ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Ηλεκτρολόγων Μηχ. Και Μηχ. Υπολογιστών

Στατιστική Μοντελοποίηση και Αναγνώριση Προτύπων (ΤΗΛ311)

Φυλλάδιο Ασκήσεων 3

Οδηγίες:

1. Σας παρακαλώ να σεβαστείτε τον παρακάτω κώδικα τιμής τον οποίον θα θεωρηθεί ότι προσυπογράφετε μαζί με τη συμμετοχή σας στο μάθημα και τις εργασίες του:
   a) *Οι απαντήσεις στις εργασίες, τα quiz και τις εξετάσεις, ο κώδικας και γενικά οτιδήποτε αφορά τις εργασίες θα είναι προϊόν δικής μου δουλειάς.*
   b) *Δεν θα διαθέσω κώδικα, απαντήσεις και εργασίες μου σε κανέναν άλλο.*
   c) *Δεν θα εμπλακώ σε άλλες ενέργειες με τις οποίες ανέντιμα θα βελτιώνω τα αποτελέσματα μου ή ανέντιμα θα αλλάζω τα αποτελέσματα άλλων.*
2. Η εργασία είναι ατομική
3. Ημερομηνία παράδοσης: **Κυριακή, 24/5/2020 στις 23:00**
4. **Παραδοτέα:** α) Κώδικας και β) Αναφορά με τις απαντήσεις, παρατηρήσεις, πειράματα, αποτελέσματα και οδηγίες χρήσης του κώδικα.

Θέμα 1: Κατηγοριοποίηση κειμένου με k-NN

Η κατηγοριοποίηση κειμένου είναι η διαδικασία κατάταξης κειμένων φυσικής γλώσσας σε ένα προκαθορισμένο αριθμό κατηγοριών (π.χ. με βάση το θέμα). Στην άσκηση αυτή, η βάση δεδομένων που θα χρησιμοποιηθεί είναι ένα υποσύνολο της WebKB, και περιέχει ιστοσελίδες από 4 πανεπιστήμια στην Αμερική. Οι ιστοσελίδες ταξινομήθηκαν από ανθρώπους σε 4 κατηγορίες: *student*, *course*, *faculty*, *project*.

Στον φάκελο "**exercise3_1/data**" βρίσκονται 4 αρχεία (*course_TDM.csv, faculty_TDM.csv, project_TDM.csv, student_TDM.csv*) ένα για κάθε μία από τις παραπάνω κατηγορίες. Το κάθε ένα από αυτά περιέχει στην πρώτη γραμμή το λεξικό των κειμένων (λέξεις σε αλφαβητική σειρά που έχουν βρεθεί στα κείμενα) και στις υπόλοιπες γραμμές στην πρώτη θέση το όνομα της ιστοσελίδας το οποίο ακολουθείται από αριθμητικές τιμές χωρισμένες με το σύμβολο «;». Επειδή τα αρχικά δείγματα (ιστοσελίδες) είναι κείμενα, για να μπορέσουν να χρησιμοποιηθούν από ένα σύστημα μηχανικής μάθησης χρειάζεται να γίνει κατάλληλη επεξεργασία και εξαγωγή αριθμητικών χαρακτηριστικών. Στα αρχεία που σας δίνονται τα δείγματα περιγράφονται με τη μορφή ενός term-document πίνακα. Σε έναν term-document πίνακα $M \in \mathbb{R}^{nD \times nT}$, όπου *nD* είναι το πλήθος των γραμμών και *nT* το πλήθος των στηλών, η πληροφορία για τα κείμενα αναπαρίσταται ως εξής: κάθε γραμμή $i, 1 \leq i \leq nD$, αναπαριστά ένα κείμενο ενώ κάθε στήλη $j, 1 \leq j \leq nT$, αναπαριστά μια λέξη από το λεξικό των κειμένων. Το στοιχείο m_{ij} του M είναι η συχνότητα εμφάνισης (ακέραιος αριθμός) του όρου j στο κείμενο i την οποία για ευκολία συμβολίζουμε και ως f_{ij} , όπου $f_{ij=}m_{ij}$.

1) Γενικά, οι πίνακες term-document είναι αραιοί (sparse) γιατί έχουν πολλά μηδενικά στοιχεία και έχουν μεγάλο μέγεθος γιατί συνήθως ο αριθμός των λέξεων στα κείμενα είναι μεγάλος. Επιπλέον δεν είναι όλες οι λέξεις σημαντικές καθώς ορισμένες λέξεις (π.χ. συζευκτικά, άρθρα κ.α.) δεν μας δίνουν πληροφορία για την κατηγορία ενός κειμένου. Ένας τρόπος για να περιορίσουμε τις λέξεις αξιοποιώντας μόνο όσες είναι πιο σημαντικές είναι να χρησιμοποιήσουμε το μέγεθος της εντροπίας της κάθε λέξης και με βάση αυτή να επιλέξουμε τις NW πιο σημαντικές λέξεις:
   Εντροπία: Αν ορίσουμε την ποσότητα $p_{ij} = f_{ij} / \sum_{i=1}^{nD} f_{ij}$ ως την κανονικοποιημένη συχνότητα της λέξης τότε η εντροπία της j-οστής λέξης ορίζεται ως:

$$e_j = 1 + \frac{\sum_{i=1}^{n} p_{ij} \log(p_{ij})}{\log(nD)}$$

Συμπληρώστε τον απαραίτητο κώδικα στο αρχείο `Entropy_Transformation.m` ώστε να επιστρέφει την εντροπία της κάθε λέξης και με βάση αυτή επιλέξτε τις σημαντικότερες **300 λέξεις** για να χρησιμοποιήσετε ως διάνυσμα χαρακτηριστικών. Ακολούθως χρησιμοποιώντας τις συχνότητες από τον term-document πίνακα θα δημιουργήσετε διανύσματα χαρακτηριστικών, ίδιου μεγέθους με τις λέξεις που επιλέξατε, και με τιμές το TF-IDF που τους αντιστοιχεί. Το TF-IDF χρησιμοποιεί 2 όρους για να υπολογιστεί. Ο πρώτος είναι το Term Frequency (TF) που εκφράζει την κανονικοποιημένη συχνότητα εμφάνισης ενός όρου (λέξης) σε ένα κείμενο Ο δεύτερος όρος είναι το Inverse Document Frequency (IDF) ο οποίος ποσοτικοποιεί το πόσο σημαντικός είναι ένας όρος για την διάκριση των κειμένων. Για τον υπολογισμό του TF-IDF θα χρησιμοποιήσετε την έτοιμη συνάρτηση **`tfidf`** που σας δίνεται. Περισσότερα για το TFIDF μπορείτε να βρείτε εδώ: http://www.tfidf.com/

2) Υλοποιήστε τον k-NN αλγόριθμο στο αρχείο (`KNN_classify`) ο οποίος θα κάνει ταξινόμηση σε δεδομένα ελέγχου. Η συνάρτηση θα μπορεί να δέχεται στην είσοδο τις εξής παραμέτρους:
   a) Τον αριθμό των κοντινότερων γειτόνων (θα δοκιμάσετε να τρέξετε για $k = 1, 3, 5, 10$.
   b) Τον τύπο της μετρικής απόστασης. Δοκιμάστε να χρησιμοποιήσετε τις ακόλουθες μετρικές απόστασης, συμπληρώνοντας τον απαραίτητο κώδικα στα αντίστοιχα αρχεία.
      - Ευκλείδεια απόσταση: $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} (x_i - y_i)^2$
      - Vector product (cosine similarity):

$$\cos(\theta) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||} = \frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} (x_i)^2} \sqrt{\sum_{i=1}^{p} (y_i)^2}}$$

3) Ελέγξτε την επίδοση του συστήματος σας σε κάθε περίπτωση εφαρμόζοντας K-Fold Cross-validation με 5 folds. Σχολιάστε τα αποτελέσματα.

Θέμα 2: K-means clustering

Ο K-means είναι ένας αλγόριθμος που ομαδοποιεί όμοια δεδομένα. Σε αυτή την άσκηση θα τον υλοποιήσετε και θα τον χρησιμοποιήσετε για να συμπιέσετε μια εικόνα. Θα αρχίσετε με ένα δισδιάστατο, 2D, σύνολο δεδομένων για να καταλάβετε πως λειτουργεί ο K-means. Έστω ότι έχουμε ένα σύνολο $X = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$ με m n-διάστατα δείγματα ($x^{(i)} \in \mathbb{R}^n$), τα οποία θέλουμε να τα ομαδοποιήσουμε σε K-κλάσεις. Ο K-means είναι ένας επαναληπτικός αλγόριθμος. Αρχικά θέτει τυχαίες τιμές στα κέντρα των κλάσεων, και στην συνέχεια βελτιώνει την αρχική του επιλογή τοποθετώντας τα παραδείγματα στην κλάση με την μικρότερη απόσταση μεταξύ του παραδείγματος και του κέντρου της κλάσης και ξαναϋπολογίζοντας τα κέντρα των κλάσεων.

Συγκεκριμένα ο αλγόριθμος K-means όπως θα τον βρείτε στο αρχείο `ex3_kmeans.m` συνοψίζεται ως εξής:

```
% Initialize centroids
centroids = kMeansInitCentroids(X, K);
for iter = 1:iterations
    % Cluster assignment step: Assign each data point to the
    % closest centroid. idx(i) corresponds to cˆ(i), the index
    % of the centroid assigned to example i
    idx = findClosestCentroids(X, centroids);
    % Move centroid step: Compute means based on centroid
    % assignments
    centroids = computeMeans(X, idx, K);
end
```

Για να τρέξει στο Matlab/Octave το παραπάνω script θα πρέπει να συμπληρώσετε με δικό σας κώδικα τα επόμενα μέρη του προγράμματος.

a) Να συμπληρώσετε την συνάρτηση `findClosestCentroids.m` η οποία υπολογίζει την κοντινότερη απόσταση $c^{(i)} := j$ that minimizes $\left\|x^{(i)} - \mu_j\right\|^2$ από τα κέντρα των κλάσεων, όπου $c^{(i)}$ είναι ο δείκτης του κοντινότερου κέντρου στο $x^{(i)}$, και μ_j είναι το διάνυσμα τιμών (συντεταγμένων) του κέντρου j. Το $c^{(i)}$ αντιστοιχεί στο `idx(i)` του παραπάνω κώδικα.

b) Να συμπληρώσετε την συνάρτηση `computeCentroids.m` η οποία υπολογίζει τα κέντρα των κλάσεων ως εξής:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x^{(i)}$$

όπου C_k είναι το σύνολο των παραδειγμάτων που έχουν αντιστοιχηθεί στην κλάση k.

c) Συμπληρώστε την εντολή `X_recovered = …` στο Matlab/Octave script `ex3_kmean.m` και δείτε την συμπιεσμένη εικόνα για διάφορες τιμές του K (αριθμός κλάσεων).

Θέμα 3: GMMs και εκτίμηση με τον αλγόριθμο Expectation Maximization

Στην άσκηση αυτή θα υλοποιήσετε τον αλγόριθμο Expectation Maximization (EM) για να εκπαιδεύσετε τις παραμέτρους ενός μοντέλου Gaussian Mixture (GMM). Τα βήματα του αλγορίθμου υπάρχουν στις σελίδες 438 και 439 του βιβλίου του Bishop και στις σημειώσεις του φροντιστηρίου.

https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Επίσης μπορείτε να δείτε τα βήματα του αλγορίθμου στο άρθρο του Bilmes που υπάρχει στα αρχεία της άσκησης.

Συμπληρώστε τα αρχεία κώδικα της άσκησης και δοκιμάστε τον αλγόριθμο στα λουλούδια Fisher Iris. Γνωρίζοντας ότι τα παραδείγματα 1 έως 50 ανήκουν στην κλάση «setosa» τα παραδείγματα 51 έως 100 στην κλάση «versicolor» και τα παραδείγματα 101 έως 150 στην κλάση «virginica», πόσα λουλούδια ταξινομούνται σε λάθος ομάδα με το GMM και πόσα με τον kmeans;