Θεωρία Αποφάσεων

Εργασία Μέρος Α’ & Β’

Σεργιάννης Παρασκευάς – Βασίλειος

ΑΜ: 1067467

**Ερώτημα 1.**

Τα χαρακτηριστικά κάθε δείγματος είναι 11 και είναι τα ακόλουθα:

1. Age: Age of the patient  
2. Gender: Gender of the patient  
3. TB: Total Bilirubin  
4. DB: Direct Bilirubin  
5. Alkphos: Alkaline Phosphotase  
6. Sgpt: Alamine Aminotransferase  
7. Sgot: Aspartate Aminotransferase  
8. TP: Total Protiens  
9. ALB: Albumin  
10. A/G: Ratio Albumin and Globulin Ratio  
11. Selector field used to split the data into two sets (labeled by the experts)

*(Note: Στα columns έχω περάσει λίγο διαφορετικά τις ονομασίες των πεδίων επειδή δεν είχα παρατηρήσει οτι στο αριστερό μέρος βρίσκονται οι συντομογραφίες των χαρακτηριστικών.)*

Με τα 10 πρώτα columns να είναι τα γνωρίσματα του ασθενή, και το τελευταίο το αν πάσχει ή όχι από ασθένεια στο συκώτι.

Και τα δείγματα εκπαίδευσης που περιέχει το αρχείο είναι 583.

Στο αρχείο *preprocessing.py,* έχει γίνει η προεπεξεργασία των δεδομένων, όπου:

1. Οι τιμές Male και Female έχουν αντικατασταθεί με τις τιμές 0 και 1 αντίστοιχα
2. Στην στήλη όπου πάσχει ή δεν πάσχει ο ασθενής, το 2 έχει αντικατασταθεί με το 0
3. Έχει γίνει normalization των τιμών του dataset στο διάστημα [-1,1]

**Ερώτημα 2.**

Οι **Naive Bayes classifiers** είναι μια οικογένεια απλών «[πιθανοτικών ταξινομητών](https://en.wikipedia.org/wiki/Probabilistic_classification)» που βασίζονται στην εφαρμογή [του θεωρήματος του Bayes](https://en.wikipedia.org/wiki/Bayes%27_theorem) με ισχυρές (αφελείς) υποθέσεις. Είναι από τα απλούστερα μοντέλα [δικτύων Bayes,](https://en.wikipedia.org/wiki/Bayesian_network) αλλά σε συνδυασμό με [την εκτίμηση της πυκνότητας του πυρήνα](https://en.wikipedia.org/wiki/Kernel_density_estimation) , μπορούν να επιτύχουν υψηλά επίπεδα ακρίβειας.

Πιο συγκεκριμένα, το Naive Bayes είναι μια απλή τεχνική για την κατασκευή ταξινομητών: μοντέλα που εκχωρούν ετικέτες κλάσεων σε στιγμιότυπα προβλημάτων, που αναπαρίστανται ως διανύσματα τιμών [χαρακτηριστικών](https://en.wikipedia.org/wiki/Feature_vector) , όπου οι ετικέτες κλάσεων προέρχονται από κάποιο πεπερασμένο σύνολο. Δεν υπάρχει ένας μόνο [αλγόριθμος](https://en.wikipedia.org/wiki/Algorithm) για την εκπαίδευση τέτοιων ταξινομητών, αλλά μια οικογένεια αλγορίθμων που βασίζεται σε μια κοινή αρχή: όλοι οι απλοί ταξινομητές Bayes υποθέτουν ότι η τιμή ενός συγκεκριμένου χαρακτηριστικού είναι [ανεξάρτητη](https://en.wikipedia.org/wiki/Independence_(probability_theory)) από την τιμή οποιουδήποτε άλλου χαρακτηριστικού, δεδομένης της μεταβλητής κλάσης. Για παράδειγμα, ένα φρούτο μπορεί να θεωρηθεί ότι είναι μήλο εάν είναι κόκκινο, στρογγυλό και έχει διάμετρο περίπου 10 cm. Ένας αφελής ταξινομητής Bayes θεωρεί ότι καθένα από αυτά τα χαρακτηριστικά συμβάλλει ανεξάρτητα στην πιθανότητα ότι αυτό το φρούτο είναι ένα μήλο, ανεξάρτητα από τις [συσχετίσεις](https://en.wikipedia.org/wiki/Correlation_and_dependence) μεταξύ των χαρακτηριστικών χρώματος, στρογγυλότητας και διαμέτρου.

Οι ταξινομητές Naive Bayes και οι ταξινομητές Bayes είναι και οι δύο πιθανολογικά μοντέλα που κάνουν προβλέψεις με βάση την πιθανότητα να συμβούν ορισμένα γεγονότα. Η κύρια διαφορά μεταξύ των δύο είναι ότι ένας απλός ταξινομητής Bayes υποθέτει ότι όλα τα χαρακτηριστικά είναι ανεξάρτητα το ένα από το άλλο, ενώ ένας ταξινομητής Bayes δεν κάνει αυτή την υπόθεση.

Ένας Naive Bayes ταξινομητής, είναι ένας γρήγορος και απλός ταξινομητής που χρησιμοποιείται συχνά για εργασίες ταξινόμησης με μεγάλο αριθμό χαρακτηριστικών ή μεγάλο σύνολο δεδομένων. Λειτουργεί καλά όταν τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους, αλλά μπορεί να μην έχει καλή απόδοση όταν τα χαρακτηριστικά εξαρτώνται το ένα από το άλλο.

Ένας ταξινομητής Bayes, από την άλλη πλευρά, δεν υποθέτει ότι τα χαρακτηριστικά είναι ανεξάρτητα το ένα από το άλλο. Αυτό του επιτρέπει να χειρίζεται τις εξαρτήσεις μεταξύ των χαρακτηριστικών και να κάνει πιο ακριβείς προβλέψεις. Ωστόσο, μπορεί να είναι πιο αργός και πιο περίπλοκος από έναν αφελή ταξινομητή Bayes.

Γενικά, ένας ταξινομητής Bayes μπορεί να έχει καλύτερη απόδοση από έναν Naive Bayes ταξινομητή όταν τα χαρακτηριστικά εξαρτώνται το ένα από το άλλο, αλλά ένας απλός ταξινομητής Bayes μπορεί να είναι πιο γρήγορος και απλούστερος στην εφαρμογή. Ένα επιπλέον πλεονέκτημα του Naive Bayes είναι ότι απαιτεί έναν μικρό αριθμό δεδομένων εκπαίδευσης για την εκτίμηση των παραμέτρων που είναι απαραίτητες για την ταξινόμηση. Αυτό συμβαίνει επειδή ένας ταξινομητής Naive Bayes κάνει ισχυρές υποθέσεις σχετικά με την ανεξαρτησία των χαρακτηριστικών, γεγονός που του επιτρέπει να εκτιμήσει τις απαραίτητες παραμέτρους χρησιμοποιώντας μικρότερο όγκο δεδομένων.

**Ερώτημα 3.**

Με βάση το μοντέλο στο αρχείο *model.py,* στο οποίο έχει χρησιμοποιηθεί ο Bernoulli Naive Bayes, ο οποίος είναι κατάλληλος για binary classification με μικρό αριθμό χαρακτηριστικών (10 σε αυτήν την περίπτωση) και σχετικά λίγα δείγματα (582) (Γενικά είναι ιδιαίτερα αποτελεσματικός όταν ο αριθμός των χαρακτηριστικών είναι μεγάλος σε σχέση με τον αριθμό των δειγμάτων), έχουμε τα ακόλουθες τιμές των μετρικών που ζητούνται:

Accuracy: [0.6239316239316239, 0.7094017094017094, 0.717948717948718, 0.6637931034482759, 0.7241379310344828]

Geometric mean: [0.3742588494760676, 0.44818474451657914, 0.5913223842897901, 0.5231483637805969, 0.39128094151619297]

Sensitivity: [0.8170731707317073, 0.8369565217391305, 0.8625, 0.8421052631578947, 0.9186046511627907]

Specificity: [0.17142857142857143, 0.24, 0.40540540540540543, 0.325, 0.16666666666666666]

Mean accuracy: 0.69

Mean geometric mean: 0.47

Mean sensitivity: 0.86

Mean specificity: 0.26

(Η τιμή της ακρίβειας δεν ζητείται στο ερώτημα 4 αλλά την πρόσθεσα γιατί την θεωρώ βασική για την αξιολόγηση του μοντέλου)

Mean accuracy: Η μέση ακρίβεια 0,69 υποδεικνύει ότι το μοντέλο κάνει σωστές προβλέψεις περίπου στο 69% του χρόνου κατά μέσο όρο. Αυτή δεν είναι ιδιαίτερα υψηλή ακρίβεια, αλλά μπορεί να είναι επαρκής ανάλογα με τις απαιτήσεις της εργασίας.

Mean geometric mean: Η τιμή 0,47 είναι ένα μέτρο της συνολικής απόδοσης του μοντέλου που λαμβάνει υπόψη τόσο την ευαισθησία όσο και την ειδικότητα του μοντέλου. Η τιμή 0,47 είναι σχετικά χαμηλή, γεγονός που μπορεί να υποδηλώνει ότι το μοντέλο δεν έχει καλή απόδοση συνολικά.

Mean sensitivity: Η μέση ευαισθησία 0,86 υποδεικνύει ότι το μοντέλο ανιχνεύει θετικές περιπτώσεις (π.χ. ταυτοποίηση ατόμων με ασθένεια) με υψηλό επίπεδο ακρίβειας. Αυτό είναι ένα καλό αποτέλεσμα και μπορεί να είναι σημαντικό σε ορισμένες εφαρμογές.

Mean specificity: Η μέση εξειδίκευση 0,26 δείχνει ότι το μοντέλο δεν είναι πολύ καλό στη διάκριση των αρνητικών περιπτώσεων (π.χ. στον εντοπισμό ατόμων χωρίς ασθένεια) από τις θετικές περιπτώσεις. Αυτή είναι μια σχετικά χαμηλή εξειδίκευση και μπορεί να είναι ανησυχητική ανάλογα με τις απαιτήσεις της εργασίας.

**Ερώτημα 4**

Διαφορετικοί ταξινομητές μπορούν να οδηγήσουν δωσουν διαφορετικά αποτελέσματα επειδή χρησιμοποιούν διαφορετικούς αλγόριθμους και κάνουν διαφορετικές υποθέσεις για να εκτελέσουν τις προβλέψεις τους. Στην περίπτωσή μας, ο Naive Bayes, ο k-Nearest Neighbors και ο Support Vector Machines είναι όλοι διαφορετικοί αλγόριθμοι και καθένας από αυτούς κάνει διαφορετικές υποθέσεις σχετικά με τη σχέση μεταξύ των χαρακτηριστικών και των μεταβλητών-στόχων. Ως αποτέλεσμα, κάθε μέθοδος μπορεί να έχει διαφορετικά πλεονεκτήματα και αδυναμίες σε διαφορετικά σύνολα δεδομένων, οδηγώντας σε διαφορετικά αποτελέσματα απόδοσης.

Επίσης κάνω μία παραδοχή στο SVM, επειδή μου βγάζει error οτι η τιμή του gamma δεν μπορει να είναι 0, προσεγγιστικά έχω δώσει αρχικά την τιμή 0.001 στο gamma το οποιο κάνει iteration μεχρι το 10.001 μα βήμα 0,5.

Τα αποτελέσματα που παίρνουμε από κάθε μοντέλο είναι:

*Naive Bayes:*

*Mean accuracy: 0.69*

*Mean geometric mean: 0.47*

*Mean sensitivity: 0.86*

*Mean specificity: 0.26*

*KNN:*

*Best n\_neighbors: 5*

*Mean accuracy: 0.70*

*Mean geometric mean: 0.38*

*Mean sensitivity: 0.82*

*Mean specificity: 0.19*

*SVM:*

*Best C: 1*

*Best\_gamma: 3.001*

*Mean accuracy: 0.73*

*Mean geometric mean: 0.46*

*Mean sensitivity: 0.87*

*Mean specificity: 0.25*

Με βάση τα αποτελέσματα που έχουμε, ο ταξινομητής Support Vector Machine (SVM) φαίνεται να έχει την καλύτερη απόδοση όσον αφορά το accuracy, με τιμή 0,73. Ωστόσο, η ακρίβεια είναι μόνο μία μέτρηση και μπορεί να μην είναι πάντα ο καλύτερος δείκτης απόδοσης, ειδικά όταν πρόκειται για μη ισορροπημένα σύνολα δεδομένων όπου η μία κλάση είναι πολύ πιο συχνή από την άλλη. Παρ’ολα αυτά επειδή οι τιμές του sensitivity, specificity και geometric mean δεν μας δίνουν ξεκαθαρη εικόνα, γι αυτό προσέθεσα και το accuracy σαν μετρική για να μππορούμε να δούμε απευθείας ποια μετρική κάνει τις καλύτερες προβλέψεις.

**Ερώτημα 5**

Στη συνέχεια έχω χρησιμοποιήσει το t-test , κρατώντας τα 5 σημαντικότερα χαρακτηριστικά και χρησιμοποίησα πάλι το SVM μοντέλο.

Η τελική απόδοση είναι η επακόλουθη:

*SVM:*

*Best C: 1*

*Best\_gamma: 9.501*

*Mean accuracy: 0.72*

*Mean geometric mean: 0.40*

*Mean sensitivity: 0.81*

*Mean specificity: 0.22*

Παρατηρούμε οτι τα αποτελέσματα είναι σχεδόν τα ίδια μεταά την εφαρμογή του t-test χωρις να υπαρχει σημαντικη διαφορα σε σχεση με προηγουμένως. Αυτό σημαίνει πως οι έξτρα τιμές που είχαμε για τους ασθενείς έπαιξαν ρόλο για την καλύτερη πρόβλεψη αλλά όχι τόσο σημαντικό έτσι ώστε να επιφέρουν μεγάλη βελτίωση στο αποτέλεσμα. Άρα κρατώντας τα μισά (και πιο σημαντικά) attributes μπορούμε να έχουμε παρόμοια αποτελέσματα στις προβλέψεις μας.

Εδώ αξίζει να σημειωθεί πως εφαρμόζοντας το t-test και με τους άλλους δύο classifiers, παρατήρησα ότι με το knn ήταν τα αποτελέσματα ήταν τα καλύτερα που έχω πάρει μέχρι στιγμής, με τις εξής μετρήσεις:

*KNN:*

*Best n\_neighbors: 6*

*Mean accuracy: 0.79*

*Mean geometric mean: 0.52*

*Mean sensitivity: 0.84*

*Mean specificity: 0.34*

Αυτό συμβαίνει επειδή ο αλγόριθμος λαμβάνει υπ’όψη του μόνο τις αποστάσεις των σημείων που επηρρεάζουν θετικά το αποτέλεσμα των προβλέψεων.