Καραγιαννίδης Χρήστος 4375

Άσκηση 1

Η συνάρτηση πιθανότητας για μια εκθετική κατανομή με παράμετρο λ και ένα σύνολο παρατηρήσεων X = {x1,...,xn} δίνεται από :

$L(\lambda \mid X) = \Pi(\lambda * e^{(-\lambda xi)})$, $i \in [1,n]$

Λογαριθμίζοντας την συνάρτηση πιθανότητας παίρνουμε :

$\ln(L(\lambda \mid X)) = \Sigma(\ln(\lambda) - \lambda xi)$, $i \in [1,n]$

Μετά παίρνουμε την παράγωγο του λογαρίθμου της συνάρτησης πιθανότητας ως προς το λ και το βάζουμε ίσο με 0 για να βρούμε την maximum likelihood estimate (MLE) :

$d/d\lambda \ln(L(\lambda \mid X)) = (n/\lambda) - \Sigma(xi) = 0$

Λύνουμε ως προς λ:

$\lambda = n / \Sigma(xi)$

Άρα η παράμετρος που ταιριάζει στα δεδομένα των παρατηρήσεων X χρησιμοποιώντας την Maximum Likelihood Estimation τεχνική είναι η $\lambda = n / \Sigma(xi)$

Β)

Ο ΕΜ αλγόριθμος για την εκτίμηση των παραμέτρων ενός mixture model είναι μια επαναληπτική διαδικασία που αποτελείται από 2 βήματα το βήμα Ε και το βήμα Μ.

Στο βήμα Ε, χρησιμοποιούμε τις τρέχουσες εκτιμήσεις των παραμέτρων για να υπολογίσουμε τις πιθανότητες ανάθεσης P(Lk|xi) για κάθε παρατήρηση xi και κάθε μέρος του mixture model Lk. Οι πιθανότητες ανάθεσης μας δίνουν την πιθανότητα κάθε παρατήρηση xi να παράχθηκε από την εκθετική κατανομή Lk.

Στο βήμα Μ , χρησιμοποιούμε τις πιθανότητες ανάθεσης για να εκτιμήσουμε τις νέες τιμές των παραμέτρων λ1, λ2, π1, π2. Χρησιμοποιούμε τις εξής συναρτήσεις:

1) Αφού οι παρατηρήσεις xi έχουν παραχθεί από μια εκθετική κατανομή με παράμετρο λ1, η αναμενόμενη τιμή του xi για το L1 είναι 1/λ1. Επομένως μπορούμε να εκτιμήσουμε το λ1 παίρνοντας τον σταθμισμένο μέσο όρο των παρατηρήσεων xi, όπου τα βάρη είναι οι πιθανότητες ανάθεσης P(L1|xi). Αυτό έχει ως αποτέλεσμα την συνάρτηση **$\lambda 1 = \Sigma (P(L1|xi) * xi) / \Sigma (P(L1|xi))$**

2) Με τον ίδιο τρόπο μπορούμε να εκτιμήσουμε το λ2 παίρνοντας τον σταθμισμένο μέσο όρο των παρατηρήσεων xi όπου τα βάρη είναι η πιθανότητα ανάθεσης P(L2|xi), το οποίο μας δίνει την εξίσωση **λ2 = Σ (P(L2|xi) * xi) / Σ (P(L2|xi))**
3) Η πιθανότητα μίξης π1 αναπαριστά το ποσοστό των παρατηρήσεων που παράχθηκαν από το L1. Συνεπώς μπορούμε να εκτιμήσουμε το π1 παίρνοντας το άθροισμα των πιθανοτήτων ανάθεσης (όσον παράχθηκαν από το L1) P(L1|xi) δια τον αριθμό όλων των παρατηρήσεων n. Άρα έχουμε **π1 = Σ (P(L1|xi)) / n**
4) Παρομοίως για το π2 παίρνουμε στον αριθμητή το άθροισμα όλων των πιθανοτήτων ανάθεσης (όσων παράχθηκαν από το L2) P(L1|xi) δια τον αριθμό όλων των παρατηρήσεων n. Άρα έχουμε **π2 = Σ (P(L2|xi)) / n**

Άσκηση 2

Βήμα 3:

Παρατηρούμε ότι οι 2 τιμές RMSE είναι αρχικά πολύ κοντά η μια στην άλλη και επίσης και οι 2 δείχνουν ότι οι αλγόριθμοι στους οποίους αντιστοιχούν (UA , BA) είναι πολύ καλοί στο να προβλέπουν τις τιμές των test δεδομένων.

Βήμα 4:

Η τιμή για την οποία ο αλγόριθμος πετυχαίνει το ελάχιστο RMSE είναι το κ=100 με RMSE = 3.2815251570587254. Παρατηρούμε ότι όσο μεγαλώνει το k τόσο μειώνεται το RMSE διότι μια μεγαλύτερη τιμή του K οδηγεί σε μια πιο πολύπλοκη αναπαράσταση των δεδομένων και κατά συνέπεια μια καλύτερη εκτίμηση του αρχικού πίνακα.

Βήμα 7:

Παρατηρούμε ότι το RMSE των UA και BA ήταν με διαφορά καλύτερο από τους άλλους αλγορίθμους, ακόμα και από την καλύτερη εκτέλεση του SVD (για κ=100) και σίγουρα ήταν πιο αποδοτικό από άποψη χρόνου (αυτό εν μέρει μπορεί να μην ευθύνεται στην πολυπλοκότητα του αλγορίθμου αλλά στην δικιά μου υλοποίηση αυτού). Επίσης το ότι όσο μεγαλώναμε το κ στον SVD αλγόριθμο μας μείωνε το RMSE δε σημαίνει ότι θα συνεχιζόταν αυτή η μείωση για κ>>100 καθώς θα είχαμε φτάσει σε ένα σημείο που θα είχαμε overfitting.

Άσκηση 3

Σημείωση : έκανα filtering το dataframe με δυο διαφορετικούς τρόπους και οι 2 όμως μου έβγαζαν 1073 rows αντί για 951 που θα έπρεπε. Έκανα print τα 'categories' (τις λέξεις 'Restaurant', 'Japanese', 'Italian', 'Burgers') και καμία γραμμή από τις 1073 δεν είχε παραπάνω από 2 λέξεις (Restaurant + άλλη μια) και καμία δεν είχε λιγότερες από 2. Οπότε τα επόμενα αποτελέσματα είναι σίγουρα ως ένα βαθμό επηρεασμένα από το διαφορετικό dataframe.

4) Η καλύτερη τιμή για τον αριθμό των clusters είναι κ=9 καθώς το SSE μετράει την απόσταση κάθε δεδομένου από τον cluster στον οποίο ανατέθηκε άρα όσο μικρότερο είναι το SSE τόσο καλύτερο το clustering. Ταυτόχρονα το Silhouette Coefficient το οποίο μετράει την ομοιότητα των δεδομένων σε ένα cluster σε σχέση με αλλά clusters στο k=9 φτάνει τη μέγιστη τιμή του ~0.22 (με πεδίο τιμών [-1,1]) . Ο λόγος που δεν παίρνουμε το k=10 είναι γιατί παρόλο που μειώνεται λίγο ακόμα το SSE μειώνεται σχεδόν 10% και το Silhouette Coefficient.