1η & 2η ΥΠΟΧΡΕΩΤΙΚΗ ΕΡΓΑΣΙΑ ΣΤΟ ΜΑΘΗΜΑ «ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ – ΣΤΑΤΙΣΤΙΚΗ ΜΑΘΗΣΗ»

Α. Να γραφεί πρόγραμμα σε οποιαδήποτε γλώσσα προγραμματισμού το οποίο να υλοποιεί/χρησιμοποιεί ένα **Support Vector Machine** για πρόβλημα διαχωρισμού κλάσεων ή για πρόβλεψη τιμών συνάρτησης (regression).

Β. Να γραφεί πρόγραμμα σε οποιαδήποτε γλώσσα προγραμματισμού το οποίο να υλοποιεί την μέθοδο **Kernel Principal Component Analysis plus Linear Discriminant Analysis (KPCA+LDA)** και στην συνέχεια κατηγοριοποίηση για πρόβλημα διαχωρισμού πολλών κλάσεων.

Οι ταξινομητές αυτοί θα εκπαιδευτούν, θα δοκιμαστούν και θα συγκριθούν σε **δυο διαφορετικά προβλήματα (βάσεις δεδομένων)** επιλογής σας από τα παρακάτω:

Βάσεις Δεδομένων

Α. η βάση δεδομένων Cifar-10 ή CIFAR100 ή SVHN που υπάρχουν στις παρακάτω διευθύνσεις:

https://www.cs.toronto.edu/~kriz/cifar.html
http://ufldl.stanford.edu/housenumbers/

Β. Οποιαδήποτε από τις βάσεις δεδομένων που υπάρχουν στις ιστοσελίδες:
http://www.cs.toronto.edu/~roweis/data.html
http://www.cs.cmu.edu/~cil/v-images.html
https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
https://www.kaggle.com/datasets

και αφορούν προβλήματα κατηγοριοποίησης πολλών κλάσεων. Όπου δεν υπάρχει σύνολο ελέγχου χωρίζεται η βάση τυχαία σε σύνολο εκπαίδευσης (60%) και ελέγχου (40%) ή ακολουθείται τεχνική cross-validation.

Εξαγωγή Χαρακτηριστικών

Για το διαχωρισμό των δειγμάτων μπορεί να μειώνεται πρώτα η διάσταση των δεδομένων χρησιμοποιώντας PCA ώστε να κρατήσετε περισσότερο από 90% της πληροφορίας.

Έκθεση αποτελεσμάτων

Θα πρέπει να γραφεί έκθεση στην οποία να περιγράφονται: ο αλγόριθμος, να δίνονται χαρακτηριστικά παραδείγματα ορθής και εσφαλμένης κατηγοριοποίησης καθώς και ποσοστά επιτυχίας στα στάδια της εκπαίδευσης (training) και του ελέγχου (testing), χρόνος εκπαίδευσης και ποσοστά επιτυχίας για διαφορετικούς πυρήνες, γραμμικό και μη γραμμικούς καθώς και διαφορετικές τιμές των παραμέτρων εκπαίδευσης. Να συγκριθεί η απόδοση του SVM και της KPCA+LDA σε σχέση με την κατηγοριοποίηση πλησιέστερου γείτονα (Nearest Neighbor) και πλησιέστερου κέντρου κλάσης (Nearest Class Centroid). Να σχολιασθούν τα αποτελέσματα και ο κώδικας. Η δεύτερη έκθεση θα είναι επέκταση της πρώτης με αποτελέσματα και για KPCA+LDA καθώς και σύγκριση με τις επιδόσεις των SVMs.

ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΔΟΣΗΣ

10 Δεκεμβρίου 2024 (SVMs) - 30 Δεκεμβρίου 2024 (KPCA+LDA)

Για κάθε ημέρα αργοπορημένης υποβολής της εργασίας και για πέντε ημέρες μειώνεται η βαθμολογία κατά 10%. Μετά από την παράδοση της εργασίας θα ακολουθήσει προφορική εξέταση πάνω στην εργασία, στην οποία θα περιλαμβάνεται **και προφορική εξέταση του κώδικα.**

Βοηθητικές σελίδες

http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (LibSVM)
http://en.wikipedia.org/wiki/Support_vector_machine (Σύνδεσμοι για λογισμικό)

e-mail: tefas@csd.auth.gr, Τηλ. 2310-991932