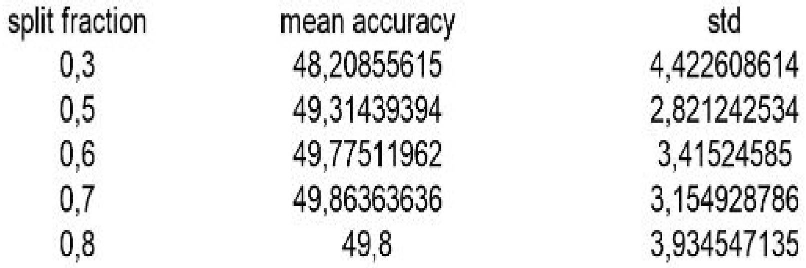
Naive bayes classification

The justification of the naive assumption

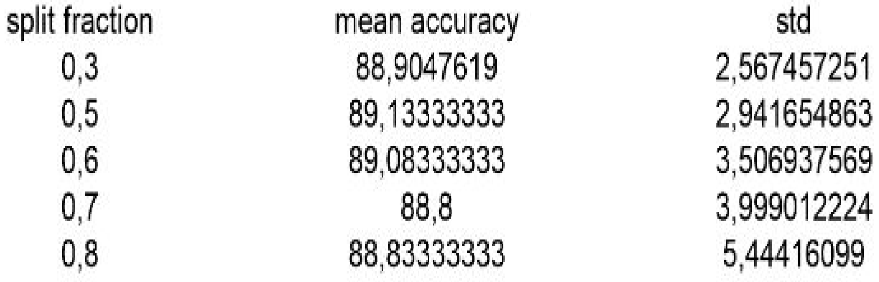
The naive bayes assumption assumes that every observed features are independent of each other given the target class. This assumption is justified if the underlying feature of the data are truly independent. For example there are some correlation between age | length | vocabulary capacity, age and vocabulary 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) and P(voc|length) are independent.

As described before the NBC is a very good and robust classifier if the observed features are probabilistic independent of each other. In the cases when that independence is not hold then a high rate of misclassification can be observed. That can occur in text classification tasks when the probability to find one word is not independent of finding another word in the text. for example in New York, so is New and York dependent of each other in that 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

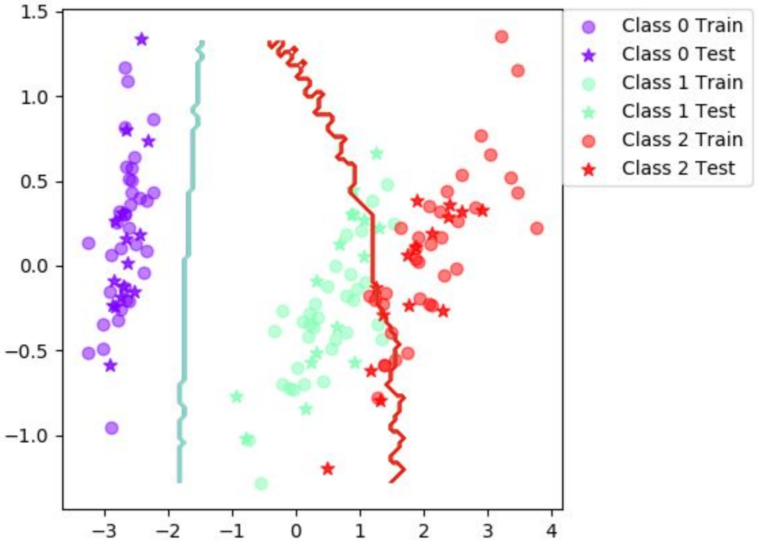
VOWEL DATA



split fraction	mean accuracy	std
0,3	88,9047619	2,567457251
0,5	89,13333333	2,941654863
0,6	89,08333333	3,506937569
0,7	88,8	3,999012224
0,8	88,83333333	5,44416099

IRIS DATA

<u>Improvement regarding the decision boundary</u>



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