Εργασία - Αναγνώριση Προτύπων & Μηχανική Μάθηση

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

2025-2026

Μέρος Α (2 Μονάδες)

Σε αυτή την άσκηση θα χρησιμοποιήσετε το σύνολο δεδομένων στο αρχείο dataset1.csv. Το σύνολο αυτό αποτελείται από 300 γραμμές δειγμάτων δύο διαστάσεων (οι δύο πρώτες στήλες αντιστοιχούν στα δεδομένα και η τελευταία στήλη στην ετικέτα), τα οποία ανήκουν σε 3 κλάσεις. Οι πρώτες 100 γραμμές ανήκουν στην κλάση 0, οι επόμενες 100 στην κλάση 1 κ.ο.κ. Θεωρήστε ότι τα δείγματα κάθε κλάσης προέρχονται από διαφορετική κανονική κατανομή. Σας ζητείται να χρησιμοποιήσετε την τεχνική της Μέγιστης Πιθανοφάνειας για να βρείτε τις παραμέτρους αυτών των τριών διαφορετικών κανονικών κατανομών. Δεν επιτρέπεται να χρησιμοποιήσετε συναρτήσεις βιβλιοθηκών. Επιτρέπεται να χρησιμοποιήσετε βασικές συναρτήσεις για πράξεις με πίνακες, όπως np.sum κ.λ.π. Σχεδιάστε και τις 3 κατανομές σε ένα ενιαίο 3D plot.

Μέρος Β (2 Μονάδες)

Σε αυτή την άσκηση θα εργαστείτε με το σύνολο δεδομένων στο αρχείο dataset2.csv. Περιλαμβάνει 200 μονοδιάστατα δείγματα (200 γραμμές, 1 στήλη). Το έργο σας είναι να υλοποιήσετε τη μέθοδο παραθύρων $Parzen$ για να εκτιμήσετε τη συνάρτηση πυκνότητας πιθανότητας της κατανομής των δεδομένων.

Αφού υλοποιήσετε τον κώδικα της παραπάνω μεθόδου, υποθέστε ότι το σύνολο δεδομένων που έχετε προέρχεται από την μονοδιάστατη κανονική κατανομή $N(1, 4)$. Βρείτε την καταλληλότερη τιμή για το h βασιζόμενοι σε αυτή τη γνώση. Δημιουργήστε ένα ιστόγραμμα των δεδομένων για να επιβεβαιώσετε ότι πράγματι προέρχονται από την παραπάνω κατανομή. Για κάθε h στο εύρος $[0.1, 10]$ με βήμα $= 0.1$, υπολογίστε: την προβλεπόμενη πιθανοφάνεια για κάθε σημείο στα δεδομένα, την πραγματική πιθανοφάνεια (μπορείτε να χρησιμοποιήσετε την συνάρτηση της κανονικής κατανομής) και το τετραγωνικό σφάλμα μεταξύ των δύο. Επαναλάβετε αυτή τη διαδικασία και για τα δύο $kernels$ (υπερκύβος και γκαουσιανή). Τέλος, εκτυπώστε την καταλληλότερη τιμή του h για κάθε $kernel$ και δημιουργήστε ένα $plot$ (ένα για κάθε kernel) το οποίο να δείχνει τις τιμές του h στον άξονα x και το σφάλμα στον άξονα y.

Μέρος Γ (2 Μονάδες)

Σε αυτή την άσκηση θα αναπτύξετε έναν ταξινομητή k Nearest Neighbors (KNN). Θα χρησιμοποιήσετε το σύνολο δεδομένων dataset3.csv για εκπαίδευση και το testset.csv για δοκιμή. Τα αρχεία αποτελούνται από 50 δείγματα δύο διαστάσεων το καθένα, μαζί με τις ετικέτες τους στην τελευταία στήλη (50 γραμμές και 3 στήλες).

Αρχικά, υλοποιήστε μια συνάρτηση eucl(x, trainData) που να επιστρέφει την ευκλείδεια απόσταση του x από όλα τα σημεία στο $trainData$. Μην χρησιμοποιείτε συναρτήσεις βιβλιοθηκών, εκτός από βασικές πράξεις με πίνακες.

Κατόπιν, υλοποιήστε μια συνάρτηση neighbors(x, trainData, k), όπου k είναι ο αριθμός των γειτόνων. Η συνάρτηση πρέπει να υπολογίζει την απόσταση του x από όλα τα σημεία στο $trainData$ να ταξινομεί τις αποστάσεις σε φθίνουσα σειρά, και να επιστρέφει τα k κορυφαία σημεία από το $trainData$.

Τέλος, υλοποιήστε μια συνάρτηση predict(testData, trainData, k), όπου θα καλεί τη συνάρτηση neighbors για κάθε σημείο στο testData, στη συνέχεια θα υπολογίζει την πιθανότητα κάθε σημείου να ανήκει στην κλάση 0 ή στην κλάση 1 (οι πιθανότητες αυτές πρέπει να αθροίζουν στο 1). Τέλος, πρέπει να επιστρέφει δύο πιθανότητες για κάθε σημείο στο testData.

Τώρα θα χρησιμοποιήσετε το test set για να βρείτε την καλύτερη τιμή του k και την ακρίβειά του. Για k στο εύρος $[1, 30]$, υπολογίστε την ακρίβεια κάθε ταξινομητή χρησιμοποιώντας τα δεδομένα από το αρχείο testset.csv. Εκτυπώστε την καλύτερη τιμή του k μαζί με την ακρίβειά της και δημιουργήστε ένα plot με τις τιμές του k στον άξονα x και την ακρίβεια στον άξονα y.

Τέλος, θα σχεδιάσετε τα όρια απόφασης του καλύτερου ταξινομητή (χρησιμοποιώντας την τιμή του k που βρήκατε προηγουμένως). Μπορείτε να χρησιμοποιήσετε τη συνάρτηση contourf για να σχεδιάσετε τις περιοχές απόφασης.

Μέρος Δ (4 Μονάδες)

Σε αυτό το μέρος θα εργαστείτε με το datasetTV.csv το οποίο θα χρησιμοποιήσετε ως training set. Τα training δεδομένα σας έχουν 8743 δείγματα και 224 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 datasetTest.csv (6955 δείγματα) 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: Τετάρτη 14 Ιανουαρίου, 2026, 23:59.

ΚΑΛΗ ΕΠΙΤΥΧΙΑ!