_{ik}x_i\lambda_k \rightarrow \lambda_k = \frac{\sum_{i=0}^{N} r_{ik}}{\sum_{i=0}^{N} r_{ik}x_i}, k = \{1,2\}$$