Εργασία - Αναγνώριση Προτύπων & Μηχανική Μάθηση

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

2024

Μέρος Α (2 Μονάδες)

Εργάζεστε σε μια εταίρεία που παράγει βιντεοπαιχίδια και συγκεκριμένα σε ένα τμήμα που πρόσφατα δημιουργήθηκε στην εταιρεία και ασχολείται με την αναγνώριση του επιπέδου του στρές στους χρήστες με βάση τα μοτίβα συχνότητας και δύναμης πίεσης των πλήκτρων της κονσόλας. Ένας συνάδελφός σας από το ίδιο τμήμα αναλύοντας αυτά τα μοτίβα εξήγαγε έναν δείκτη-αριθμό x και ισχυρίζεται ότι αυτός ο δείκτης μπορεί να χρησιμοποιηθεί σε ένα σύστημα ταξινόμησης για να διαπιστωθεί κάθε φορά αν ο χρήστης αισθάνεται στρες ή όχι. Επίσης από μελέτες που διεξήγαγε ο συνάδελφός σας, παρατήρησε ότι η κατανομή πυκνότητας πιθανότητας που ακολουθεί αυτός ο δείκτης και για τις δυο κλάσεις (χωρίς στρες = κλάση ω_1, με στρες = κλάση ω_2) είναι:

$$p(x|\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$$

με τη παράμετρο θ να είναι άγνωστη. Για να διαπιστώσετε αν όντως ο δείκτης που εξήγαγε ο συνάδελφός σας είναι αξιόπιστος δείκτης για το επίπεδο του στρες των χρηστών, ζητήσατε από 12 άλλους συναδέλφους σας στο ίδιο τμήμα να παίξουν ένα συγκεκριμένο παιχνίδι στην κονσόλα που παράγει η εταιρεία και υπολογίσατε τον εν λόγω δείκτη για κάθε χρήστη. Στη συνέχεια ρωτήσατε τους συναδέλφους σας να σας πουν αν ένιωσαν στρες κατά τη διάρκεια του παιχνιδιού ή όχι. Από τους 12 οι 7 δήλωσαν ότι δεν ένιωσαν στρες ενώ οι 5 δήλωσαν ότι το παιχνίδι τους δημιούργησε έντονο στρες. Καλήστε λοιπόν εσείς να υλοποιήσετε ένα ταξινομητή μέγιστης πιθανοφάνειας. Συγκεκριμένα:

1. Εκτιμήστε τις παραμέτρους $\hat{\theta}_1$ και $\hat{\theta}_2$ με την μέθοδο της μέγιστης πιθανοφάνειας και για τις δύο κλάσεις αν για την κλάση ω_1 οι δείκτες είναι $D_1 = [2.8, -0.4, -0.8, 2.3, -0.3, 3.6, 4.1])$ ενώ για την κλάση ω_2 είναι $D_2 = [-4.5, -3.4, -3.1, -3.0, -2.3]$. Απεικονίστε τις $\log p(\mathcal{D}_1|\theta)$ και $\log p(\mathcal{D}_2|\theta)$ σε συνάρτηση με τη θ.

2. Χρησιμοποιήστε τη συνάρτηση διάκρισης

$$g(x) = \log P(x|\hat{\theta}_1) - \log P(x|\hat{\theta}_2) + \log P(\omega_1) - \log P(\omega_2)$$

και ταξινομήστε τα δύο σύνολα τιμών. Τι παρατηρείτε για το πρόσιμο της $g(x)$ σε σχέση με τα δεδομένα σας (απεικονίστε την); Περιγράψτε τον κανόνα απόφασης. Τι παρατηρείτε σε σχέση με την ταξινόμηση των δεδομένων σας με βάση αυτόν τον κανόνα; *Βοήθεια: Μπορείτε να δημιουργήσετε μια κλαση Classifier με α) μια συνάρτηση fit που θα λαμβάνει ως όρισμα ένα σύνολο D και ένα διάνυσμα από υποψήφιες τιμές θ και θα υπολογίζει τις τιμές μέγιστης πιθανοφάνειας για το θ, β) μια συνάρτηση predict που θα λαμβάνει ως όρισμα ένα σύνολο D και τις a priori πιθανότητες των κλάσεων και θα επιστρέφει τις τιμές της συνάρτησης g*

Μέρος Β (2 Μονάδες)

Σε αυτό το μέρος καλείστε να υλοποιήσετε ένα νέο ταξινομητή εκτιμώνατας την άγνωστη παράμετρο θ με την μέθοδο εκτίμησης κατά Bayes.

Μετά από επίπονο πειραματισμό, διαπιστώσατε ότι οι τιμές τις παραμέτρου θ μπορούν να μοντελοποιηθούν με την συνάρτηση πυκνότητας πιθανότητας (prior)

$$p(\theta) = \frac{1}{10\pi}\frac{1}{1+(\theta/10)^2}.$$

Έχοντας αυτό το μοντέλο και με βάση τη θεωρία είσαστε τώρα σε θέση να υπολογίσετε την a posteriori πιθανότητα $p(\theta|\mathcal{D})$ και την πυκνότητα πιθανότητας $p(x|\mathcal{D}_j), j = 1, 2$.

1. Απεικονίστε τις εκ των υστέρων πυκνότητες πιθανότητας $p(\theta|\mathcal{D}_1)$ και $p(\theta|\mathcal{D}_2)$. Τι παρατηρείτε σε σχέση με την (prior) $p(\theta)$. *Βοήθεια: Για τον υπολογισμό των ολοκληρωμάτων μπορείτε να χρησιμοποιήσετε τον κανόνα του τραπεζίου*

2. Υλοποιήστε μια συνάρτηση predict που θα υπολογίζει τις τιμές μιας συνάρτησης διάκρισης

$$h(x) = \log P(x|\mathcal{D}_1) - \log P(x|\mathcal{D}_2) + \log P(\omega_1) - \log P(\omega_2).$$

   Τι παρατηρείτε τώρα για τις τιμές τις h σε σχέση με τα σύνολά δεδομένων σας (απεικονίστε την); Πως αξιολογείτε την μέθοδο εκτίμησης παραμέτρων κατα Bayes σε σχέση με τη μέθοδο της μέγιστης πιθανοφάνειας για το συγκεκριμένο παράδειγμα; Που πιστεύετε ότι οφείλεται η διαφορά των δύο προσεγγίσεων για το συγκεκριμένο παράδειγμα; *Βοήθεια:Μπορείτε να υιοθετήσετε παρόμοια υλοποίηση με το μέρος Α*

Μέρος Γ (2 Μονάδες)

1η ενότητα Εργάζεστε ως βοηθός έρευνας στο Εργαστήριο Ανθοκομίας του Τμήματος Γεωπονίας, ΑΠΘ, με ειδίκευση στην ανάλυση δεδομένων. Ένα τμήμα έρευνας στο εργαστήριο αφορά την αυτοματοποιημένη αναγνώριση διαφορετικών ειδών από ένα συγκεκριμένο φυτό, της Ίριδας. Τρία συγκεκριμένα είδη η Iris setosa, η Iris versicolor, και η Iris virginica παρουσιάζον διαφορές στο μήκος και πλάτος των σεπάλων και των πετάλων του ανθους τους. Από τη βιβλιοθήκη sklearn μπορείτε να κατεβάσετε μια βάση από 150 (50 για κάθε είδος) μετρήσεις τους μήκους και του πλάτους των σεπάλων και των πετάλων του άνθους κάθε είδους. Απομονώνοντας μόνο τα δύο πρώτα χαρακτηριστικά της βάσης, χρησιμοποιήστε τον έτοιμο αλγόριθμο DecisionTreeClassifier από τη βιβλιοθήκη sklearn και ταξινομήστε το 50% τυχαίων δειγμάτων του συνόλου αφού πρώτα έχετε εκπαιδεύσει τον αλγόριθμο με το υπόλοιπο 50%.

1. Τι ποσοστό σωστής ταξινόμησης λαμβάνετε; Ποιο βάθος δέντρου σας δίνει το καλύτερο ποσοστό;

2. Απεικονήστε τα όρια απόφασης του ταξινομητή για το καλύτερο αποτέλεσμα (*Βοήθεια: χρησιμοποιήστε τη συνάρτηση contourf*)

2η ενότητα Τώρα θα δημιουργήστε ένα Random Forest ταξινομητή 100 δέντρων με την τεχνική Bootstrap. Πιο συγκεκριμένα, το 50% των δειγμάτων που χρησιμοποιήσατε για εκπαίδευση στην προηγούμενη ενότητα (έστω ότι το ονομάζετε σύνολο Α) χρησιμοποιήστε το τώρα για την δημιουργία 100 νέων συνόλων εκπαίδευσης ένα για κάθε δέντρο όπου κάθε φορά θα χρησιμοποιήται το $\gamma = 50\%$ του συνόλου Α. Το σύνολο που ταξινομήσατε στο προηγούμενο μέρος χρησιμοποιήστε το και εδώ για αξιολόγηση του αλγορίθμου. Όλα τα δέντρα να έχουν το ίδιο μέγιστο βάθος.

1. Τι ποσοστό σωστής ταξινόμησης λαμβάνετε; Ποιο βάθος δέντρου σας δίνει το καλύτερο ποσοστό;

2. Απεικονήστε τα όρια απόφασης του ταξινομητή για το καλύτερο αποτέλεσμα. Τι παρατηρείτε σε σχέση με τον απλό ταξινομητή της προηγούμενης ενότητας;

3. Πώς πιστεύετε ότι επηρεάζει το ποσοστό γ την απόδοση του αλγορίθμου; Δώστε παραδείγματα.

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 Χ 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 8 Ιανουαρίου, 2025, 23:59.

ΚΑΛΗ ΕΠΙΤΥΧΙΑ!