ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

UNIVERSITY OF PATRAS
SCHOOL OF ENGINEERING
Department of Computer Engineering & ***Informatics***
Division of Applications and Foundations
of Computer Science
Pattern Recognition Laboratory
Director: S. *Likothanassis, Professor,*
e-mail : *likothan@ceid.upatras.gr*
URL: http://prlab.ceid.upatras.gr/~likothan/

Εργαστηριακή Άσκηση
για το μάθημα Θεωρία Αποφάσεων

Διδάσκοντες:

Σ. Λυκοθανάσης, Α. Ανδρικόπουλος

2019-2020

ΑΝΤΙΚΕΙΜΕΝΟ:

"Σχεδιασμός και Αξιολόγηση Συστήματος, από Δημογραφικά και Εργαστηριακά Δεδομένα, για Πρόβλεψη της Ασθένειας του Διαβήτη"

Σε αυτή την εργαστηριακή άσκηση καλείστε να χρησιμοποιήσετε ένα σύνολο δεδομένων για να εκπαιδεύσετε και αξιολογήσετε ένα σύνολο ταξινομητών για την πρόβλεψη του σακχαρώδη διαβήτη.

Το σύνολο δεδομένων που σας δίνεται στο αρχείο που επισυνάπτεται, παρέχει ένα σύνολο δημογραφικών και εργαστηριακών μετρήσεων για κάθε ασθενή καθώς και την πληροφορία για το αν αυτός πάσχει ή όχι από διαβήτη. Κάθε γραμμή στο αρχείο αυτό περιέχει πληροφορία για διαφορετικό ασθενή. Η πρώτη στήλη αναφέρεται σε εγκυμοσύνες (άρα υπονοεί και το φύλλο), η προτελευταία την ηλικία (από 21 ετών και μεγαλύτερες), ενώ η τελευταία δείχνει το αν η καταγραφή αφορά μέτρηση ασθενούς ατόμου (πάσχει από διαβήτη) ή υγιούς ατόμου (τιμή false). Όλες οι άλλες στήλες αντιστοιχούν σε εργαστηριακές μετρήσεις (όλα τα χαρακτηριστικά είναι αριθμοί) που θα πρέπει επίσης 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://github.com/LamaHamadeh/Pima-Indians-Diabetes-DataSet-UCI) από την οποία προέρχονται.

Ερώτημα 1. Προεπεξεργασία δεδομένων

Να αναφέρετε, πόσα είναι τα χαρακτηριστικά κάθε δείγματος, πόσα δείγματα εκπαίδευσης περιέχει το αρχείο και πόσα δείγματα έχει η κάθε κατηγορία.

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 όποια άλλη εφαρμογή θέλετε) τόσο για το διάβασμα του αρχείου που σας δίνεται όσο και για την κανονικοποίηση των δεδομένων εισόδου στο εύρος τιμών [-1,1].

Ερώτημα 2. Στο μάθημα συζητήθηκε εκτεταμένα ο ταξινομητής Bayes. Στη βιβλιογραφία, υπάρχει μια παραλλαγή του που λέγεται Αφελής Ταξινομητής Bayes (Naïve Bayes), με την υπόθεση ότι τα χαρακτηριστικά είναι στατιστικά ανεξάρτητα. Αναζητήστε τη σχετική βιβλιογραφία στο Internet, και να κάνετε μια σύντομη παρουσίαση του αλγορίθμου. Στη συνέχεια να κάνετε μια σύγκριση με τον Ταξινομητή Bayes.

Ερώτημα 3. Με χρήση της μεθόδου 5-fold cross validation, να εκπαιδεύσετε τον Naïve Bayes ταξινομητή, να παρουσιάστε και να σχολιάσετε την απόδοσή του. Μπορείτε να χρησιμοποιήσετε κατάλληλες συναρτήσεις του matlab ή οποιαδήποτε εφαρμογή επιθυμείτε (π.χ. το WEKA (http://www.cs.waikato.ac.nz/ml/weka/) ή να

υλοποιήσετε δικό σας κώδικα). Για την αξιολόγηση της απόδοσης του ταξινομητή να χρησιμοποιήσετε τις μετρικές του ερωτήματος 4, παρακάτω.

Ερώτημα 4. Με χρήση της μεθόδου 5-fold cross validation και του matlab (ή οποιαδήποτε εφαρμογή επιθυμείτε π.χ. το Microsoft Azure Machine Learning Studio https://studio.azureml.net/), να εκπαιδεύσετε τους παρακάτω ταξινομητές, να παρουσιάστε και να σχολιάσετε την απόδοσή τους:

- Support Vector Machines (με Radial Basis Function 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 Κ-Κοντινότερου Γείτονα
    - Ρυθμίστε την παράμετρο Κ με διαδοχική αναζήτηση της βέλτιστης τιμής στο διάστημα 3-10.

Ποιο συγκεκριμένα ζητείται να παρουσιάσετε για κάθε ταξινομητή την μέση απόδοση 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 αποτελέσματα; Με βάση τα παραπάνω αποτελέσματα ποιά από τις μεθόδους προτείνετε εσείς να χρησιμοποιηθεί για το παραπάνω πρόβλημα και γιατί;

Παρατηρήσεις:

- Η αξιολόγηση του Β΄ μέρους της εργασίας θα έχει βαρύτητα 20% του τελικού βαθμού.
- Η εργασία σας θα πρέπει να παραδοθεί (με ανάρτηση στο e-class) την παραμονή της εξέτασης του μαθήματος στις 23.55.
- Για απορίες σχετικές με την εργασία σας μπορείτε να επικοινωνείτε με email με τον κο Θ. Αμοργιανιώτη (amorgianio@ceid.upatras.gr) και τον κο Α. Ανδριόπουλο (a.andriopoulos@upatras.gr).
- Σύντομα θα ανακοινωθούν και ώρες γραφείου.