Θεωρία Αποφάσεων Εργαστηριακή άσκηση

Όνομα/επώνυμο: Κωνσταντίνος Καραΐσκος

Α.Μ.: 1072636

Έτος: 4ο

**Ερώτημα 1.** Προεπεξεργασία δεδομένων

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

Η δεύτερη στήλη περιέχει το φύλο του ανθρώπου που συμμετείχε στο δείγμα. Στο αρχείο όμως είναι σημειωμένη με Male για αρσενικό και Female για θηλυκό. Προκειμένου να την χρησιμοποιήσουμε σαν είσοδο θα πρέπει να αντιστοιχίσετε το Male με την τιμή 0 και το Female με την τιμή 1.

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].

Το αρχείο Indian Liver Patient Dataset περιέχει 583 δείγματα από τα οποία αφαιρούμε τα 4 τα οποία περιέχουν μηδενικές τιμές σε κάποιο από τα χαρακτηριστικά τους και καταλήγουμε με 579 δείγματα για να εκπαιδεύσουμε και να ελέγξουμε τον ταξινομητή μας. Το καθ’ ένα από αυτά τα δείγματα έχει 10 χαραακτηριστικά (Age, Gender,…), και μπορεί να ανήκει σε μία από τις δύο κλάσεις ω1: έχει καρκίνο του ήπατος, ω2, δεν έχει καρκίνο του ήπατος.

Αναθέτουμε στη στήλη Gender όπου male 0 και όπου female 1 (μέσω του excel. Για την κανονικοποίηση των δεδομένων υπολογίζουμε την μέση τιμή κάθε στήλης χαρακτηριστικών όπως και τη διαφορά d = max\_value – min\_value. Για κάθε τιμή της εκάστοτε στήλης χαρακτηριστικών υπολογίζουμε new\_value= (old\_value – E(x))/d και γράφουμε τα κανονικοποιημένα δεδομένα σε ένα καινούριο csv αρχείο από όπου και θα τα διαβάζουμε για της εκπαίδευση των ταξινομητών στη συνέχεια.

Το head() του αρχείου με κανονικοποιημένα δεδομένα:

Age Gender TB DB ... TP ALB AGR Dataset

0 0.2351 1 -0.0351 -0.0711 ... 0.0461 0.0351 -0.0188 1

1 0.2002 0 0.1017 0.2044 ... 0.1476 0.0134 -0.0828 1

2 0.2002 0 0.0534 0.1330 ... 0.0751 0.0351 -0.0228 1

3 0.1537 0 -0.0310 -0.0558 ... 0.0461 0.0568 0.0212 1

4 0.3165 0 0.0078 0.0258 ... 0.1186 -0.1605 -0.2188 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.

Naïve Bayes

Ο Naïve Bayes είναι ένας εποπτευόμενος αλγόριθμος μάθησης, ο οποίος βασίζεται στο θέωρημα Bayes και χρησιμοποιείται για την επίλυση προβλημάτων ταξινόμησης.

Έστω ότι το πρόβλημα αυτό αντιπροσωπεύεται από ένα διάνυσμα χαρακτηριστικών x=(x1,…,xn), ανεξάρτητα μεταξύ τους.

Έστω ένα σύνολο φυσικών κλάσεων {ω1,ω2,…,ωc}. Έστω p(x|ωj) η συνάρτηση πυκνότητας πιθανότητας για το x υπό την συνθήκη ότι η πραγματική κατάσταση της φύσης είναι η ωj. Τότε, από τον τύπο του Bayes, μπορώ να βρω την εκ των υστέρων πιθανότητα:

P(ωj|x) =

ο παρονομαστής είναι ανεξάρτητος του ωj και των τιμών των χαρακτηριστικών xi, i=1,…n. Ο αριθμητής είναι ισοδύναμος με το μοντέλο κοινής πιθανότητας: p(ωj,x1,…,xn), το οποίο από τον κανόνα της αλυσίδας ισούται με:

p(ωj,x1,…,xn) = p(x1,…,xn,ωj)

= P(x1|x2,…,xn,ωj)\*p(x2,…,xn,ωj)

= p(x1|x2,…xn,ωj)\*p(x2|x3,…,xn,ωj)\*p(x3,…,xn,ωj)

= …

= p(x1,x2…,xn,ωj)\*p(x2|x3,...,xn,ωj)\*…\*p(xn-1|xnωj)\*p(xn|ωj)\*p(ωj)

Θεωρώντας ότι όλα τα χαρακτηριστικά στο διάνυσμα x είναι αμοιβαίως ανεξάρτητα, για μια κλάση ωj, έχουμε:

p(xi|xi+1,xn,ωj) = p(xi|ωj)

Άρα, από τις δύο παραπάνω εξισώσεις, έχουμε:

p(ωj|x1,…,xn) p(ωj,x1,…,xn)

p(ωj)p(x1|ωj)p(x2|ωj)p(x3|ωj)…

p(ωj)

Άρα, σύμφωνα με τις παραπάνω παραδοχές ανεξαρτησίας, η υπό συνθήκη κατανομή στην μεταβλητή κλάσης ωj, είναι:

όπου, .

Τέλος, ο Naïve Bayes συνδυάζει το παραπάνω μοντέλο πιθανοτήτων με ένα κανόνα απόφασης, ώστε να επιλέγουμε την υπόθεση που είναι πιο πιθανή, ελαχιστοποιώντας την πιθανότητα εσφαλμένης ταξινόμησης. Ο κανόνας αυτός εκχωρεί μία ετικέτα κλάσης ŷ = Ck, ως εξής:

ŷ =

Ταξινομητής Bayes

* Έστω διάνυσμα χαρακτηριστικών x διάστασης d
* σύνολο φυσικών καταστάσεων (κλάσεις) {ω1,…,ωc}
* σύνολο δυνατών δράσεων {α1,…,αI}
* συνάρτηση κόστους λ(αi|ωj): κόστος της δράσης αi όταν η πραγματική φυσική κατάσταση είναι ωj

Ο κανόνας του Bayes σε διανυσματική μορφή : ,

όπου

Έστω ότι παρατηρείται ένα συγκεκριμένο x και λαμβάνεται η ενέργεια αi. Αν η πραγματική κατάσταση της φύσης είναι ωj, θα έχουμε κόστος ίσο με λ(αi|ωj). Αφού p(ωj|x) είναι η πιθανότητα η πραγματική κατάσταση της φύσης να είναι η ωj, το αναμενόμενο κόστος ή συνολικό ρίσκο, που σχετίζεται με την ενέργεια αi, θα είναι:

R(αi|x) =

Σύμφωνα με τον κανόνα απόφασης του Bayes, για να ελαχιστοποιηθεί το συνολικό ρίσκο, υπολογίζουμε το συνολικό ρίσκο με βάση τον παραπάνω τύπο για κάθε i = 1,2,…I, και μετά επιλέγουμε την ενέργεια αi για την οποία το συνολικό ρίσκο R(αi|x) είναι ελάχιστο.

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

3. Οι αποδόσεις του ταξινομητή GaussianNB εκπαιδευμένου με k-fold cross validation είναι:

---------------------------------------------------------

5-Fold cross validation scores (Gaussian Naive Bayes):

[0.5862069 0.61206897 0.42241379 0.5 0.70434783]

0.57 accuracy with standard deviation of 0.10

---------------------------------------------------------

, έχουμε δηλαδή, σύμφωνα με την default scoring method, 0.57% ακρίβεια σωστής πρόβλεψης.

>>> Χρησιμοποιώντας την make\_score για να μετατρέψουμε τις μετρικές sensitivity\_score και specificity\_score σε κλητά αντικείμενα για την cross\_val\_score για την αξιολόγηση του μοντέλου μας, παίρνουμε τα εξής αποτελέσματα:

---------------------------------------------------------

Specificity scores:

[0.90909091 0.90909091 0.93939394 1. 1. ]

0.95 accuracy with standard deviation of 0.04

---------------------------------------------------------

Sensitivity scores:

[0.45783133 0.4939759 0.21686747 0.30120482 0.58536585]

0.41 accuracy with standard deviation of 0.13

---------------------------------------------------------

>>> Ο ταξινομητής μας είναι δηλαδή κατα 0.95% ικανός να ταξινομεί αυτούς που δεν έχουν καρκίνο του ήπατος στους υγιείς και 0.41% ικανός να ταξινομεί αυτούς που έχουν καρκίνο του ήπατος στου ασθενείς.

>>>Τέλος, βλέπουμε οτι χρησιμοποιώντας τη μετρική του γεωμετρικού μέσου (με το ίδιο τρόπο με τις δύο παραπάνω), έχουμε μια βελτίωση στην απόδωση του ταξινομητή της τάξης του 0.05:

---------------------------------------------------------

Geometric mean scores:

[0.64514362 0.67012611 0.45135794 0.5488213 0.76509206]

0.62 accuracy with standard deviation of 0.11

---------------------------------------------------------