ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ
ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΑΝΑΛΥΣΗΣ ΗΛΕΚΤΡΙΚΩΝ ΚΥΚΛΩΜΑΤΩΝ
ΜΑΘΗΜΑ: ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ – 9ο ΕΞΑΜΗΝΟ

Εργασία 2

(Προθεσμία: Κυριακή 19 Νοεμβρίου 2023, 23:59)

1. Κατεβάστε από εδώ το αρχείο δεδομένων Data_ex1.txt που περιέχει δεδομένα δύο χαρακτηριστικών $\mathbf{x}=\left[x_1\ x_2\right]$ από 3 κλάσεις (ω1, ω2, ω3). Κάθε γραμμή του αρχείου περιέχει δεδομένα στη μορφή: x1,x2,class_label .

Α. Να γράψετε κατάλληλο κώδικα ώστε να εκτιμήσετε τις πυκνότητες πιθανότητας $p(\mathbf{x}|\omega_1)$, $p(\mathbf{x}|\omega_2)$ και $p(\mathbf{x}|\omega_3)$ με τη μέθοδο παραθύρων Parzen, για συνάρτηση παραθύρου:

$$\varphi(\mathbf{x}-\mathbf{x}_i)=\frac{1}{h_N\sqrt{2\pi}}e^{-\frac{\|\mathbf{x}-\mathbf{x}_i\|_2^2}{2h_N{}^2}}$$

και για h_N=0.3 να απεικονίσετε κατάλληλα τις κατανομές σε κοινό γράφημα. Δοκιμάστε να κάνετε εκτίμηση για h_N=0.7 και h_N=0.1. Τι παρατηρείτε? Εάν είχατε στη διάθεσή σας μόνο το 25% των δεδομένων, τι τιμή θα έπρεπε να έχει το h_N για να κάνετε εκτίμηση με παρόμοια λεπτομέρεια με την αρχική? Επιβεβαιώστε την απάντησή σας πειραματικά.

Β. Να γράψετε κατάλληλο κώδικα ώστε να εκτιμήσετε τις πυκνότητες πιθανότητας $p(\mathbf{x}|\omega_1)$, $p(\mathbf{x}|\omega_2)$ και $p(\mathbf{x}|\omega_3)$ με τη μέθοδο k-NN, και για k=10 να απεικονίσετε κατάλληλα τις κατανομές σε κοινό γράφημα. Δοκιμάστε να κάνετε εκτίμηση για k=3 και k=30. Τι παρατηρείτε;

Γ. Θεωρώντας ότι όλες οι κλάσεις έχουν ίδιες apriori πιθανότητες, να απεικονίσετε τα δεδομένα και τις περιοχές απόφασης για ταξινόμηση σύμφωνα με τον κανόνα του Bayes και κατανομές που εκτιμώνται με τη μέθοδο των παραθύρων parzen για h_N=0.1 , h_N=0.3 , h_N=1.5. Πως επηρεάζονται οι περιοχές απόφασης από την παράμετρο παραθύρου; Σχολιάστε.

Δ. Να απεικονίστε τα δεδομένα και τις περιοχές απόφασης για ταξινόμηση σύμφωνα με τον κανόνα k-NN για k=3, k=8 και k=30. Πως επηρεάζονται οι περιοχές απόφασης από την παράμετρο k; Σχολιάστε.

Ε. Ποια τα κυριότερα πλεονεκτήματα και μειονεκτήματα των δύο τεχνικών τόσο στη μεταξύ τους σύγκριση, όσο και σε σύγκριση με γεωμετρικές τεχνικές ταξινόμησης (π.χ. linear discriminants, SVMs κλπ.);

2. Έστω δυο κατηγορίες ω1 και ω2 , των οποίων τα δείγματα σε 2-διαστάσεις ακολουθούν κατανομές που περιγράφονται από τις ακόλουθες πυκνότητες πιθανότητας:

$$p(\mathbf{x}\,|\,\omega_1) = N(\mu_1, \Sigma_1) \quad \text{και} \quad p(\mathbf{x}\,|\,\omega_2) = N(\mu_2, \Sigma_2),$$

με $\mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, $\Sigma_1 = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 1 \end{pmatrix}$ και $\mu_2 = \begin{pmatrix} -8 \\ 2 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$

Γράψτε κώδικα ώστε να παράξετε 150 τυχαία δείγματα από καθεμία από αυτές τις κατηγορίες και απεικονίστε τα δείγματα κατάλληλα ώστε να διακρίνονται οι κλάσεις.

A. Υλοποιείστε τον αλγόριθμό του Batch Perceptron και με αυτόν υπολογίστε έναν γραμμικό ταξινομητή για τις κλάσεις αυτές. Απεικονίστε την επιφάνεια απόφασης που προέκυψε επάνω στο προηγούμενο γράφημα.

B. Χρησιμοποιείστε γραμμικό SVM (από κατάλληλη βιβλιοθήκη της επιλογής σας) για να υπολογίσετε έναν νέο γραμμικό ταξινομητή για τα ίδια δεδομένα. Απεικονίστε τα δεδομένα, τα support vectors και το επίπεδο απόφασης σε νέο διάγραμμα

Γ. Σχολιάστε το αποτέλεσμα των δύο τεχνικών. Ποιες οι διαφορές και που οφείλονται?

3. Κατεβάστε το Wine Dataset από εδώ. Τα δεδομένα αυτά αποτελούν τα αποτελέσματα μιας χημικής ανάλυσης κρασιών από τρεις διαφορετικές καλλιεργητικές ποικιλίες {c1,c2,c3}, και περιλαμβάνουν τιμές για 13 χημικά συστατικά που μετρήθηκαν σε κάθε κρασί. Η πρώτη στήλη του αρχείου των δεδομένων περιλαμβάνει την ετικέτα της ποικιλίας του κάθε κρασιού, και οι επόμενες στήλες τις τιμές των συστατικών που μετρήθηκαν. Στόχος μας είναι να διερευνήσουμε το πρόβλημα της πρόβλεψης της ποικιλίας από τα αποτελέσματα της χημική ανάλυσης.

Βήματα:

A. Θεωρείστε το υποσύνολο των δεδομένων που περιέχει τιμές μόνο για τα 5 πρώτα συστατικά και για τα κρασιά από τις ποικιλίες c2 και c3. Να χωρίσετε το παραπάνω σε σύνολο εκπαίδευσης, επικύρωσης και δοκιμής (training, validation & test sets) με αναλογία 50%, 25% και 25% αντίστοιχα, με τυχαία επιλογή δεδομένων και ίδια αναλογία μεταξύ των κλάσεων σε κάθε σύνολο.

B. Χρησιμοποιήστε γραμμικό SVM για να εκπαιδεύσετε ταξινομητή που να διαχωρίζει την κλάση c2 από τη c3. Χρησιμοποιήστε το validation set για να ρυθμίσετε την παράμετρο C (box constraint) κατάλληλα, και για την καλύτερη τιμή εφαρμόστε τον ταξινομητή που εκπαιδεύσατε στο test set. Τι σφάλμα ταξινόμησης πετύχατε?

Γ. Επαναλάβετε το προηγούμενο για 5 νέους τυχαιοποιημένους διαμερισμούς των δεδομένων, και υπολογίστε τη μέση τιμή και την τυπική απόκλιση του σφάλματος ταξινόμησης στο test set.

Δ. Επαναλάβετε το Γ για μη-γραμμικό SVM δοκιμάζοντας διάφορες συναρτήσεις πυρήνα (RBF, polynomial κλπ). Τι σφάλμα πετύχατε? Ποιος είναι ο καλύτερος ταξινομητής για το πρόβλημα? Σχολιάστε.

Ε. Χρησιμοποιήστε γραμμικό SVM για να εκπαιδεύσετε ταξινομητές για το πλήρες πρόβλημα των 3 κλάσεων, με την προσέγγιση ένας-εναντίον-ενός (one-vs-one) και καταμέτρηση ψήφων. Μπορείτε να αξιοποιήσετε τις σχετικές αυτοματοποιημένες λειτουργίες που έχουν κάποιες βιβλιοθήκες SVM. Θέτοντας το C=1 και ακολουθώντας πρωτόκολλο 5-fold cross validation, να υπολογίστε τη μέση τιμή του σφάλματος ταξινόμησης χρησιμοποιώντας α) τα 5 πρώτα χαρακτηριστικά όπως και παραπάνω, και β) όλα τα διαθέσιμα χαρακτηριστικά. Σχολιάστε τα αποτελέσματα. Υπολογίστε για κάθε περίπτωση τον πίνακα σύγχυσης (confusion matrix) της ταξινόμησης και σχολιάστε ποιες κλάσεις ομοιάζουν περισσότερο.

Οδηγίες: *Για τον SVM μπορείτε να χρησιμοποιήσετε τη βιβλιοθήκη LIBSVM* (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) *που παρέχει διεπαφές για τις περισσότερες γλώσσες προγραμματισμού ή όποια άλλη υλοποίηση επιθυμείτε αρκεί να δηλώσετε στην αναφορά ή τον κώδικά σας την πηγή της υλοποίησης που χρησιμοποιήσατε. Υποβάλετε τις απαντήσεις σας στην πλατφόρμα του e-class στην ενότητα εργασίες, πριν τη λήξη της προθεσμίας υποβολής. Η υποβολή σας πρέπει να αποτελείται από* ***ένα μόνο συμπιεσμένο αρχείο*** *(.rar, .zip κλπ.) το οποίο θα περιέχει όλα τα απαραίτητα αρχεία για την υποβολή σας. Καταγράφετε τις απαντήσεις σας σε μία αναφορά που θα αποστείλετε σε μορφή pdf ή docx ή εναλλακτικά ενσωματωμένες σε Python notebooks ή Matlab live scripts. Οι αποδεκτές γλώσσες προγραμματισμού είναι Matlab ή Python, όπου και θα πρέπει να συμπεριλάβετε στο συμπιεσμένο αρχείο και τα αρχεία του πηγαίου κώδικα (αρχεία .m, .py κλπ) με τα απαραίτητα σχόλια για την τεκμηρίωση εντός του κώδικα. υποβάλλοντας το πολύ ένα αρχείο για κάθε άσκηση. Σε περίπτωση που χρησιμοποιήσετε Matlab, μπορείτε να συμπεριλάβετε τον κώδικα για όλες τις ασκήσεις σε ένα αρχείο, χωρίζοντας τις απαντήσεις σε διαφορετικά cells. Επίσης, για τις ασκήσεις που περιλαμβάνουν κώδικα, αντί απάντησης στην αναφορά μπορείτε να υποβάλετε Matlab live scripts ή Python notebooks κατάλληλα δομημένα ώστε να περιλαμβάνουν και τις απαντήσεις/σχολιασμούς που ζητούνται.*

Ξάνθη, 5/11/2023