Εργασία - Αναγνώριση Προτύπων & Μηχανική Μάθηση

Διδάσκων: Επικ. Καθ. Παναγιώτης Πετραντωνάκης (ppetrant@ece.auth.gr)
Βοηθός διδασκαλίας: Υπ. Διδ. Στέφανος Παπαδόπουλος (stefpapad@iti.gr)

2023

Για τα ερωτήματα στα μέρη Α-Γ χρησιμοποιήστε το dataset.csv. Για το μέρος Δ χρησιμοποιήστε το datasetC.csv και το datasetCTest.csv.

Μέρος Α (2 Μονάδες)

Φορτώστε τα δεδομένα σας[1] και χωρίστε τα σε training και test sets με αναλογία 50%-50%. Κάθε ένα από τα δείγματα των δεδομένων σας είναι ένα $2-D$ διάνυσμα μαζί με μια ετικέτα (label) (1, 2, ή 3) για κάθε ένα από αυτά στην τελευταία στήλη. Εκπαιδεύστε με τη Maximum Likelihood τεχνική ένα Bayes ταξινομητή με τα δεδομένα σας και εξετάστε δύο διαφορετικές περιπτώσεις (χρησιμοποιώντας κανονικές κατανομές):

1. Ίδιο πίνακα συνδιασποράς για όλες τις κλάσεις.
2. Διαφορετικό πίνακα συνδιασποράς για κάθε κλάση.
3. Απεικονίστε τα test δεδομένα σας και μαρκάρετε με ειδικό χρώμα (ή σύμβολο) αυτά που ταξινομήθηκαν λανθασμένα και στις δυο περιπτώσεις. Επίσης απεικονίστε τις περιοχές (regions) που αντιστοιχούν σε κάθε κλάση.

Υπολογίστε το μέσο σφάλμα ταξινόμησης στο test set. Ποια από τις δύο παραπάνω προσεγγίσεις δίνει το καλύτερο αποτέλεσμα;

Μέρος Β (2 Μονάδες)

Σε αυτό το μέρος 'εκπαιδεύστε' έναν ταξινομητή $k-NN$ χρησιμοποιώντας το ίδιο training set με το προηγούμενο μέρος.

1. Υπολογίστε το μέσο σφάλμα ταξινόμησης στο ίδιο test set με το προηγούμενο μέρος.
2. Απεικονίστε τα δεδομένα σας καθώς και τις περιοχές (regions) που αντιστοιχούν σε κάθε κλάση.
3. Επαναλάβετε τα δυο προηγούμενα ερωτήματα για $k = 1, ..., 10$.

Πώς συγκρίνονται τα αποτελέσματα αυτού του μέρους με τα αντίστοιχα του προηγούμενου;

Μέρος Γ (2 Μονάδες)

Σε αυτό το μέρος χρησιμοποιήστε έναν SVM ταξινομητή για να ταξινομήσετε τα ίδια δεδομένα διαχωρίζοντας σε train και test sets με τον ίδιο τρόπο όπως και στα προηγούμενα μέρη.

1. Χρησιμοποιήστε γραμμικό SVM.
2. Χρησιμοποιήστε RBF kernel SVM και πειραματιστείτε με τις υπερπαραμέτρους του. Σχολιάστε τα αποτελέσματα σας (μέσο σφάλμα στο test set) σε κάθε περίπτωση.
3. Απεικονίστε τα δεδομένα σας χρησιμοποιώντας διακριτό συμβολισμό για τα train set, test set, support vectors και δείγματα με λάθος ταξινόμηση. Απεικονίσετε επίσης και τις περιοχές για κάθε κλάση.

Πώς συγκρίνονται τα αποτελέσματα αυτού του μέρους με τα αντίστοιχα των προηγούμενων;

[1] Αν θέλετε χρησιμοποιήστε: data = np.loadtxt("dataset.csv", delimiter=",", dtype=np.float64)

Μέρος Δ (4 Μονάδες)

Σε αυτό το μέρος θα εργαστείτε με το datasetC.csv το οποίο θα χρησιμοποιήσετε ως training set. Τα training δεδομένα σας έχουν 5000 δείγματα και 400 χαρακτηριστικά (features) ανα δείγμα (sample) που συνοδεύονται από μια ετικέτα (label), $1, ..., 5$ στην τελευταία στήλη. Με αυτά τα δεδομένα αναπτύξτε ένα αλγόριθμό ταξινόμησης με όποια μέθοδο εσείς επιθυμείτε. Μπορείτε επίσης να διαχειριστείτε τις τιμές των χαρακτηριστικών σας όπως νομίζετε.

Ακολούθως θα χρησιμοποιήσετε τα δεδομένα του αρχείου datasetCTest.csv σαν test set (σε αυτό **δεν** δίνονται οι ετικέτες). Σε αυτά τα δεδομένα θα εφαρμόσετε το **τελικό, εκπαιδευμένο** μοντέλο σας και θα εξάγετε ένα διάνυσμα με το όνομα labelsX (δείτε στις οδηγίες παρακάτω την επεξήγηση για το X) το οποίο και θα υποβάλετε σε numpy μορφή.

Στις ομάδες με τα καλύτερα αποτελέσματα (ελάχιστο σφάλμα ταξινόμησης) από αυτό το μέρος θα δοθεί προσθετική bonus βαθμολόγηση.

Οδηγίες

- Η Υλοποίηση της εργασίας θα γίνει σε Python. Επιλέξτε ένα notebook (π.χ., Jupyter, Collab) και γράψτε τον κώδικα όσο και τα σχόλιά σας.
- Για την παράδοση θα ανεβάσετε ΕΝΑ αρχείο με όνομα: TeamX.zip με όλα τα απαραίτητα αρχεία (αν είστε ομάδα δύο ατόμων, ΜΟΝΟ ένας κατεθέτει την εργασία). Πρέπει μέσα στο αρχείο .zip να περιέχονται:
  1. το αρχείο TeamX-AC.ipynb με τον κώδικα για τα μέρη Α-Γ.
  2. το αρχείο TeamX-D.ipynb με τον κώδικα για το μέρος Δ.
  3. το αρχείο labelsX.npy το οποίο θα αφορά το διάνυσμα των ετικετών που έχετε εξάγει από το μέρος Δ. (**πολύ σημαντικό:** βεβαιωθείτε ότι το αποθηκευμένο labelsX.npy μπορεί να διαβαστεί με την numpy.load() και ότι έχει διάσταση N (N ο αριθμός των samples στο test set))
  4. ένα αρχείο TeamX.pdf σε μορφή διαφανειών όπου θα περιγράφονται (σε μορφή παρουσίασης) όλα τα μέρη της εργασίας (μέρος Α έως Δ).

  Σε όλα τα παραπάνω αρχεία, όπου X βάλτε τον αύξοντα αριθμό της ομάδας σας (1, 2, 3 κτλ., **ΟΧΙ** 01, 02, 03, κτλ). Το αρχείο της παρουσίασης πρέπει να ειναι (αυστηρά!) μέχρι 50 διαφάνειες (10 για καθένα από τα μέρη Α-Γ και 20 για το τελευταίο). Σε κάθε αρχείο .ipynb, .pdf θα αναγράφονται (**σημαντικό!**) μέσα τα στοιχεία σας (ονοματεπώνυμο, ΑΕΜ).
- Κάθε ένα από τα ερωτήματα των μερών Α-Γ θα απαντηθεί (κώδικας) σε ξεχωριστό κελί. Και ο κώδικας σε κάθε κελί θα συνοδεύεται από σύντομα σχόλια (σημαντικό!). Τον κώδικα για το μέρος Δ μπορείτε να τον δομήσετε όπως θέλετε αλλά τα σχετικά σχόλια είναι κι εδώ απαραίτητα.
- Η βαθμολογία σας θα προκύψει από την ποιότητα του κώδικα και των σχετικών σχολίων, από την ποιότητα της αντίστοιχης παρουσίασης του κάθε μέρους και από την ορθότητα των προσεγγίσεων και των αποτελεσμάτων. Οι καλύτερες εργασίες που θα προκύψουν από το μέρος Δ θα παρουσιάσουν τον ταξινομητή τους δια ζώσης. (η δια ζώσης παρουσίαση είναι υποχρεωτική για την bonus βαθμολόγηση).
- Τελική ημερομηνία υποβολής: Παρασκευή 5 Ιανουαρίου, 2024, 23:59.

ΚΑΛΗ ΕΠΙΤΥΧΙΑ!