# Machine Learning Assignment 1

Iain Souttar

February 12, 2018

## 1 Introduction

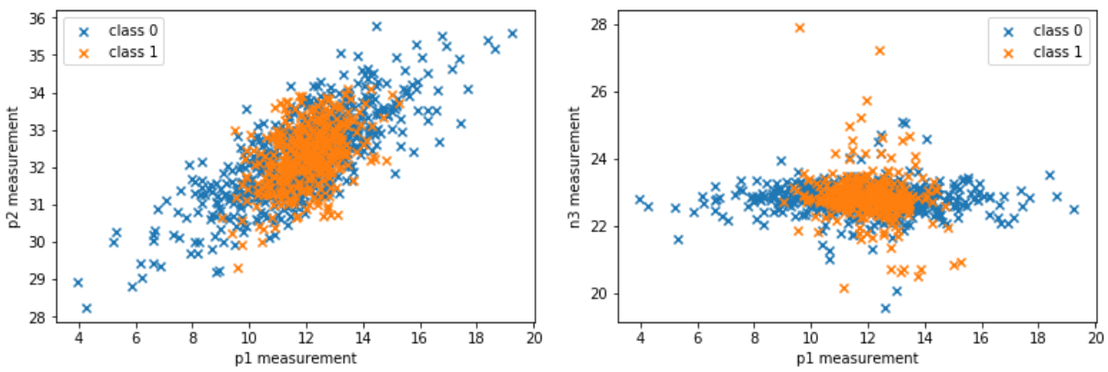
To begin, it is important to explain and justify the preprocessing of the data that I have utilised throughout this assignment. When initially plotting the distributions of each feature, I noticed that there were some outliers many standard deviations away from the mean. Upon reviewing the data, it was clear that some entries didn't belong. It appeared that some had been swapped. I attributed this to a fault in the machine so decided to write a function to swap them back, ensuring that each data entry was in the corrrect column.

In terms of scaling, I used the tried and tested mean, standard deviation method- fitting each feature to a more natural normal distribution. Throughout my testing phase, I also experimented with normalising the rows to ensure the entries lie on a hypershpere, but it proved to have no positive difference on the score of each model. Another thing that caught my eye was that the measurements in probe B seem to be somewhat shifted. Certainly this machine doesn't seem reliable so it seems necessary to scale probe A and B throughout the assignment. This will prevent any conflict of magnitude between the two.

## 2 Importance of features

### 2.1 Class

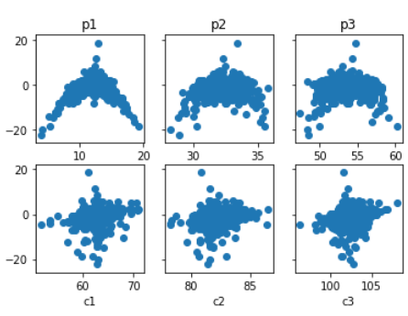
It was not immediately clear which measurements were most useful to look at when wanting to determine class. When plotted against class, each feature is very mixed- the class is not determined in any meaningful way by the value of any one feature. A more fruitful venture was when I looked at all combinations of two measurements plotted against eachother. Some were fairly informative. I used a scatter graph for this, with different colour points indicating the class. Seen below, the data for class 1 is less spread out. This would mean, for example, that if we had a $p1$ measurement of 18 and a $p2$ measurement of 35, we could make a reasonable guess for that being class 0. However, we still have little to no information on the class of data towards the middle- this is why any decent model would have to use more than just two values. The scatter of $p1$ and $n3$ is perhaps even more revealing, but with the same problem in the middle of our data. Using the Decision tree classifier, I also was able to calculate the '$feature\ importances$', as a quantitative measurement of how important each feature is. This rated $n3$ and $p1$ as the most important, with several offering negligible information in terms of class.




### 2.2 TNA

I began by plotting each feature against TNA. Most of the plots produced seemed to display no real sign of correlation. However (as shown below $right$) both the $p$ and $c$ measurements indicated some sign of importance. The most striking is $p1$, where a clear curve can be identified. This indicated to me a potentially quadratic relation- something which I used when deciding upon my feature expansion and model. The correlation in $c$ plots are less noticeable, but there is still some real relationship there. I reinforced this by producing a table of correlation (seen below $left$) coefficients. Interestingly the $p$ values were lower, but this is because the correlation is a linear measure. The $c$ values were the highest of all features, but still relatively low around 0.25.

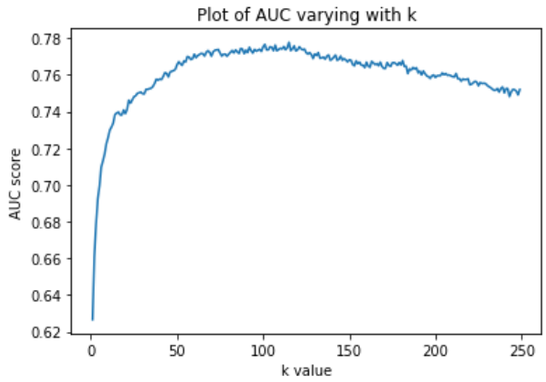
| | c1 | c2 | c3 | p1 | p2 | p3 | tna |
|---|---|---|---|---|---|---|---|
| **c1** | 1.000000 | 0.688252 | 0.670122 | 0.000527 | -0.016717 | 0.003907 | 0.223412 |
| **c2** | 0.688252 | 1.000000 | 0.491119 | -0.020999 | -0.007824 | 0.006509 | 0.157044 |
| **c3** | 0.670122 | 0.491119 | 1.000000 | -0.012925 | -0.029027 | -0.010548 | 0.309319 |
| **p1** | 0.000527 | -0.020999 | -0.012925 | 1.000000 | 0.703235 | 0.720972 | 0.112145 |
| **p2** | -0.016717 | -0.007824 | -0.029027 | 0.703235 | 1.000000 | 0.500621 | 0.092050 |
| **p3** | 0.003907 | 0.006509 | -0.010548 | 0.720972 | 0.500621 | 1.000000 | 0.042546 |
| **tna** | 0.223412 | 0.157044 | 0.309319 | 0.112145 | 0.092050 | 0.042546 | 1.000000 |

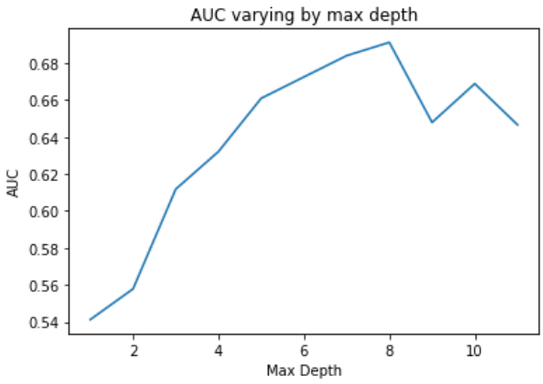


# 3 Classifier

The two options here were a decision tree model or K-nearest neighbour. I tested a decision tree first. I was able to tinker with the criterion (gini or entropy), the max depth of the tree, minimum impurity decrease to create a new node and many other things. Throughout classification, I used stratified 10 fold cross validation while testing, meaning the proportion of classes was preserved in each fold. This ensured I obtained a truer, more representative AUC. Below can be seen how the decision tree fared when changing the max depth.

The other option was knn. There was less to tinker with here- only the value of k and whether the distance was weighted when taking into account the 'votes' of neighbours. I also tested something fairly novel- before passing my dataframe into the knn model. It had occurred to me that this model would take into account all features the same- due to the nature of the $L^p$ norm. I multiplied each column with the feature importances found in the decision tree. This has the same affect as using a weighted metric- differences in the important features are what count the most since they are now much larger so the nearest neighbours are ones with similar values of these features. I was wary of potential for overfitting so calculated the feature importances ten different times and took the mean. I also, as well as the stratified K-fold mentioned previously, used a repeated (50 times) 10-fold to double check that it didn't only produce good results under certain folds. Below the graph of how the AUC varied according to the K value can be seen. The best k-nearest value was 114. This model produced the highest score out of all classifier methods tested and doesn't seem, to me, to be overfitting so was the one I chose.




# 4 Regressor

There were quite a few options for regression models. OLS, Ridge, Lasso and regression types of Decision trees/KNN were the ones I considered. I used repeated 10-fold cross validation as explained before for the regression part of this assignment. Looking back on my data exploration, I already had an intuition for needing more than just one degree of features. I identified a possible quadratic relation between the $p$'s and tna. This led me to test on the power set of both degree 2 and 3 polynomials of the features. With this in mind, I decided not to use OLS since it has no real ability to set useless features to 0 weight. This would mean I am likely to overfit the data using it.

The regression models for usual-classifiers KNN and Decision trees were the first ones I tested. They were okay, producing a $r^2$ score of around 0.6 after some tinkering, but again could not create a model of the right simplicity while retaining the vital pieces of information. Seen below $left$ is the graph of the best k value in the KNN regressor. Clearly, it is not a fantastic model considering the $r^2$, but also produces a low k-value. This tells us that in the data, the tna value can be determined by the very closest neighbours.

Lasso was the best performing in my tests. The model depends on a given alpha which varies how much it tries to reign in feature coefficients. This means that a high alpha will result in a much simpler model- sacrificing on accuracy but preventing overfitting. This was particularly useful in my situation because I, as mentioned before, wanted to capture potential quadratic/cubic relationships but with over 10 features to start with overfitting would be very likely. The alpha I picked that prevented overfitting while not dropping too much in $r^2$ was 0.001. This, with feature expansion of degree 2, made my lasso model. Degree 3 did give a couple of percent better scores, but I believe it was overfitting the data due to the sheer number of features in the expansion. Given more time I would have utilised some form of feature selection, reducing the amount of noise in the data.

Below $right$ is another indicator of why the swap was needed. The distributions of the chemical measurement clearly show that some pieces of data have been misplaced. Who knows why it happened, but it did.