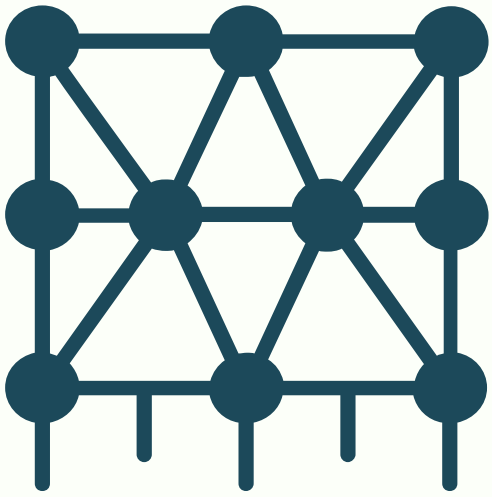
24 ΔΕΚ 2022

Αναφορά Project

Pattern Recognition Laboratory

Karampidis Konstantinos



Ψαλτάκης Γιώργος TH20027

STEP 1 – Data Selection

Απο τα δύο dataset που μας δίνονται επιλέγω το dataset με τα δεδομένα του καρκίκου του μαστού. "breast-cancer-wisconsin.data"

Διαβάζοντας τα δεδομένα που μας δίνει οσο αφορά το dataset μπορούμε να βρούμε τα χαρακτηριστικά της κάθε στήλης και να αναγνωρίσουμε τι τύπου δεδομένα θα περιμένουμε για το dataset μας

#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(2 for benign, 4 for malignant)

Οπότε ολοκληρώνουμε την εισαγωγή του dataset μας στον κωδικά με την πρωσθήκη στηλών τις ονομασίες απο πάνω.

Step 2 – Data inspection

Διαβάζοντας πάλι τα χαρακτηριστικά του dataset μας βλέπουμε οτι μας λέει οτι υπάρχουν δεκαεξι τιμές ποτ έχουν τουλάχιστον κάποια τιμή που να είναι μή υπαρκτή. Η τιμή αυτή χαρακτηρίζεται με το αγγλικό ερωτηματικό. Οπότε κάνουμε έλενχο σε όλο τον κώδικα μας για την ευρεση τιμών με ερωτιματικών για να επιβεβαιώσουμε οτι βρίσκουμε 16. Οταν βρίσκουμε τις 16 τιμές τις κάνουμε drop. και επιβαιώνουμε την αλλαγη στο dataset μας. Με βάση την περιγραφή του συνόλου δεδομένων για τον καρκίνο του μαστού του Ουισκόνσιν που παρέχεται παραπάνω, φαίνεται ότι οι ακόλουθες στήλες ενδέχεται να περιέχουν περιττές πληροφορίες σχετικά με τον ταξινομητή:

"Κωδικός αριθμός δείγματος": Αυτή η στήλη φαίνεται να είναι ένα αναγνωριστικό για κάθε δείγμα και δεν περιέχει χρήσιμες πληροφορίες για τον ταξινομητή.

"Μιτώσεις": Αυτή η στήλη περιέχει πληροφορίες σχετικά με τον αριθμό των μιτώσεων (δηλαδή των κυτταρικών διαιρέσεων) που παρατηρήθηκαν σε κάθε δείγμα, αλλά δεν είναι σαφές πώς αυτές οι πληροφορίες θα ήταν χρήσιμες για την ταξινόμηση των δειγμάτων ως καλοήθων ή κακοήθων που είναι ο target μας.

Οπότε κάνουμε drop και τις δύο αυτές στήλες.

STEP 3 – Data Prepossessing

Τα boxplots παρέχουν μια οπτική αναπαράσταση της κατανομής των τιμών για κάθε στήλη στο DataFrame, ενώ τα συνοπτικά στατιστικά στοιχεία θα παρέχουν αριθμητικές πληροφορίες σχετικά με τη μέση τιμή, την τυπική απόκλιση, το ελάχιστο, το μέγιστο και άλλες στατιστικές ιδιότητες των στηλών.

Παρατηρούμε οτι όλες οι στήλες έχουν παρόμοιο ευρος τιμών με μικρές διαφορές αλλα μερικές έχουν διαφορετική διακύμανση που υποδηλώνει οτι είναι πιο σημαντικές για την ταξινόμηση των δειγμάτων ώς καλοήθων η κακοήθων. Ενας καλύτερος τρόπος να καταλάβουμε αν αυτα τα δεδομένα μπορούν να βοηθήσουν στην κατανοήση μας ειναι να σπλιτάρουμε το dataframe στα δύο βασιζόμενοι στο target αν ειναι καλοηθών η κακοήθων. Βλέπουμε οτι στις περιπτώσεις καλοηθών οι τιμές είναι πιο standardized και πολλές φορές μόναδικής τιμής σε σύγκριση με τών κακοηθών που έχουνε μεγαλύτερο range στις τιμές τους. Μπορόυμε να τα δούμε αυτα αναλυτικά στα τρία boxplot στο jupyter.

Τα δεδομένα τώρα τα κανονικοποιούμε με χρήση του minmaxscaler και αφαιρούμε τους outlier με την χρήση του z_score. Κάνουμε σπλιτ τα δεδομένα μας σε target και το κανονικλο μάς σετ βάση της κλάσης. Και τα πλοτάρουμε σε ένα απλό scatterplot.

Step 4 – Supervised learning (Classification)

Χρησιμοποιούμε τρία διαφορετικά ποσοστά διαχώρισης δεδομένων απο 90/10 σε 80/20 και 70/30 κανοντας χρήση της κλασσικής εντολής xtrain xtest ytrain ytest. Στην συνέχεια παρουσιάζουμε σε scatterplot ολα τα xtest ytest xtrain ytrain για όλες τις περιπτώσεις διαχωρισμού δεδομένων.

Επειτα με την χρήση του classification report μπορούμε να βρούμε ποια μέθοδος απο τις τρείς είναι καλύτερη

80% -20%

```
LinearDiscriminantAnalysis
Classification error:  0.058394160583941646
report:
              precision    recall  f1-score   support

           2       0.93      0.98      0.96        87
           4       0.96      0.88      0.92        50

    accuracy                           0.94       137
   macro avg       0.95      0.93      0.94       137
weighted avg       0.94      0.94      0.94       137

QuadraticDiscriminantAnalysis
Classification error:  0.06569343065693434
report:
              precision    recall  f1-score   support

           2       0.98      0.92      0.95        87
           4       0.87      0.96      0.91        50

    accuracy                           0.93       137
   macro avg       0.92      0.94      0.93       137
weighted avg       0.94      0.93      0.93       137
```

70% -30%

```
LinearDiscriminantAnalysis
Classification error:  0.06341463414634141
report:
              precision    recall  f1-score   support

           2       0.93      0.98      0.95       130
           4       0.96      0.87      0.91        75

    accuracy                           0.94       205
   macro avg       0.94      0.92      0.93       205
weighted avg       0.94      0.94      0.94       205

QuadraticDiscriminantAnalysis
Classification error:  0.05853658536585371
report:
              precision    recall  f1-score   support

           2       0.97      0.94      0.95       130
           4       0.90      0.95      0.92        75

    accuracy                           0.94       205
   macro avg       0.93      0.94      0.94       205
weighted avg       0.94      0.94      0.94       205
```

90% -10%

```
LinearDiscriminantAnalysis
Classification error:  0.01449275362318836
report:
              precision    recall  f1-score   support

           2       1.00      0.98      0.99        46
           4       0.96      1.00      0.98        23

    accuracy                           0.99        69
   macro avg       0.98      0.99      0.98        69
weighted avg       0.99      0.99      0.99        69

QuadraticDiscriminantAnalysis
Classification error:  0.04347826086956519
report:
              precision    recall  f1-score   support

           2       1.00      0.93      0.97        46
           4       0.88      1.00      0.94        23

    accuracy                           0.96        69
   macro avg       0.94      0.97      0.95        69
weighted avg       0.96      0.96      0.96        69
```

Με βάση το σφάλμα ταξινόμησης και τις παρεχόμενες μετρικές αξιολόγησης (ακρίβεια, ανάκληση και f1-score), φαίνεται ότι ο ταξινομητής Linear Discriminant Analysis (LDA) αποδίδει καλύτερα για τα διαχωρίσματα 90/10 και 80/20, ενώ ο ταξινομητής Quadratic Discriminant Analysis (QDA) αποδίδει ελαφρώς καλύτερα για το διαχωρισμό 70/30.

Όταν η υπόθεση των ίσων συνδιακυμάνσεων μεταξύ των κλάσεων είναι λογική, η LDA τείνει να αποδίδει καλύτερα, ωστόσο η QDA μπορεί να αποδίδει καλύτερα όταν η υπόθεση δεν είναι ακριβής ή όταν οι κλάσεις δεν είναι καλά διαχωρισμένες.

Αξίζει να σημειωθεί ότι το σφάλμα ταξινόμησης τόσο για την LDA όσο και για την QDA είναι σχετικά χαμηλό και για τις τρεις διαχωριστικές κατηγορίες, γεγονός που υποδηλώνει ότι και οι δύο ταξινομητές είναι σε θέση να προβλέψουν με ακρίβεια τα αποτελέσματα του class για τα δεδομένα δοκιμής.

Step 5 – Unsupervised learning (k-means)

Με την χρήση του elbow και silhouette analysis βρίσκουμε οτι για το dataset μας έχει ιδανικό k που στο συγκεκριμένο dataset με τις συγκεκριμένες απώλειες δεδομένω μή χρήσιμα για τον έλενγχο μας βρίσκουμε οτι το K=4. Κάνουμε scatter plot για 4 clusters του kmeans για την οπτικοποίηση μας.

Με την χρήση του elbow και silhouette analysis βρίσκουμε οτι για το dataset μας έχει ιδανικό k που στο συγκεκριμένο dataset με τις συγκεκριμένες απώλειες δεδομένω μή χρήσιμα για τον έλενγχο μας βρίσκουμε οτι το K=4. Κάνουμε scatter plot για 4 clusters του kmeans για την οπτικοποίηση μας.

Για να εκτιμηθεί η απόδοση του αλγορίθμου k-means, μπορούμε να συγκρίνουμε τις predicted cluster labels με τα original class labels για κάθε δείγμα στα δεδομένα μας. Ένας τρόπος για να το κάνετε αυτό είναι να χρησιμοποιήσετε τη βαθμολογία προσαρμοσμένης αμοιβαίας πληροφορίας (AMI), η οποία μετρά τη συμφωνία μεταξύ των προβλεπόμενων και των πραγματικών cluster labels.

Στην συγκεκριμένη περίτωση του αποτέλεσμα της ami είναι 0.60093 . Γνώριζοντας οτι το ami κυμαίνετε απο 0 εως 1 με όσο ψηλότερη βαθμολογία τόσο μεγαλύτερη βλέπουμε οτι σε γενικές γραμμές το 0.6 πάραγει κάπως καλά αποτελέσματα αλλα το ιδανικό θα ήταν γύρω στα 0.8. Οπότε δέν συγκρίνεται με της περιπτώσεις supervised learning.

Step 6 – Unsupervised learning (Hierarchal Clustering)

Κάνουμε το κλάσσικο AgglomerativeClustering με 4 clusters και για να βρώ πια μέθοδος clustering είναι καλύτερη βρίσκω το inconsictency coefficient για όλες τις μεθόδους σύνδεσης απο single,complete,average αλλα και ward.

Βάση της μεταβλητης αυτής βλέπουμε ξεκάθαρα οτι η απλή σύνδεση αποδίδει καλύτερα αποτέλέσματα και αμέσως μετά η μέση.

Δεν είναι σαφές πώς ένα δενδρόγραμμα θα ήταν χρήσιμο για την ανάλυση αυτού του συνόλου δεδομένων. Τα δενδρογράμματα χρησιμοποιούνται συνήθως για την οπτικοποίηση της ιεραρχικής δομής ενός συνόλου δεδομένων, αλλά το συγκεκριμένο σύνολο δεδομένων δεν φαίνεται να έχει ιεραρχική δομή. Αντ' αυτού, αποτελείται από μια συλλογή παρατηρήσεων δειγμάτων ιστού καρκίνου του μαστού, καθένα από τα οποία χαρακτηρίζεται είτε ως καλοήθης είτε ως κακοήθης

Παρόλα αυτά βλέπουμε αναλυτικά τα δενδρογράμματα που μας βγαίνουν για αυτό το set. βλέπουμε οτι όλα τα δενδρογράμματα παράγουν κάπως καλά δεδομένα όμως παρατηρούμε οτι το δενδρογραμμα απλής single σύνδεσης έχει μικρότερη αποστασή στο κοψιμό y στις τελευταίες κορυφές μας έχουν διακριτό σχήμα και είναι σχετικά συμπαγείς, παρόμοια αποτελέσματα βλέπουμε και στο average που έχει το αμέσως πιο μικρό inconcistency index μετά το single

Step 7 – Naïve Bayes

Κάνουμε την διαδικάσια για το naive bayes τώρα και βρίσκουμε και τα αποτελέσματα βάση ακρίβειας ανάκλησης αλλα και βαθμολογίας f1.

Accuracy: 0.94
Precision: 0.88
Recall: 0.99
F1 score: 0.93

Με βάση των παραπάνων τιμών μπορούμε να συμπεράνουμε οτι το μοντελο του naive bayes κάνει αρκετά ακριβές προβλέψεις και οτι παράγει ελάχιστες σχετικά ψευδεις προβλέψεις και έχει έναν αρμονικό μέσο όρο f1 0.93 που αποδεικνύει οτι κάνει ακριβείς προβλέψεις με υψηλη ακρίβεια αλλα και άνακληση.

Στην συνέχεια κάνουμε το scatterplot για την οπτικοποίηση των αποτελεσμάτων της naive bayes.

Step 8 – Conclusion

Απο οτι βλέπουμε με όλες τις μεθόδους μάθησης που δοκιμάσαμε η απόδοση στις περιπτώσεις supervised learning όπως Classification με τον quadratic classifier αλλα και τον linear classifier αλλα και τον naive bayes βλέπουμε αποτελέσματα τις τάξης πάνω απο 0.85 που παράγουν πάρα πολύ σωστά και ευστοχα αποτελέσματα με μερικές μόνο αστοχίες.

Σε αντίθεση οι μέθοδοι Unsupervised learning παρόλο που παράγουν αποτελέσματα αρκετά σωστά δεν συγκρίνεται με την ακριβεία των supervised learning με αποτελέσματα κλάσης 0.6

Γενικά, οι αλγόριθμοι μάθησης με επίβλεψη μπορούν να είναι αποτελεσματικοί για εργασίες ταξινόμησης όπως αυτή που περιγράφεται σε αυτή τη βάση δεδομένων, όπου ο στόχος είναι να προβλεφθεί εάν μια δεδομένη περίπτωση καρκίνου του μαστού είναι καλοήθης ή κακοήθης με βάση τα χαρακτηριστικά της περίπτωσης. Οι αλγόριθμοι μάθησης χωρίς επίβλεψη, από την άλλη πλευρά, δεν χρησιμοποιούνται συνήθως για εργασίες ταξινόμησης και χρησιμοποιούνται συχνότερα για εργασίες όπως η ομαδοποίηση, όπου ο στόχος είναι ο εντοπισμός ομάδων ή συστάδων παρόμοιων σημείων δεδομένων χωρίς προηγούμενη γνώση των ετικετών τους.

Οπότε συμπαιραίνουμε γενικά οτι για το συγκεκριμένο dataset έχουμε καλύτερα αποτελέσματα με supervised learning όπως και περιμέναμε.