Μηχανική Μάθηση (Αναγνώριση Προτύπων)
Απαλλακτική Εργασία σε Python

Ημερομηνία Παράδοσης: **23/08/2020, 17.00μμ μέσω eclass**

Σε αυτή την εργασία καλείστε να υλοποιήσετε σε Python αλγορίθμους μηχανικής μάθησης και να ερμηνεύσετε τα αποτελέσματά τους σε πραγματικά δεδομένα (εικόνες) και συνθετικά δεδομένα. Προτείνετε να χρησιμοποιήσετε το Google Colab. Μπορείτε να χρησιμοποιήσετε οποιαδήποτε βιβλιοθήκη της python επιθυμείτε π.χ. scikit learn, pandas, OpenCV κτλ.

Η εργασία είναι **ατομική** και αποτελείται από δύο ερωτήματα τα οποία είναι βαθμολογικά ισοδύναμα. Ενδέχεται να πραγματοποιηθεί και προφορική εξέταση. Σε περίπτωση που διαπιστωθεί αντιγραφή θα μηδενιστούν όλα τα εμπλεκόμενα μέρη. Ωστόσο, μπορείτε να συμβουλευτείτε και να χρησιμοποιήσετε οποιοδήποτε υλικό ή/και κώδικα που είναι διαθέσιμος στο διαδίκτυο, αρκεί να αναφέρεται σωστά τη πηγή ή/και το σύνδεσμο στην ιστοσελίδα που αντλήσατε πληροφορίες.

Θα υποβάλετε **ένα μόνο αρχείο Notebook IPython (Jupiter notebook) στο eclass,** ακολουθώντας την εξής σύβαση ονομασίας για το αρχείο σας: *επώνυμο_ΑΜ.ipynb* Τόσο ο κώδικας όσο και οι απαντήσεις σας στις ερωτήσεις κατανόησης/ερμηνείας πρέπει να είναι ενσωματωμένα στο ίδιο IPython notebook. Μπορείτε να χρησιμοποιήσετε κελιά επικεφαλίδας για να οργανώσετε περαιτέρω το έγγραφό σας.

[Ερώτημα 1 – *Προ-επεξεργασία, μείωση διαστάσεων, οπτικοποίηση και ταξινόμηση εικόνων*]

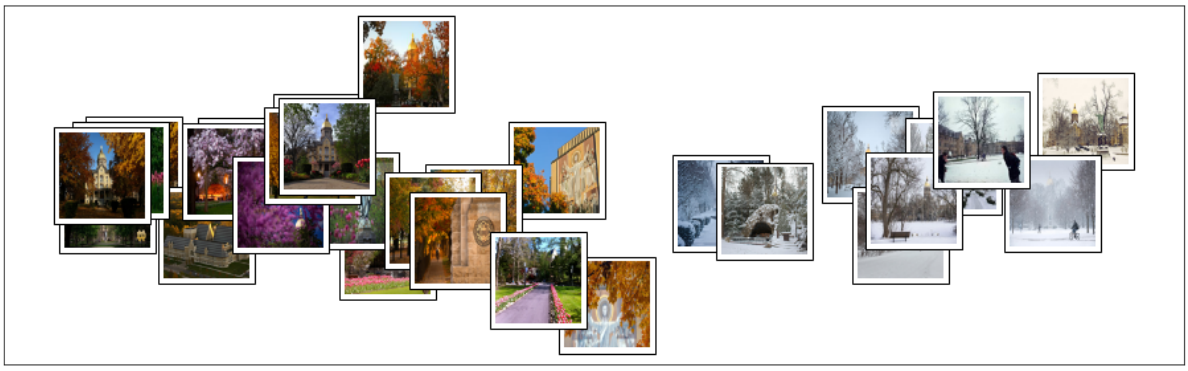
Δεδομένα: Το σύνολο δεδομένων αποτελείται από 30 έγχρωμες RGB εικόνες που καταγράφουν τοπία κατά την άνοιξη (spring), το φθινόπωρο (fall) και το χειμώνα (winter) (10 εικόνες για κάθε εποχή). Το πρώτο γράμμα στο όνομα του αρχείου της κάθε εικόνας προσδιορίζει την εποχή κατά την οποία καταγράφηκε η εικόνα, π.χ. η εικόνα F1.jpg καταγράφηκε το φθινόπωρο (fall) ενώ η εικόνα W10.jpg καταγράφηκε το χειμώνα (winter). Συνεπώς η ονοματολογία των αρχείων καθορίζει πλήρως την κατηγορία στην οποία ανήκει κάθε εικόνα. Οι εικόνες αποτελούνται από διαφορετικά σε πλήθος εικονοστοιχεία (pixels). Κάθε pixel αποτελείται από τρεις τιμές χρώματος που κυμαίνονται μεταξύ 0 και 255 και καθορίζουν την ένταση φωτεινότητας του κόκκινου, του πράσινου και του μπλε αντίστοιχα σε κάθε σημείο της εικόνας. Τα δεδομένα είναι διαθέσιμα στο αρχείο images.zip στο eclass.

Ζητούμενα:

1) Να γράψετε μία συνάρτηση loadImages(path) η οποία παίρνει ως είσοδο το path στο οποίο βρίσκεται ο φάκελος των εικόνων π.χ. loadImages("C:/images"), διαβάζει τις εικόνες, τις μετατρέπει σε διάσταση 100 x 100 pixels και επιστέφει έναν πίνακα δεδομένων 30 στηλών, όπου κάθε εικόνα αναπαρίσταται ως διάνυσμα στήλη. Η

συνάρτηση επιστέφει επίσης τις κατηγορίες (labels) στις οποίες ανήκουν οι διαφορετικές εικόνες κωδικοποιημένες με ακεραίους (π.χ. 0 για φωτογραφίες που καταγράφηκαν το χειμώνα, 1 για τις φωτογραφίες που καταγράφηκαν το φθινόπωρο και 2 για αυτές που καταγράφηκαν την άνοιξη).

2) Να γραφεί συνάρτηση PCA_ImageSpaceVisualization(X) η οποία παίρνει ως είσοδο τον πίνακα δεδομένων, υπολογίζει τις δύο πρώτες κύριες συνιστώσες (principal components) των δεδομένων και προβάλει τα δεδομένα στις δύο πρώτες κύριες συνιστώσες. Η συνάρτηση επιστέφει ένα plot στο οποίο οπτικοποιούνται οι εικόνες στο δυσδιάστατο χώρο που προκύπτει από τη προβολή των δεδομένων στις δύο πρώτες κύριες συνιστώσες. Το plot αναμένεται να είναι της μορφής:



2.1 Τι σημαίνει όταν εικόνες βρίσκονται κοντά σε αυτό το χώρο δύο διαστάσεων που απεικονίζεται στο παραπάνω plot; Τι σημαίνει όταν εικόνες απέχουν πολύ; Μπορούμε να γενικεύσουμε αυτά τα συμπεράσματα για τον αρχικό χώρο των εικόνων ο οποίος είναι πολύ μεγάλης διάστασης;
2.2 Οι εικόνες που αντιστοιχούν σε μία από τις εποχές τείνουν να ομαδοποιούνται πιο κοντά από ότι οι υπόλοιπες; Γιατί συμβαίνει αυτό;

3) Να συγκρίνετε την ακρίβεια (accuracy) του ταξινομητή πλησιέστερου γείτονα (1-NN) και της γραμμικής μηχανής διανυσμάτων υποστήριξης (linear support vector machine - SVM) στο πρόβλημα της αναγνώρισης της εποχής κατά την οποία καταγράφηκε μια εικόνα. Με άλλα λόγια να συγκρίνετε την επίδοση (ως προς την ακρίβεια ταξινόμησης) των παραπάνω ταξινομητών στην ταξινόμηση των δεδομένων εικόνων στις κατηγορίες χειμώνας, άνοιξη και φθινόπωρο.

   Καλείστε να αντιμετωπίσετε το πρόβλημα ταξινόμησης χρησιμοποιώντας 1) τις αρχικές μεγάλης διάστασης εικόνες σε μορφή διανύσματος και 2) χαρακτηριστικά χαμηλής διάστασης που θα εξάγετε μέσω της PCA.

   3.1 Να ορίσετε μαθηματικά το μέτρο της ακρίβειας ταξινόμησης (classification accuracy).
   3.2 Χρησιμοποιείστε 5-fold cross validation και αναφέρετε τη μέση ακρίβεια ταξινόμησης για τους δύο ταξινομητές τόσο για τα δεδομένα μεγάλης διάστασης όσο και για τα χαρακτηριστικά χαμηλής διάστασης.
   3.3 Πώς θα προσδιορίσετε τη διάσταση των χαρακτηριστικών που θα εξάγεται μέσω της PCA;

3.4 Ποιος ταξινομητής έχει τη καλύτερη επίδοση και γιατί;

[Ερώτημα 2 - Κανονικοποιημένη μη-αρνητική παραγοντοποίηση πινάκων**]**

Έστω το παρακάτω πρόβλημα βελτιστοποίησης για την κανονικοποιημένη μη-αρνητική παραγοντοποίηση πινάκων (regularized non-negative matrix factorization -regNMF):

$$\min_{W,C} \|\mathbf{X} - \mathbf{WC}\|_F^2 + \lambda\|\mathbf{W}\|_F^2 + \lambda\|\mathbf{C}\|_F^2 \text{ s.t. } \mathbf{W} \geq \mathbf{0}, \mathbf{C} \geq 0$$

Το πρόβλημα προφανώς δεν έχει λύση σε κλειστή μορφή και συνεπώς πρέπει να λυθεί επαναληπτικά. Να υλοποιήσετε στη συνάρτηση RegNMF(X,k,lambda,epsilon) έναν επαναληπτικό αλγόριθμο για την επίλυση του παραπάνω προβλήματος βελτιστοποίησης. Η συνάρτηση παίρνει ως είσοδο έναν μη-αρνητικό πίνακα X διαστάσεων d x N, το πλήθος των συνιστωσών k, και τη τιμή της παραμέτρου κανονικοποίησης λ (lambda) και το κατώφλι τερματισμού ε (epsilon) και επιστέφει τους μη αρνητικούς πίνακες W διαστασης d x k και C διάστασης k x N.

Για να διαπιστώσουμε εάν συγκλίνει στη βέλτιστη λύση ένας επαναληπτικός αλγόριθμος συνήθως παρακολουθούμε το σφάλμα ανακατασκευής ||X – W[t]C[t]||_F^2/||X||_F^2 σε κάθε επανάληψη και εάν η μεταβολή του ανάμεσα σε δύο διαδοχικές επαναλήψεις είναι μικρότερη από ένα κατώφλι ε (||X – W[t]C[t]||_F^2 - ||X – W[t-1]C[t-1]||_F^2)/||X||_F^2 < ε με ε = 0.01 ή 0.001 ή 0.0001) τερματίζουμε τον αλγόριθμο. Το t στις παραπάνω σχέσεις συμβολίζει το δείκτη επανάληψης.

Καλείστε να μελετήσετε την σύγκλιση του αλγορίθμου χρησιμοποιείστε συνθετικά δεδομένα. Δηλαδή, να κατασκευάστε ένα τυχαίο πίνακα X διάστασης 500 x 1000 **με μη αρνητικές τιμές** και μελετήστε τη συμπεριφορά του αλγορίθμου regNMF ως προς το πλήθος των επαναλήψεων που απαιτούνται για να συγκλίνει εάν k= 1, 10, 100 και ε = 0.1, 0.01, 0.001. Ποια είναι τα συμπεράσματά σας ως προς τη συμπεριφορά του αλγορίθμου για διαφορετικές τιμές του k και epsilon (ε);

Για τη δημιουργία των συνθετικών δεδομένων μπορείτε να χρησιμοποιήσετε τη συνάρτηση rand της numpy. Για να βεβαιωθείτε ότι τα στοιχεία του πίνακα είναι μη αρνητικά μπορείτε να εφαρμόσετε ένα τελεστή απόλυτης τιμής σε κάθε στοιχείο του τυχαίου πίνακα που παρήγαγε η συνάρτηση rand.