Εργαστηριακή Άσκηση για το μάθημα της Θεωρίας Αποφάσεων

Χειμερινό Εξάμηνο 2023-24

Θέμα:"Σχεδιασμός Ταξινομητή για Κατηγοριοποίηση και Πρόβλεψη Καρκίνου του Μαστού"

Σε αυτή την εργαστηριακή άσκηση καλείστε να χρησιμοποιήσετε ένα σύνολο δεδομένων για να εκπαιδεύσετε ένα σύνολο ταξινομητών για την κατηγοριοποίηση και πρόβλεψη του καρκίνου του μαστού.

Το σύνολο δεδομένων που σας δίνετε στο αρχείο BreastTissue.xls παρέχει ένα σύνολο βιοιατρικών μετρήσεων για κάθε ασθενή καθώς και την πληροφορία για τον αν ο ιστός του μαστού είναι Καρκίνωμα (Carcinoma), Ινο-αδένωμα (Fibro-adenoma), Μαστοπάθεια (Mastopathy), Αδενικός (Glandular), Συνδετικός (Connective), Λιπώδης (Adipose). Κάθε γραμμή στο αρχείο αυτό περιέχει πληροφορία για διαφορετικό ασθενή. Η πρώτη στήλη αφορά τον κωδικό της κάθε καταγραφής ενώ η δεύτερη (status) δείχνει τη διάγνωση, δηλ. αν η καταγραφή αφορά ασθενή με: Καρκίνωμα (Carcinoma), Ινο-αδένωμα (Fibro-adenoma), Μαστοπάθεια (Mastopathy), Αδενικός (Glandular), Συνδετικός (Connective), Λιπώδης (Adipose).

Car	Carcinoma
Fad	Fibro-adenoma
Mas	Mastopathy
Gla	Glandular
Con	Connective
Adi	Adipose

Όλες οι άλλες στήλες αντιστοιχούν σε βιοιατρικές μετρήσεις που θα πρέπει να χρησιμοποιήσετε σαν εισόδους στους ταξινομητές που θα δημιουργήσετε. Για περισσότερες πληροφορίες για τα δεδομένα αυτά μπορείτε να δείτε την αναλυτική περιγραφή τους στην βάση δεδομένων μηχανικής μάθησης UCI (https://archive.ics.uci.edu/dataset/192/breast+tissue) από την οποία προέρχονται. Σε αυτό το link, θα βρείτε μια σύντομη περιγραφή του αρχείου (χαρακτηριστικά, πλήθος, κλπ.)

Ερώτημα 1. Προεπεξεργασία δεδομένων (10 μόρια)

Το εύρος τιμών των δεδομένων που σας έχουν δοθεί διαφέρει σημαντικά ανά χαρακτηριστικό. Για αυτό τον λόγο, για να μην υπερεκτιμηθεί η συνεισφορά κάποιου χαρακτηριστικού έναντι άλλων, θα πρέπει πριν την επεξεργασία των χαρακτηριστικών εισόδου να κανονικοποιηθούν στο εύρος [-1,1]. Χρησιμοποιήστε το matlab (ή όποιο

περιβάλλον προγραμματισμού επιθυμείτε) τόσο για το διάβασμα του αρχείου που σας δίνεται όσο και για την κανονικοποίηση των δεδομένων εισόδου στο εύρος τιμών [-1,1]. Ακόμη θα πρέπει να αντιστοιχίσετε την κλάση του ασθενούς σε μία τιμή. Για αυτό το λόγο να αντιστοιχίσετε κάθε ασθενή, στην κατηγορία που ανήκει, σύμφωνα με τον παρακάτω πίνακα.

Car	Carcinoma	1
Fad	Fibro-adenoma	2
Mas	Mastopathy	3
Gla	Glandular	4
Con	Connective	5
Adi	Adipose	6

Η στήλη αυτή δεν χρειάζεται κανονικοποίηση.

Ερώτημα 2. (20 μόρια) Να κάνετε μια σύντομη παρουσίαση του ταξινομητή Support Vector Machine και του Naive Bayes ταξινομητή. Να αντλήσετε υλικό από το διαδίκτυο και να αναφέρετε τις πηγές σας. Να κάνετε μια σύντομη σύγκριση των παραπάνω ταξινομητών με τον ΚΝΝ ταξινομητή.

Ερώτημα 3. (30 μόρια) Με χρήση της μεθόδου 5-fold cross validation και του matlab (ή όποιο περιβάλλον προγραμματισμού επιθυμείτε), εκπαιδεύστε τους παρακάτω τρεις ταξινομητές και παρουσιάστε την απόδοσή τους:

- Support Vector Machine (με Radial Basis Function σαν kernel function):
    - Ρυθμίστε την παράμετρο C με διαδοχική αναζήτηση του βέλτιστου C στο διάστημα 1-200 με βήμα 5 και χρήση γραμμικών SVM. Στη συνέχεια, ρυθμίστε την παράμετρο γ με χρήση του βέλτιστου C που βρέθηκε από πριν, και διαδοχική αναζήτηση του βέλτιστου γ στο διάστημα 0-10 με βήμα 0.5 και χρήση RBFSVM.
- Ταξινομητής ΚΝΝ-Κ Κοντινότερου Γείτονα
    - Ρυθμίστε την παράμετρο Κ με διαδοχική αναζήτηση της βέλτιστης τιμής στο διάστημα 3-15.
- Αφελής Μπεϋζιανός (Naive Bayes) Ταξινομητής

Παρουσιάστε για κάθε ταξινομητή την μέση απόδοσή του με χρήση 5 fold cross validation σε σχέση με την μετρική του γεωμετρικού μέσου (Geometric Mean) της ευαισθησίας (Sensitivity) και της ειδίκευσής (Specificity) του:

Geometric Mean = sqrt (Sensitivity * Specificity)

Η μετρική αυτή χρησιμοποιείται για προβλήματα ταξινόμησης όπου παραδείγματα εκπαίδευσης της μίας κλάσης είναι περισσότερα από τα παραδείγματα εκπαίδευσης της άλλης κλάσης.

Στη συνέχεια, παρουσιάστε τα ενδιάμεσα αποτελέσματα που πήρατε από τα πειράματα για την ρύθμιση των παραμέτρων των αλγορίθμων. Γιατί η κάθε μέθοδος ταξινόμησης δίνει διαφορετικά αποτέλεσματα; Ποιά μέθοδο προτείνετε εσείς και γιατί;

Ερώτημα 4 (10 μόρια). Να κάνετε μια σύντομη παρουσίαση της μεθόδου Student t-test. Να αντλήσετε υλικό από το διαδίκτυο και να αναφέρετε τις πηγές σας.

Ερώτημα 5. (30 μόρια) Για τα δεδομένα που σας δίνονται διατάξτε τα χαρακτηριστικά εισόδου με χρήση της μεθόδου student t-test με βάση την σημαντικότητά τους στην πρόβλεψη του καρκίνου του μαστού. Στη συνέχεια, κρατείστε τα 4 πιο σημαντικά χαρακτηριστικά, και επαναλάβετε την εκπαίδευση του βέλτιστου ταξινομητή που βρήκατε από προηγούμενο ερώτημα. Σχολιάστε την τελική απόδοση.

ΠΑΡΑΤΗΡΗΣΕΙΣ:

- Η εργασία αποτελείται από 2 μέρη. Στο 1^{ο} μέρος θα πρέπει να απαντήσετε τα ερωτήματα 1 και 2. Η αναφορά του 1^{ου} μέρους της εργασίας πρέπει να αναρτηθεί στη σελίδα του μαθήματος στο e-class, στις 17/12/2023 μέχρι τις 23.55.
- Η αναφορά του 2^{ου} μέρους της εργασίας πρέπει να αναρτηθεί στη σελίδα του μαθήματος στο e-class την παραμονή της εξέτασης του μαθήματος, μέχρι τις 23.55 (μετά από αυτή την ώρα το σύστημα θα κλείσει).
- Για την αναφορά της εργασίας σας, θα χρησιμοποιήσετε το αρχείο της εκφώνησης. Μετά από κάθε υποερώτημα θα έχετε την απάντησή σας, στην οποία θα περιγράφετε τη μεθοδολογία, τα εργαλεία που χρησιμοποιήσατε κλπ. Στα ερωτήματα που αναπτύσσετε κώδικα, θα πρέπει να έχετε κάποια screen shots από τα τρεξίματα και σε χωριστό αρχείο τον κώδικα. Το όνομα του αρχείου με τον κώδικα, θα πρέπει να περιέχει τον αριθμό του αντίστοιχου υποερωτηματος.
- Θα αναρτήσετε ένα .zip αρχείο που θα περιέχει την αναφορά και τον κώδικα. Στο όνομα του αρχείου πρέπει να υπάρχει το επώνυμό σας και το αρχικό γράμμα του ονόματός σας και ο ΑΜ.