Comparing Classifiers on Breast Cancer Data

Ellie Pizey Imperial College London Mathematics

Introduction

The data I have chosen is the properties of breast cancer cells[1]. The data has 699 instances and 10 attributes which include radius, area, smoothness, symmetry and concave points. I'm using these to determine whether the tumor is malignant or benign. That means that the data would ideally be interpretable so a doctor could give an explanation to their patient.

KNN

K nearest neighbours is a method that works by finding which class (malignant or benign) the majority of the k nearest points belong to.

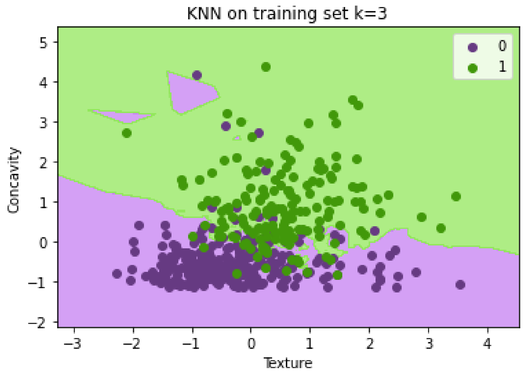


Figure 1: k=3

I started with k=3 to get a rough idea of the underlying patterns in the data. As you can see in Figure 1 there is some evidence of overfitting (which is when the classifier learns more about the training data than the underlying process) in the shapes that correspond to benign (purple) outliers among the malignant (green) points. The accuracy rate on the training data is 0.90 but is 0.77 on the test data. Performing significantly better on training data is a strong indicator of overfitting.

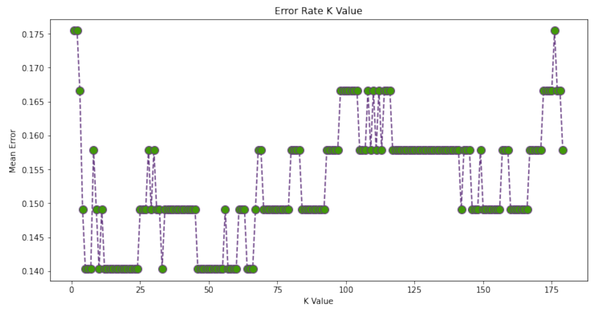


Figure 2: Error rates for different k.

I calculated the average error rate in the test data for different values of k as is shown in Figure 2. This has minima around 20 and 50.

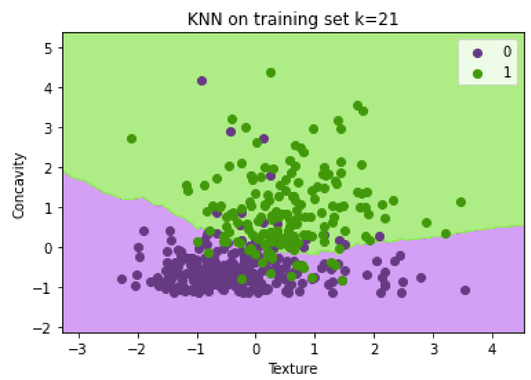


Figure 3: k=21

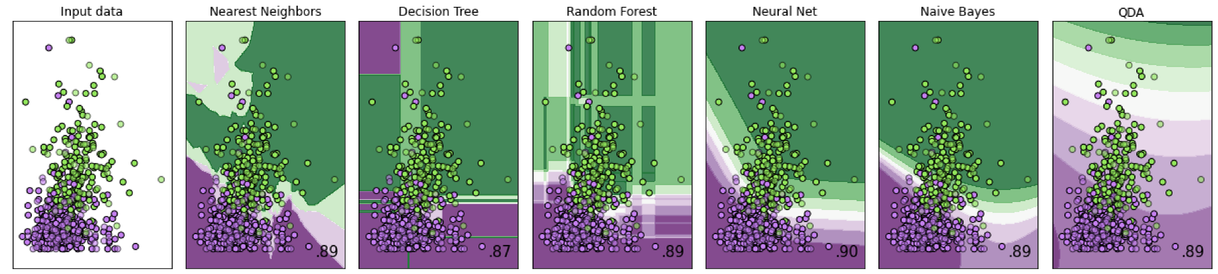
From this I chose the value k=21 as it showed little overfitting in the graph and had the lowest error rate for the smallest k.

References

[1] Wisconsin breast cancer dataset from Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

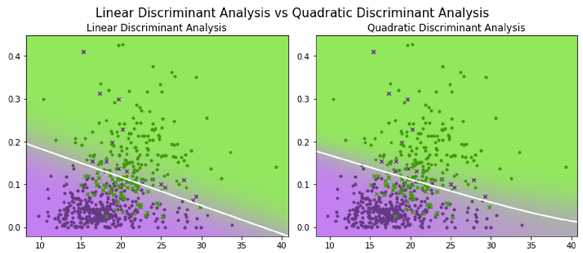
[2]Christoph Molnar Interpretable Machine Learning 2021 All the websites I used and other sources/acknowledgements are in my github repository as well as my python code, R code and extra graphs that didn't fit this poster. It is located at : https://github.com/moticilla/classificationproblem.

Other Classifiers



LDA and QDA

Linear Discriminant Analysis and Quadratic Discriminant Analysis are different classifiers that fit a line to the data (straight line for LDA).



We can see in these graphs that the line for QDA is only slightly curved so actually the LDA line will give a similar classification of the data and may be a good fit.

Evaluation

Classifier	Train error	Test error
KNN k=3	0.101	0.235
KNN k=21	0.251	0.281
LDA	0.134	0.123
QDA	0.121	0.105

Table 1: Average of error for Test and Train.

From this table we can see that KNN for k=3 overfits the most but there is only slight overfitting for k=21 and none for LDA and QDA. The reason for the higher error rate for k=21 is that I haven't used the weighted averages of the error and my malignant:benign split is 241:458.

Conclusion

In the Other Classifiers graphs I have used the built in classification functions in python and plotted them for my data. I have also calculated the accuracy on the test data which is in the corner of each graph. From these test errors, we can see that the neural network (Python's Multilayer Perceptron classifier) has the highest accuracy with 0.90. However, Nearest neighbours (k=3), Naive Bayes and QDA all have accuracy 0.89, so we will need to find another way to make a decision.

KNN is interpretable on a local level for this data as we are only using two features so you can retrieve the k neighbours that were used to make the prediction and this method can be easily explained to a patient. LDA models a linear relationship which in this case seems to be a reasonable assumption as it has a low error rate and the QDA plot is only slightly curved. This means the resulting model is easily explained but the method behind it is not.[2]

The classifier I would choose is QDA as it doesn't show signs of overfitting, has the lowest error rate and although the theory behind the classification isn't easily explained to a patient, the features we are using are easily understood and so it can be interpreted.

If I had more time I would investigate feature selection. My data has 10 features and I plotted pairs of features against each other to find any relationships, but I haven't done any statistical investigation into this, such as Recursive Feature Elimination.