Εξόρυξη Δεδομένων

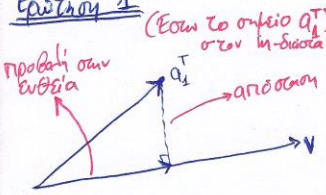
Δεύτερη Σειρά Ασκήσεων

Ευάγγελος Τζώρτζης

ΑΜ: 3088

Εξόρυξη Δεδομένων / 2η Σειρά Ασκήσεων / Ευάγγελος Τζώρτζης ΑΜ:3088

Ερώτημα 1



Λόγω του πυθαγορείου θεωρήματος για το σημείο που περιγράφεται από το a_1^T και κάποιας γραμμής πάνω στην οποία βρίσκεται το διάνυσμα v ισχύει ότι:

$$(a_1^T)^2 = (\text{προβολή})^2 + (\text{απόσταση από ευθεία})^2$$

$$(\text{απόσταση})^2 = a_{11}^2 + a_{12}^2 + a_{13}^2 + \dots + a_{1m}^2 - (\text{προβολή})^2 \quad (1)$$

Για να ελαχιστοποιηθεί το άθροισμα των τετραγώνων των αποστάσεων όλων των διανυσμάτων του A μπορούμε να ελαχιστοποιήσουμε το $\sum_{i=1}^{n} (a_{i1}^2 + a_{i2}^2 + \dots + a_{im}^2)$ (2) μείον το άθροισμα των τετραγώνων των προβολών των σημείων στη γραμμή. Όμως το (2) άθροισμα είναι σταθερό, άρα μπορούμε να μεγιστοποιήσουμε το άθροισμα των τετραγώνων των προβολών των σημείων στην ευθεία.

Το μήκος της προβολής της γραμμής a_i του πίνακα A πάνω στο μοναδιαίο διάνυσμα v είναι $|a_i \cdot v|$. Άρα το άθροισμα των τετραγώνων των προβολών είναι $|A \cdot v|^2$. Η βέλτιστη ευθεία είναι αυτή που μεγιστοποιεί το $|A \cdot v|^2$ και επομένως ελαχιστοποιεί το άθροισμα των τετραγώνων των αποστάσεων των σημείων από τη γραμμή.

Έχοντας τα παραπάνω στο νου, ορίζουμε τον πρώτο singular vector v_1 του A, το οποίο είναι διάνυσμα στήλη, ως τη καλύτερη ευθεία από την αρχή των αξόνων για τα n στοιχεία m διαστάσεων τα οποία είναι οι γραμμές του πίνακα A. Έτσι έχουμε $v_1 = \arg\max_{|v|=1} |Av|$.

Εξόρυξη Δεδομένων / 2η Σειρά Ασκήσεων / Ευάγγελος Τζιρτζής ΑΜ: 3088

Ερώτηση 2

Α. $P(X=x) = (a-1)\cdot x^{-a}$

$$P(a|X) = \frac{P(X|a)\cdot P(a)}{P(X)}$$

$$P(x_i|a) = (a-1)\cdot x_i^{-a}$$

Πιθανότητα όλων των σημείων (υποθέτουμε ανεξαρτησία των x)

$$P(X|a) = \prod_{i=1}^{n} P(x_i|a) = \prod_{i=1}^{n} (a-1)\cdot x_i^{-a} = L(a)$$

Log Likelihood:

$$LL(a) = \log\left[\prod_{i=1}^{n} (a-1)\, x_i^{-a}\right] = \sum_{i=1}^{n} \log\left(a x_i^{-a} - x_i^{-a}\right)$$

$$\frac{\partial LL(a)}{\partial (a)} = \sum_{i=1}^{n} \frac{1}{\frac{a}{x_i^a} - \frac{1}{x_i^a}} \cdot \left[\left(\frac{a}{x_i^a}\right)' - \left(\frac{1}{x_i^a}\right)'\right] =$$

$$= \sum_{i=1}^{n} \frac{1}{\frac{a}{x_i^a} - \frac{1}{x_i^a}} \cdot \left[\frac{x_i^a - a\cdot \ln(x)\cdot x_i^a}{x_i^{2a}} - \ln(x_i)\cdot x_i^{-a}\cdot(-1)\right] =$$

$$= \sum_{i=1}^{n} \frac{x_i^a}{a-1} \cdot \left[\frac{1 - a\cdot \ln(x_i)}{x_i^a} + \frac{\ln(x_i)}{x_i^a}\right] =$$

$$= \sum_{i=1}^{n} \frac{1 - a\ln(x_i) + \ln(x_i)}{a-1} = 0 \Rightarrow \frac{1}{a-1}\sum_{i=1}^{n}\left[1 - a\ln(x_i) + \ln(x_i)\right] = 0 \Leftrightarrow$$

$$\frac{1}{a-1}\left[n - a\sum_{i=1}^{n}\ln(x_i) + \sum_{i=1}^{n}\ln(x_i)\right] = 0 \Leftrightarrow \frac{1}{a-1}\left[n + (1-a)\sum_{i=1}^{n}\ln(x_i)\right] = 0 \Leftrightarrow$$

$$\frac{n}{a-1} - \sum_{i=1}^{n}\ln(x_i) = 0 \Leftrightarrow \frac{n}{a-1} = \sum_{i=1}^{n}\ln(x_i) \Leftrightarrow \boxed{a = 1 + n\cdot\left[\sum_{i=1}^{n}\ln(x_i)\right]^{-1}}$$

Εξόρυξη Δεδομένων / 2η Σειρά Ασκήσεων / Ευάγγελος Γιώργης ΑΜ: 3088

Ερώτηση 2

Β. $P(X_i|L1) = (a_1-1)X_i^{-a_1}$, $P(X_i|L2) = (a_2-1)X_i^{-a_2}$

$$P(X_i) = \Pi_1 \cdot P(X_i|L1) + \Pi_2 \cdot P(X_i|L2)$$

$$\frac{\partial P(X_i|L1)}{\partial a_1} = \frac{\ln(X_i) - \ln(X_i)\cdot a_1 + 1}{X_i^{a_1}}$$

$$P(L1|X_i) = \frac{\Pi_1 \cdot P(X_i|L1)}{P(X_i)} = \gamma_{iL1} \quad , \quad P(L2|X_i) = \gamma_{iL2}$$

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\Pi_1 \cdot P(X_i|L1) + \Pi_2 \cdot P(X_i|L2)\right) \quad , \theta = (a_1, a_2, \Pi_1, \Pi_2)$$

Για το a_1 θα υπολογίσω το $\frac{\partial LL(\theta)}{\partial a_1} = 0$, και για το a_2 $\frac{\partial LL(\theta)}{\partial a_2}$

$$\frac{\partial LL(\theta)}{\partial a_1} = \sum_{i=1}^{n} \frac{1}{\Pi_1 \cdot P(X_i|L1) + \Pi_2 \cdot P(X_i|L2)} \cdot \Pi_1 \cdot \frac{\partial P(X_i|L1)}{\partial a_1} =$$

$$= \sum_{i=1}^{n} \frac{\Pi_1}{P(X_i)} \cdot \left(\frac{\ln(X_i) - \ln(X_i)\cdot a_1 + 1}{X_i^{a_1}}\right) = \sum_{i=1}^{n} \frac{\Pi_1}{P(X_i)} \cdot \left(\ln(X_i)\cdot(a_1-1) + 1\right)\cdot X_i^{-a_1} =$$

$$= \sum_{i=1}^{n} \frac{\Pi_1}{P(X_i)} \cdot \left[-\ln(X_i)\cdot \overbrace{(a_1-1)\cdot X_i^{-a_1}}^{P(X_i|L1)} + X_i^{-a_1}\right] = \sum_{i=1}^{n} \left[\frac{\Pi_1 \cdot P(X_i|L1)}{P(X_i)} \cdot \left(-\ln(X_i)\right)\right] + \sum_{i=1}^{n} \frac{X_i^{-a_1}\cdot \Pi_1}{P(X_i)} =$$

$$= \sum_{i=1}^{n} -\ln(X_i)\cdot \gamma_{iL1} + \sum_{i=1}^{n} \frac{\overbrace{(a_1-1)\ X_i^{-a_1}}^{P(X_i|L1)}\cdot \Pi_1}{\underbrace{(a_1-1)}_{\gamma_{iL1}}\ P(X_i)} = 0 \iff$$

$$\Rightarrow \sum_{i=1}^{n} \frac{\gamma_{iL1}}{(a_1-1)} = \sum_{i=1}^{n} \ln(X_i)\cdot\gamma_{iL1} \Rightarrow \frac{1}{(a_1-1)} = \frac{\sum_{i=1}^{n} \ln(X_i)\cdot\gamma_{iL1}}{\sum_{i=1}^{n}\gamma_{iL1}} \Rightarrow$$

$$\Rightarrow a_1 - 1 = \frac{\sum_{i=1}^{n}\gamma_{iL1}}{\sum_{i=1}^{n}\ln(X_i)\cdot\gamma_{iL1}} \Rightarrow \boxed{a_1 = 1 + \frac{\sum_{i=1}^{n}\gamma_{iL1}}{\sum_{i=1}^{n}\ln(X_i)\cdot\gamma_{iL1}}}$$

παρομοίως $\boxed{a_2 = 1 + \frac{\sum_{i=1}^{n}\gamma_{iL2}}{\sum_{i=1}^{n}\ln(X_i)\cdot\gamma_{iL2}}}$

Εξόρυξη Δεδομένων / 2η Σειρά Ασκήσεων / Ευάγγελος Τζίρτζης ΑΜ: 3088

Ερώτηση 2

Β. (συνέχεια)

Υπολογισμός των Π_1, Π_2 με χρήση πολλαπλασιαστών Lagrange

• Ισχύει $\Pi_1 + \Pi_2 = 1$

$$f(\theta, \lambda) = LL(\theta) - \lambda(\Pi_1 + \Pi_2 - 1) \qquad \frac{\partial f}{\partial \lambda} = 0 \Rightarrow \Pi_1 + \Pi_2 = 1$$

$$\frac{\partial f}{\partial \Pi_1} = \sum_{i=1}^{n} \frac{1}{P(X_i)} \cdot P(X_i | L_1) - \lambda = 0 \Rightarrow$$

$$\lambda = \sum_{i=1}^{n} \frac{P(X_i | L_1)}{P(X_i)} \Rightarrow \Pi_1 \cdot \lambda = \sum_{i=1}^{n} \frac{\Pi_1 \cdot P(X_i | L_1)}{P(X_i)} = \sum_{i=1}^{n} P(L_1 | X_i) \quad (1)$$

αντίστοιχα: $\Pi_2 \cdot \lambda = \sum_{i=1}^{n} P(L_2 | X_i)$ (2)

Από (1) και (2):

$$(\Pi_1 + \Pi_2)\lambda = \sum_{i=1}^{n} [P(L_2 | X_i) + P(L_1 | X_i)] = n \Rightarrow \boxed{\lambda = n}$$

(1 πάνω στο $\Pi_1 + \Pi_2$, 1 πάνω στο άθροισμα)

άρα:

$$\boxed{\Pi_1 = \frac{1}{n} \sum_{i=1}^{n} P(L_1 | X_i)}$$

$$\boxed{\Pi_2 = \frac{1}{n} \sum_{i=1}^{n} P(L_2 | X_i)}$$