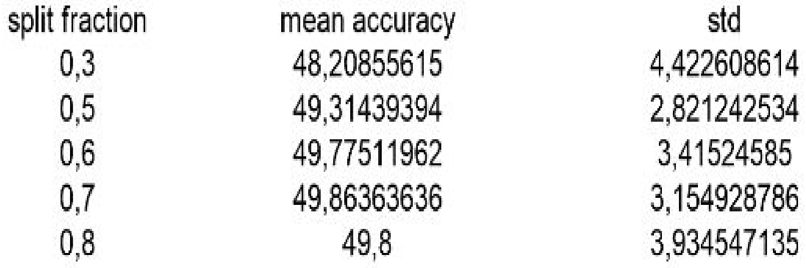
Naive bayes classification

The justification of the naive assumption

The naive bayes assumption that every observed features are independent of each other given the target class is justified if the underlying feature of the data are truly independent. For example there are some correlation of age | length | vocabulary capacity, age and vocabulary capacity is not independent if we analyse them together but there truly independent if the target class is fixed. Given a fixed target class ‘length’ then P(age|length)P(voc|length).

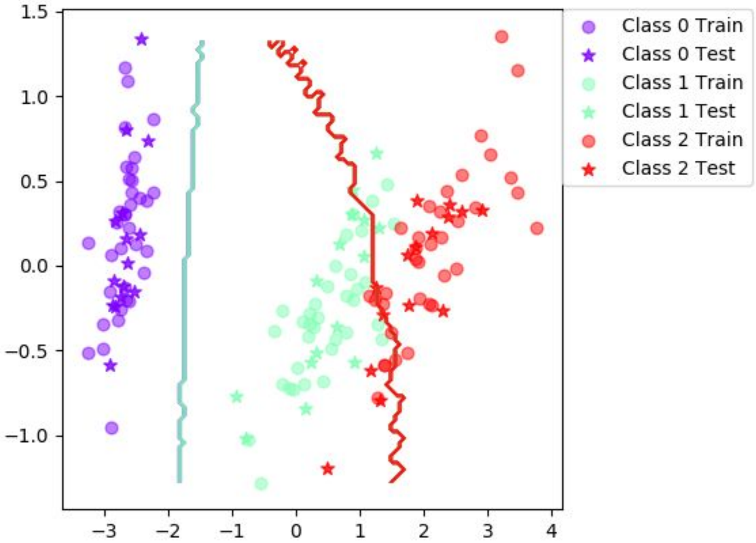
As said before the NBC is a very good and robust if the observed features are probabilistic independent of each other. In the cases when that independence is not hold then we can have a high rate of misclassification. That can occur in text classification tasks when the probability to find a word is not independent of find the other. for example in New York, so is New and York dependent of each other in this context. In those cases the NBC may performs poorly as observed in the Classification of vowels dataset.



split fraction	mean accuracy	std
0,3	48,20855615	4,422608614
0,5	49,31439394	2,821242534
0,6	49,77511962	3,41524585
0,7	49,86363636	3,154928786
0,8	49,8	3,934547135

VOWEL DATA

Improvement regarding the decision boundary



We can see that the decision boundary between the class 1 and class 2 is very difficult to determine. It seems there some kind of overfitting and the presence of outliers. We could change the classifier and use for example *the support vector machine* ***(SVM)*** *with non-linear kernel to transform the data and make more separable*. We can have ***more data*** to find out exactly if the outliers are single errors or if they really represent a pattern in the dataset. We can also try to transform the data by *the* ***PCA*** *method* and find out if there are some dimensionality there we can have a more separable case and higher predictive accuracy.