Lab 3

Summary

Assignment 3 Questions

Q: When can a feature independence assumption be reasonable and when not?

In general, it is **not realistic** to believe that features are independent.However, it may be a useful assumption in order to reduce complexity.

It is *not reasonable* to make such an assumption when there is:

- **high dependence between features** (i.e. features "temperature: hot/cold" and "outlook: sunny/cloudy") OR
- When there is a **high risk involved in a bad model** (i.e. testing a drug to cure an illness).

Naive Bayes Classifier

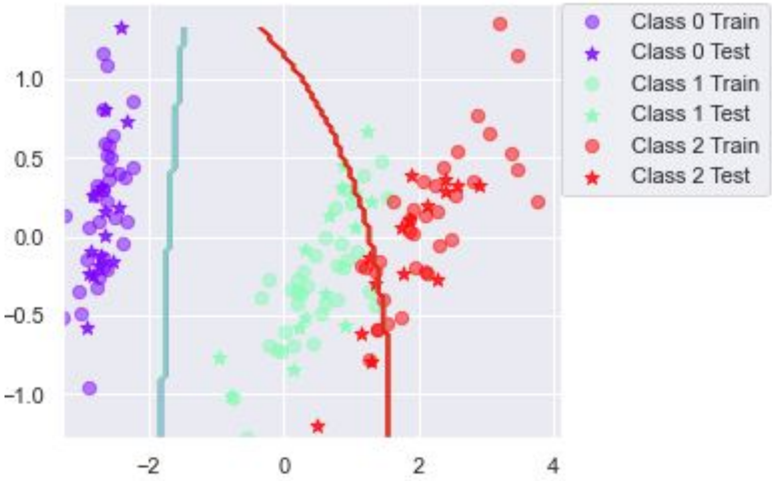
Iris Dataset

Q: How could one improve the classification results for this scenario by changing classier or, alternatively, manipulating the data?

The classes 0 and 1 are successfully divided (green curve)

The classes 1 and 2 seem to not properly be divided (red curve).

Alternative classifier: Support vector machine SVM (with slack).



Naive Bayes Classifier
Iris Dataset

Without boosting:

Mean accuracy = 89

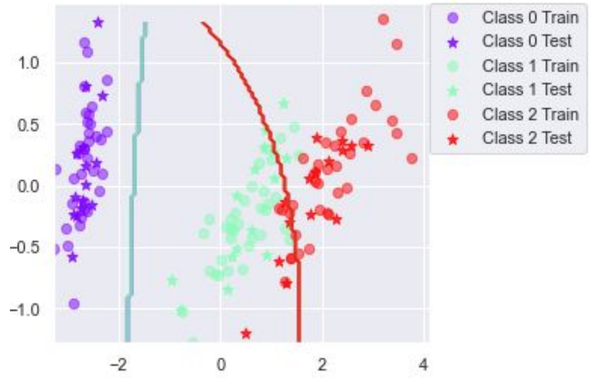
With boosting:

Mean accuracy = 94.1

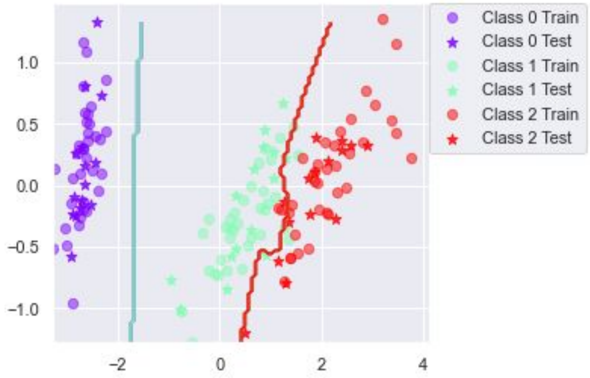
Boosting **reduces bias** by focusing on minimizing the error of poor predictions and trying to model them better in the next iteration.

The algorithm also **lowers the variance** by taking a weighted average of many weak models (but not too many - to not overfit)

Without Boosting



With Boosting



Decision Tree Classifier
Iris Dataset

Without boosting:
Mean accuracy = 92.4

With boosting:
Mean accuracy = 94.6



Without Boosting

With Boosting

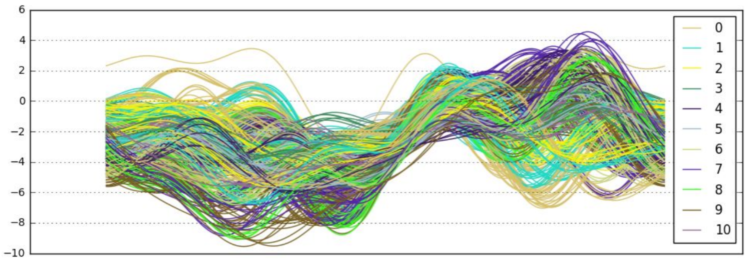
Naive Bayes Classifier
Vowel Dataset

Without boosting:

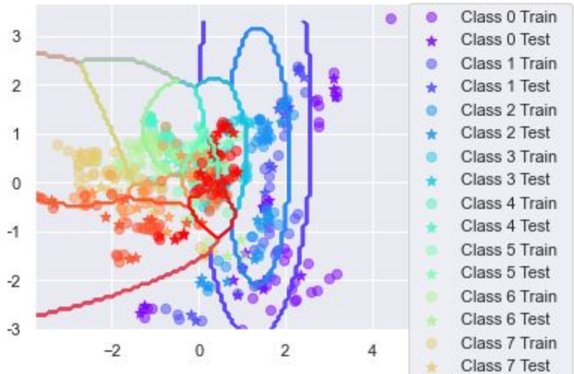
Mean accuracy = 64.7

With boosting:

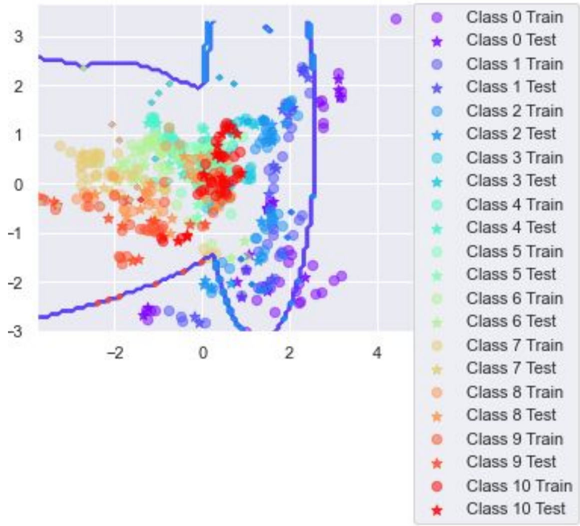
Mean accuracy = 80.2



Without Boosting



With Boosting



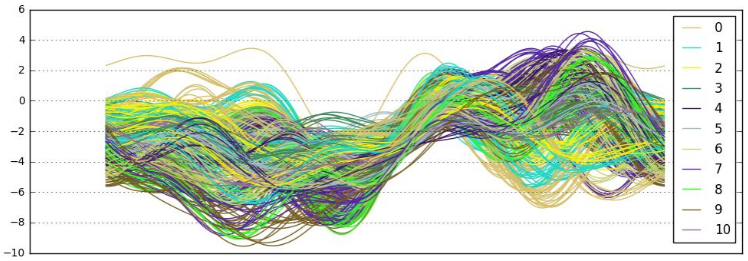
Decision Tree Classifier
Vowel Dataset

Without boosting:

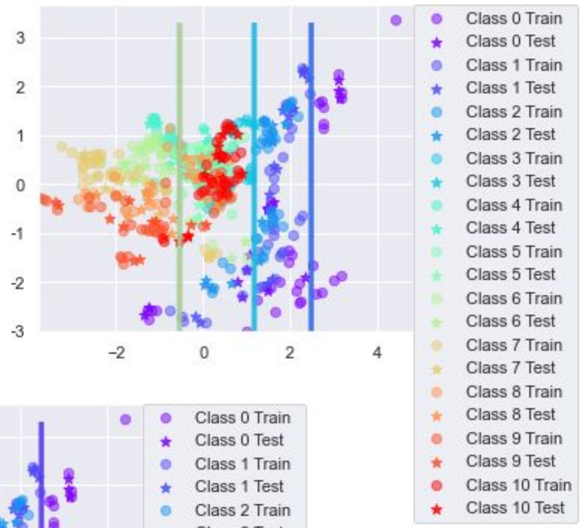
Mean accuracy = 64.1

With boosting:

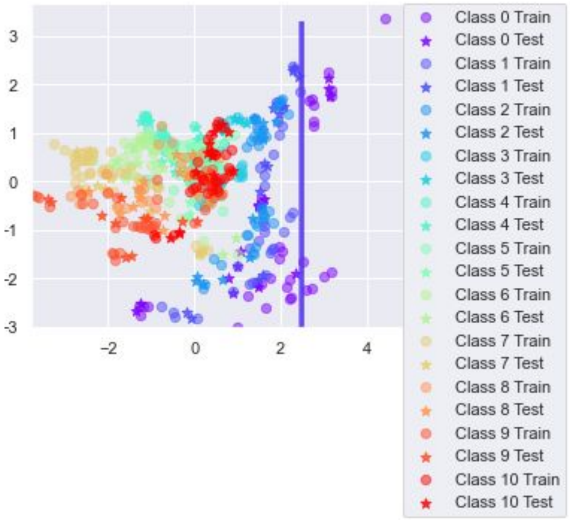
Mean accuracy = 86.6



Without Boosting



With Boosting



Pick a Classifier

Context	Winner!	
Outliers	Decision Tree	And if the data set has a lot of outliers we choose Decision Trees because they are more robust to outliers compared to Naive Bayes.
Irrelevant Inputs	Decision Tree	We choose Decision Trees because the irrelevant data can be pruned from the data set or the irrelevant features will be placed lower in the tree, affecting the classification less. In Naive Bayes each event contributes equally to classifying the outcome, which is too sensitive to irrelevant inputs.
Predictive Power	Naive Bayes with Boosting	We choose Naive Bayes with Boosting because in terms of predictive power, which had better accuracy for the Iris Data Set
Mixed Types of Data	Either Model	Both models can handle mixed types of data, however Naive Bayes will perform better on continuous data forms.
Scalability	Decision Tree	Decision Trees is able to handle large data sets along with colinearity of features. On the other hand, Naive Bayes is not affected by the *curse of dimensionality* and therefore it can scale well to large data sets, however we must assume independence among features.