ΕΠ08 Αναγνώριση Προτύπων – Μηχανική Μάθηση

3η Εργασία: Support Vector Machines

Τύπος εργασίας: ***Ατομική - Προαιρετική***

Ημερομηνία παράδοσης: ***Κυριακή 13/06/2021, 23:55 (Δεν θα δοθεί παράταση)***

Τρόπος παράδοσης: ***Αποκλειστικά μέσω του eclass***

Σύνολο βαθμών: **100** (Προσθετικά μέχρι 10% στον τελικό βαθμό του μαθήματος μόνο εφόσον ο βαθμός στη τελική εξέταση είναι προβιβάσιμος, δηλαδή μεγαλύτερος ή ίσος του 5).

Σε αυτή την εργασία καλείστε να διερευνήσετε την επίδοση των support vector machines (SVMs) στην ταξινόμηση εικόνων, και πιο συγκεκριμένα στην ταξινόμηση εικόνων οι οποίες απεικονίζουν χειρόγραφα ψηφία. Η εργασία είναι ατομική και έχει προαιρετικό χαρακτήρα, δίνοντας σας τη δυνατότητα να βελτιώσετε τον τελικό σας βαθμό μέχρι και 10%, εφόσον ο βαθμός στη τελική εξέταση είναι προβιβάσιμος. Καθώς η παρούσα εργασία είναι προαιρετική και μικρής δυσκολίας δεν θα υπάρχει υποστήριξη από πλευράς διδασκόντων στο χώρο συζητήσεων του eclass.

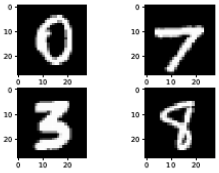
Συνιστάται ιδιαίτερα, να αφιερώσετε χρόνο ώστε να κατανοήσετε την θεμελιώδη λογική πίσω από τα ερωτήματα της εργασίας και να αποφύγετε την αναζήτηση έτοιμων λύσεων στο διαδίκτυο. Αν ωστόσο συμβουλευτείτε ή/και χρησιμοποιήσετε οποιοδήποτε υλικό ή/και κώδικα που είναι διαθέσιμος στο διαδίκτυο, πρέπει να αναφέρεται σωστά τη πηγή ή/και το σύνδεσμο στην ιστοσελίδα που αντλήσατε πληροφορίες. Σε κάθε περίπτωση, η αντιγραφή τμήματος ή του συνόλου της εργασίας δεν είναι αποδεκτή και στη περίπτωση που διαπιστωθεί αντιγραφή θα μηδενιστούν στο μάθημα όλα τα εμπλεκόμενα μέρη.

Θα πρέπει να υποβάλετε **ένα μόνο αρχείο Notebook IPython (Jupiter notebook) μέσω του εργαλείου εργασίες του eclass**, ακολουθώντας την εξής σύβαση ονομασίας για το αρχείο σας: *Επώνυμο_ΑριθμόςΜητρώου.ipynb*

Σημαντικό: Το IPython notebook που θα παραδώσετε θα πρέπει βεβαιωθείτε ότι ανοίγει και να εκτελείται στο google colab.

Ερώτημα

Καλείστε να διερευνήσετε την επίδοση των support vector machines στο πρόβλημα της αναγνώρισης χειρόγραφων ψηφίων. Για το σκοπό αυτό θα χρησιμοποιήσετε το σύνολο δεδομένων MNIST και υλοποιήσεις αλγορίθμων της βιβλιοθήκης scikit-learn. Το σύνολο δεδομένων MNIST αποτελείται από 70000 εικόνες χειρόγραφων ψηφίων και, τυπικά, χωρίζεται σε τρία υποσύνολα: training set (50000 εικόνες), validation set (10000 εικόνες), test set (10000 εικόνες). Κάθε εικόνα έχει διάσταση 28 x 28 pixels και απεικονίζει ένα χειρόγραφο ψηφίο. Παραδείγματα τέτοιων εικόνων απεικονίζονται στο παρακάτω σχήμα:



- Ζητείται να φορτώσετε τα δεδομένα του συνόλου MNIST και να μετατρέψετε κάθε εικόνα σε μορφή διανύσματος διάστασης 28 x 28 = 784. Στη συνέχεια κανονικοποιήστε (normalize) τα δεδομένα στο διάστημα [0,1].
- Στα SVMs υπάρχουν διάφορες επιλογές που μπορεί να επηρεάσουν την απόδοση τους σε προβλήματα ταξινόμησης. Παραδείγματα τέτοιων επιλογών αποτελούν ο τύπος του πυρήνα (kernel) και οι τιμές των διάφορων παραμέτρων. Ζητείται να εξετάσετε την επίδοση των SVMs για διαφορετικές τιμές παραμέτρων/πυρήνων ώστε να καθορίσετε το συνδυασμό παραμέτρων/πυρήνων που οδηγούν στη μεγαλύτερη ακρίβεια ταξινόμησης. Για αυτό το πείραμα να χρησιμοποιήσετε 60000 εικόνες για εκπαίδευση (training) και 10000 παραδείγματα για δοκιμές (test). Να αναφέρετε τις τιμές των παραμέτρων, δηλαδή τύπο πυρήνα, τιμές των C και gamma που οδηγούν στις καλύτερες επιδόσεις τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής (test set).
- Στη συνέχεια, να εφαρμόσετε PCA στα δεδομένα επιλέγοντας 3 διαφορετικές τιμές για τη διατηρούμενη διακύμανση και για κάθε τιμή διακύμανσης εκτελέστε ξανά τη μέθοδο SVM χρησιμοποιώντας τις παραμέτρους που οδήγησαν στην καλύτερη επίδοση στο παραπάνω ερώτημα. Για κάθε εκτέλεση, αναφέρετε τον αριθμό των συνιστωσών (components) που διατηρούνται καθώς και την ακρίβειας ταξινόμησης. Επίσης, καταγράψτε τους χρόνους εκτέλεσης κάθε πειράματος και εξαγάγετε συμπεράσματα σχετικά με μια πιθανή αντιστάθμιση (trade-off) μεταξύ ακρίβειας ταξινόμησης, μείωσης διαστάσεων και χρόνου εκτέλεσης του αλγορίθμου.