ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

2η ΣΕΙΡΑ ΑΣΚΗΣΕΩΝ

Ερώτηση 1

//Με χρήση ενός (1) πάσου για παράταση έως 13/1

ΙΩΑΝΝΗΣ ΓΙΑΝΝΑΚΟΣ AM4970

ΕΡΩΤΗΣΗ 1

Α. Μία power-law κατανομή ορίζεται ως $P(X=x)=(a-1)x^{-α}$, όπου $α$ είναι ο εκθέτης της κατανομής. Σας δίνεται ένα σύνολο από παρατηρήσεις $X=\{x_1,\ldots,x_n\}$ που έχουν παραχθεί από μία power-law κατανομή. Χρησιμοποιήστε την Maximum Likelihood Estimation τεχνική που περιγράψαμε στην τάξη για να βρείτε τον εκθέτη της power-law κατανομής που ταιριάζει (fits) τα δεδομένα των παρατηρήσεων.

Δεδομένα:

- Power-Law κατανομή: $P(X = x) = (a - 1) * x^{-a}$, όπου α ο εκθέτης της κατανομής
- Σύνολο παρατηρήσεων: $X=\{x_1,\ldots,x_n\}$ το οποίο έχει παραχθεί από μια Power-law κατανομή

<u>ΛΥΣΗ</u>

Για να χρησιμοποιήσουμε την MLE για τον εκθέτη α της παραπάνω κατανομής, πρέπει να βρούμε την τιμή του α που μεγιστοποιεί την συνάρτηση MLE.

Η MLE για την Power-law κατανομή είναι

L(α) = $\prod_{i=1}^{n} P(X = x_i)$ => L(α) = $\prod_{i=1}^{n} (a - 1) * {x_i}^{-a}$

Στη συνέχεια, λογαριθμίζω την παραπάνω συνάρτηση για ευκολία:
LL(α)= $(a - 1) * \sum_{i=1}^{n} \log(x_i) - n * \sum_{i=1}^{n} \log{(a - 1)}$

Έπειτα, λύνω την εξίσωση $\frac{\partial LL}{\partial a} = 0$ προκειμένου να βρω την εκτιμώμενη τιμή του εκθέτη α της Power-law κατανομής που ταιριάζει τα δεδομένα των παρατηρήσεων.

Αρχικά, θα παραγωγίσουμε τον πρώτο όρο:

$$\frac{\partial}{\partial a}[(a - 1) * \sum_{i=1}^{n} \log{(x_i)}]$$

Άρα θα έχουμε: $\sum_{i=1}^{n} \log{(x_i)}$

Στη συνέχεια παραγωγίζουμε τον δεύτερο όρο:

$\frac{\partial}{\partial \alpha}[-n * \sum_{i=1}^{n} \log{(a-1)}]$

Άρα θα έχουμε: $-n * \sum_{i=1}^{n} \frac{1}{a-1}$

Τώρα συνενωνουμε τις παραγώγους και θέτουμε το αποτέλεσμα ίσο με μηδέν για λύσουμε την εξίσωση: $\sum_{i=1}^{n} \log(x_i) - n * \sum_{i=1}^{n} \frac{1}{a-1} = 0$

Για να λύσουμε την παραπάνω εξίσωση αθροίζουμε τον πρώτο όρο $\sum_{i=1}^{n} \log(x_i)$

Ο παραπάνω όρος δεν μπορεί να απλοποιηθεί περαιτέρω

Στη συνέχεια, αθροίζουμε τον δεύτερο όρο $n * \sum_{i=1}^{n} \frac{1}{a-1}$, συνεπώς εκτελούμε το άθροισμα $n * \frac{n}{a-1}$

Έπειτα, συνενώνουμε τους δύο όρους και θέτουμε το αποτέλεσμα ίσο με μηδέν:

$\sum_{i=1}^{n} \log(x_i) - n * \frac{n}{a-1} = 0 =>$

$\sum_{i=1}^{n} \log(x_i) = n * \frac{n}{a-1} => \ (\alpha - 1) = \frac{n}{\sum_{i=1}^{n} \log(x_i)} =>$

$$a = 1 + \frac{n}{\sum_{i=1}^{n} \log{(x_i)}}$$

Συνεπώς, συμπεραίνουμε ότι το α παίρνει τιμές μεγαλύτερες του 1 (α>1).

B. Υποθέστε ότι οι παρατηρήσεις X={$x1$,…,xn} έχουν παραχθεί από ένα μείγμα δύο power-law κατανομών, $L1$,$L2$, με παραμέτρους $a1$,$a2$, και πιθανότητες μίξης (mixture probabilities) $\pi1$,$\pi2$. Θα χρησιμοποιήσουμε τον EM αλγόριθμο για να υπολογίσουμε τις παραμέτρους Θ=($a1$,$a2$,$\pi1$,$\pi2$) του mixture μοντέλου, όπως κάναμε και για την περίπτωση της μίξης από Gaussian κατανομές. Στο Μ βήμα, υποθέτουμε ότι έχουμε τις πιθανότητες ανάθεσης $P(Lk|xi)$, για k=1,2 και i=1,…,n, και θέλουμε να υπολογίσουμε τις παραμέτρους Θ. Δώστε τις εξισώσεις για τα $a1$,$a2$,$\pi1$,$\pi2$ και τους υπολογισμούς με τους οποίους τις παρήγατε.

Δεδομένα:

- Power-law κατανομή L1: $P(X = x) = (a1 - 1) * x^{-a1}$, οπού α1 ο εκθέτης της κατανομής
- Power-law κατανομή L2: $P(X = x) = (a2 - 1) * x^{-a2}$, όπου α2 ο εκθέτης της κατανομής
- Σύνολο παρατηρήσεων: X={x1,...,xn} το οποίο έχει παραχθεί από μείγμα δύο power-law κατανομών L1 και L2 με α1, α2 παραμέτρους και π1, π2 πιθανότητες μίξης

<u>ΛΥΣΗ</u>

Αρχικά, με πιθανότητα π1 το μείγμα προέρχεται κυρίως από την κατανομή L1 και αντίστοιχα με π2 από την κατανομή π2.

Συνεπώς, π1+π2=1 και π1=P(L1), π2=P(L2)

Έπειτα, έχουμε $P(x_i|L1) = P(x_i|\theta1)$ και $P(x_i|L1) = P(x_i|\theta1)$, όπου θ1=α1 και θ2=α2

Άρα έχουμε $\Theta(\pi1, \pi2, \theta1, \theta2) \rightarrow \Theta(\pi1, \pi2, \alpha1, \alpha2)$

Για την τιμή θα έχουμε $P(x_i|\Theta) = \pi1 * P(x_i|\theta1) + \pi2 * P(x_i|\theta2)$

Άρα για όλες τις τιμές X=(x1,...,xn) έχουμε $P(X|\Theta) = \prod_{i=1}^{n} P(x_i|\Theta)$

Παρακάτω θα αναλύσουμε τα βήματα του EM Αλγόριθμου

EM Αλγόριθμος: Στο Ε-Βήμα δοθέντων των παραπάνω παραμέτρων Θ υπολογίζω τις πιθανότητες $P(L1|x_i)$ και $P(L2|x_i)$.

EM Αλγόριθμος: Στο Μ-Βήμα υπολογίζω τις παραμέτρους που μεγιστοποιούν την συνάρτηση

$$LL(\Theta) = \sum_{x_i} \log\left(\pi 1 * P(x_i|\theta 1) + \pi 2 * P(x_i|\theta 2)\right)$$

Συνεπώς, οι παράμετροι που θέλουμε να υπολογίσουμε είναι τα α1, α2 που υπολογίζονται από τους τύπους:

$$\alpha 1 = \frac{\sum_{i=1}^{n} P(L1|x_i) * \log(x_i)^2}{\sum_{i=1}^{n} P(L1|X_i) * \log(x_i)}$$

και

$$\alpha 2 = \frac{\sum_{i=1}^{n} P(L2|x_i) * \log(x_i)^2}{\sum_{i=1}^{n} P(L2|X_i) * \log(x_i)}$$

καθώς και τα π1, π2 που υπολογίζονται από τους τύπους

$$\pi 1 = \frac{\sum_{i=1}^{n} P(L1|x_i)}{n}$$

και

$$\pi 2 = \frac{\sum_{i=1}^{n} P(L2|x_i)}{n}$$