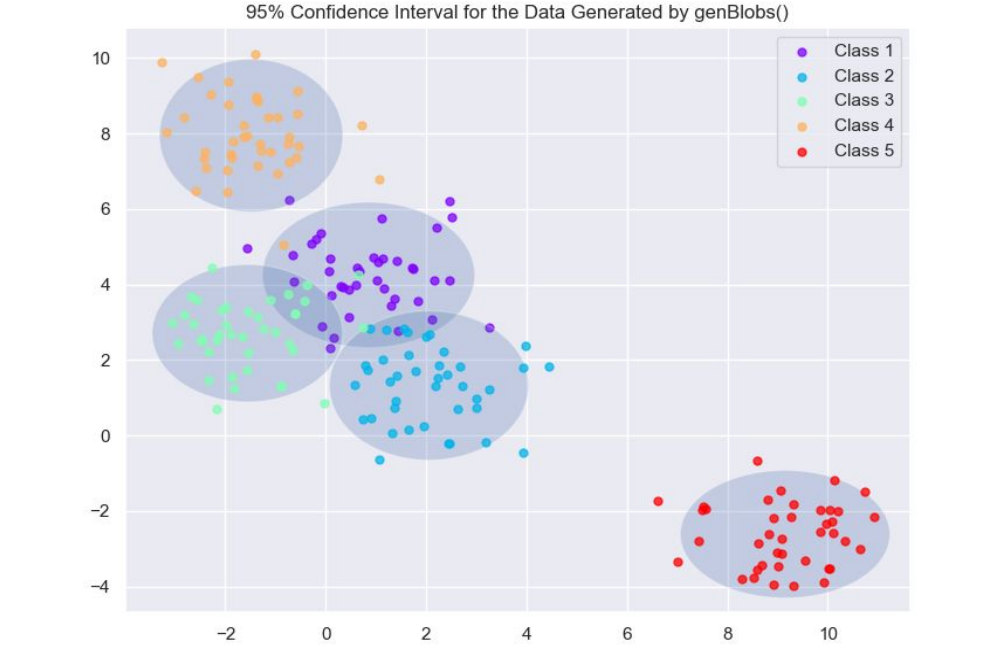
MACHINE LEARNING LAB 3

BAYESIAN LEARNING AND BOOSTING

David O'Leary & Cillian Smith

ASSIGNMENT 1

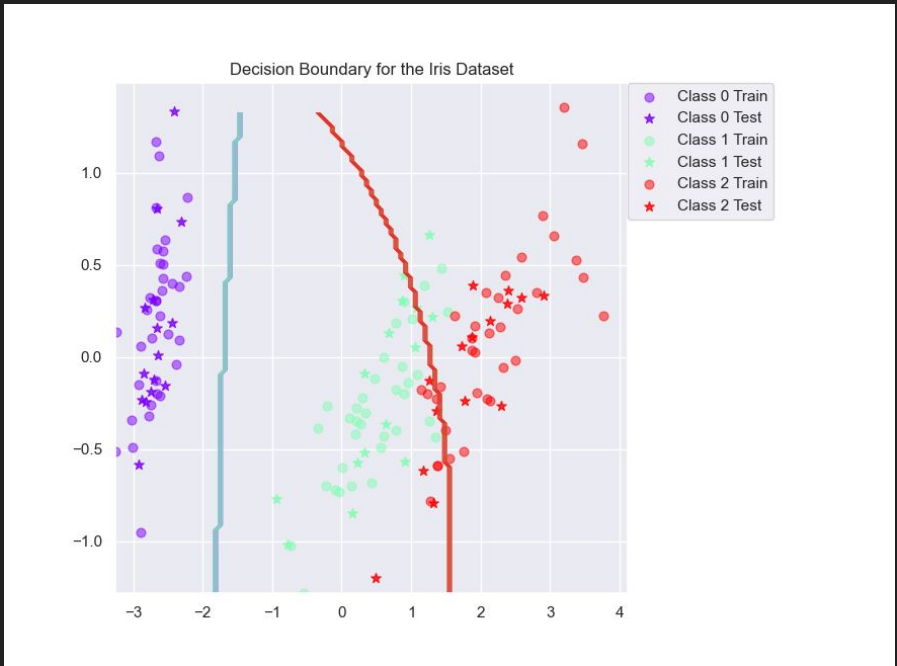


ASSIGNMENT 2

We implemented the computePrior function to estimate and return the class prior in X.

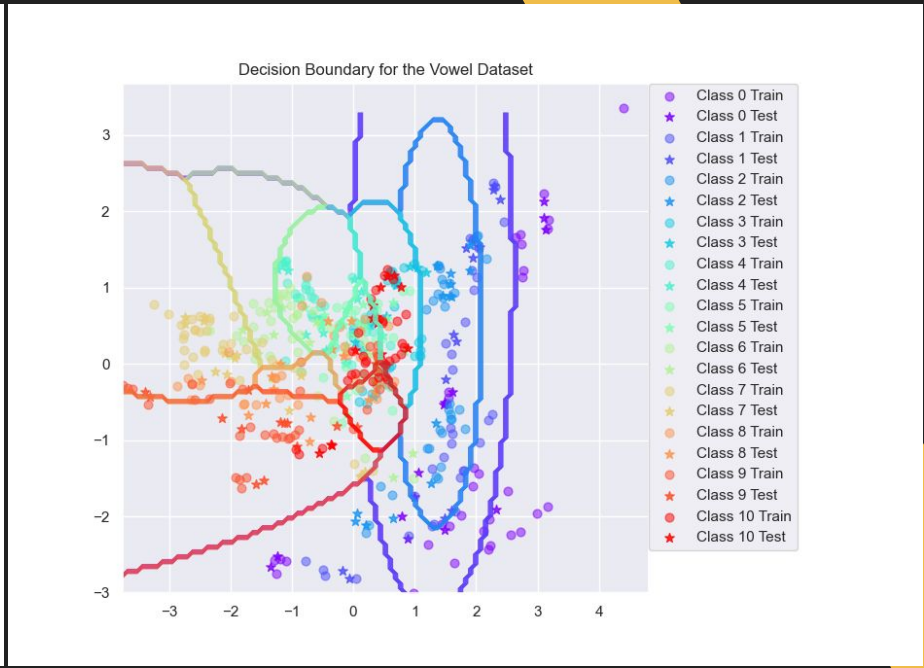
We implemented a function called classifyBayes that computes the discriminant function values for all classes and data points, and classifies each point to belong to the max discriminant value.

ASSIGNMENT 3



Mean classification accuracy: 89

Standard deviation: 4.11



Mean classification accuracy: 64.8

Standard deviation: 4.02

ASSIGNMENT 3

1. **When can a feature independence assumption be reasonable and when not?**

   The independence assumption is reasonable if the features being used are relatively simple and not strongly correlated.

   In real world scenarios, the independence assumption is often not reasonable. For example, reading from weather sensor data may be strongly correlated.

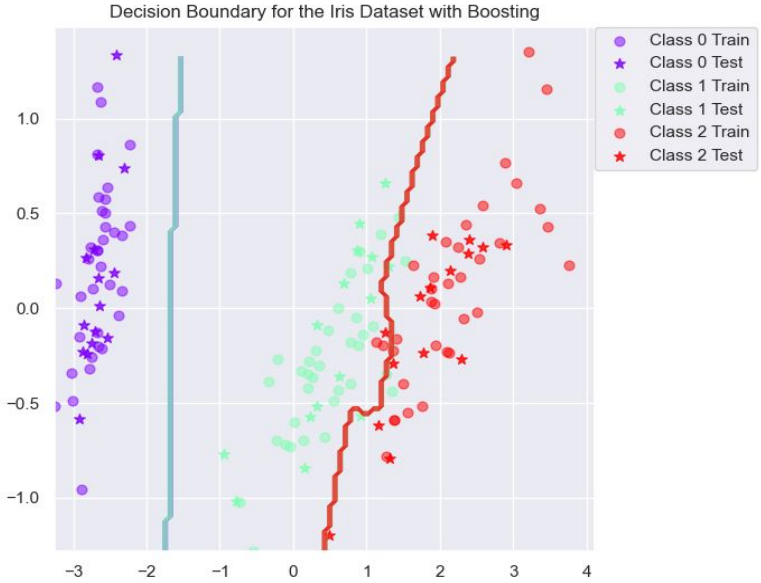
2. **How could one improve the classification results for this scenario by changing classifier or alternatively manipulating the data?**

   Dropping the independence assumption would improve accuracy - by using a Bayesian approach instead of Naive Bayes.

ASSIGNMENTS 4 & 5

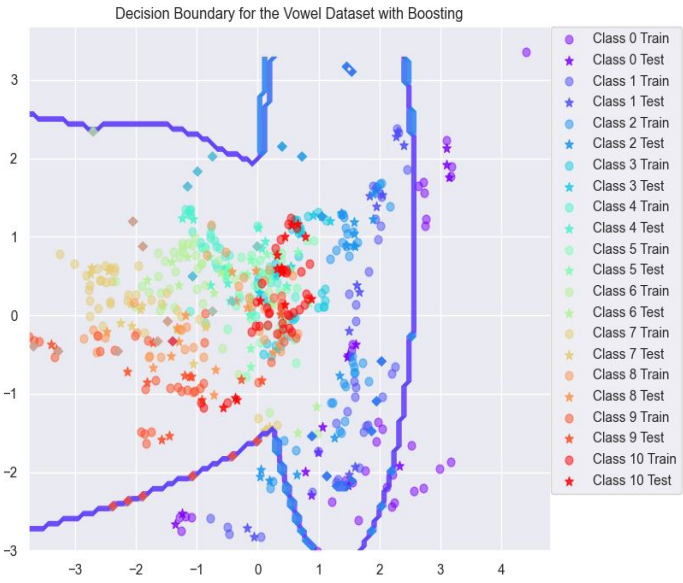
- We implemented the weight vector for ***mlParams*** and the MAP parameters were the same as those obtained with the previous version.
- We modified ***computePrior*** to take boosting weights into account
- We implemented the Adaboost algorithm and the ***classifyBoost*** function to classify instances in data by means of aggregated boosted classifier
- The accuracy of the Iris dataset improved **from 89 to 94.1** with boosting.
- The accuracy of the Vowel dataset improved **from 64.8 to 80.2 with** boosting.

ASSIGNMENT 5



Mean classification accuracy: 94.1

Standard deviation: 6.72



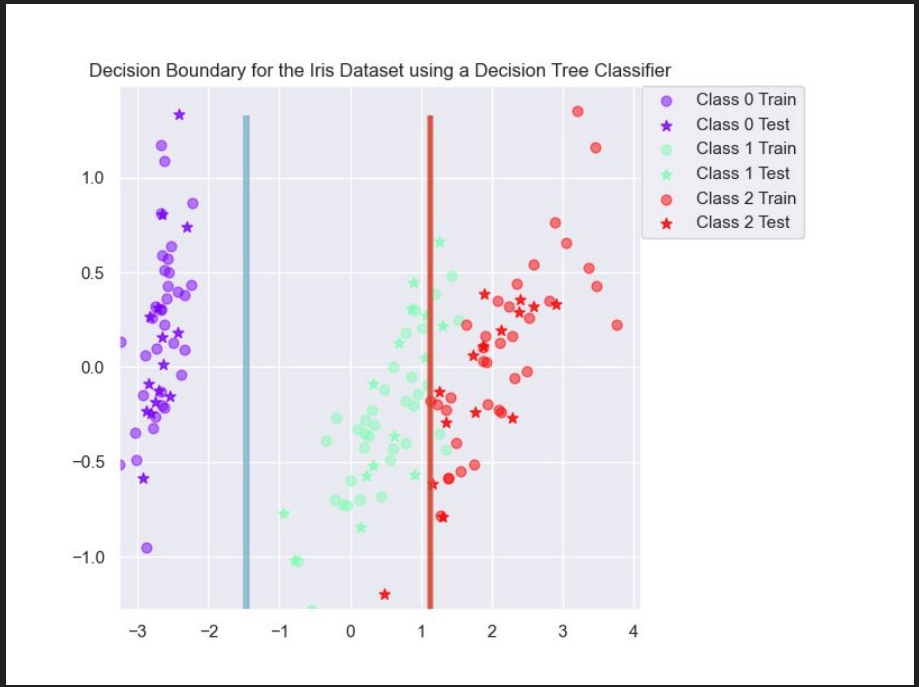
Mean classification accuracy: 80.2

Standard deviation: 3.52

ASSIGNMENT 5

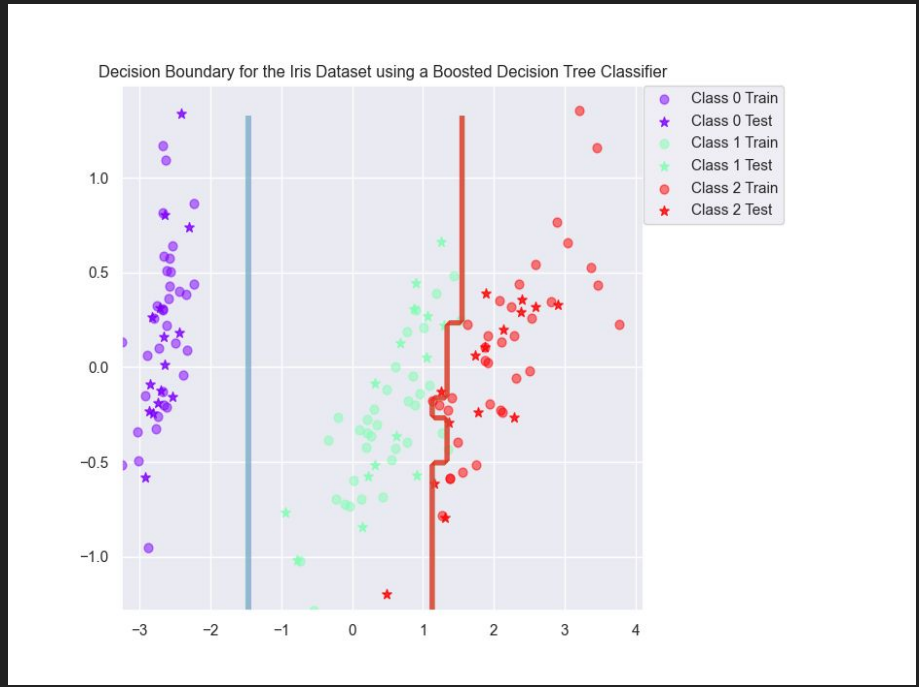
- Differences between boosted Iris decision boundary and original Iris decision boundary:
- The decision boundary of the boosted classifier is more complex and doesn’t follow as smooth of a curve as the original.
- However, it is much more accurate due to the Adaboost algorithm taking the weights of different points into account.
- We can improve a basic model by using boosting, however boosting alone is not always enough to fix a model that is too simple.
- If the model is too simple, it might be best to use a more advanced one such as a random forest or decision tree.

ASSIGNMENT 6



Mean classification accuracy: 92.4

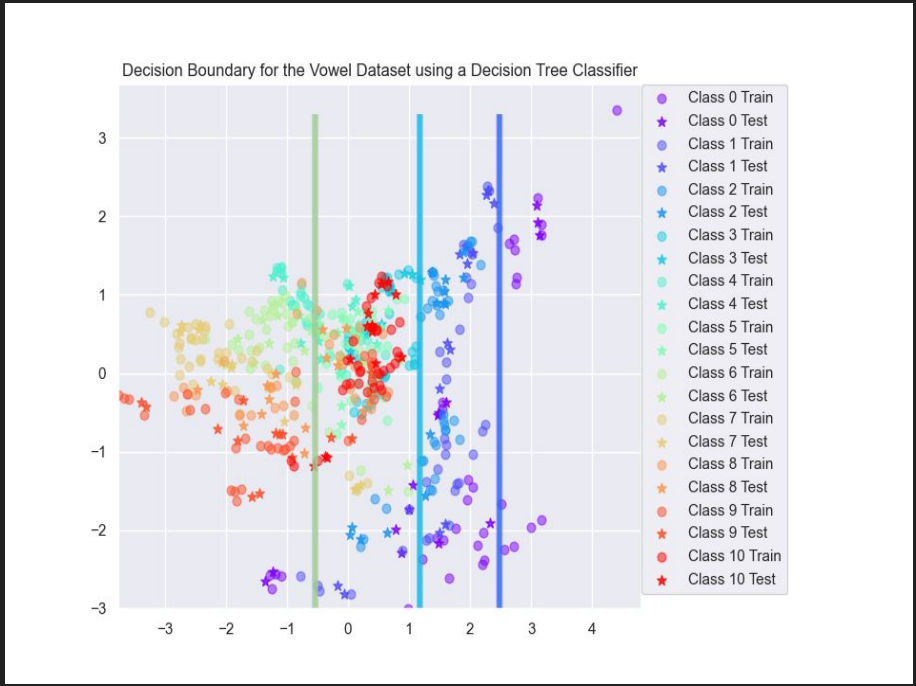
Standard deviation: 3.71



Mean classification accuracy: 94.6

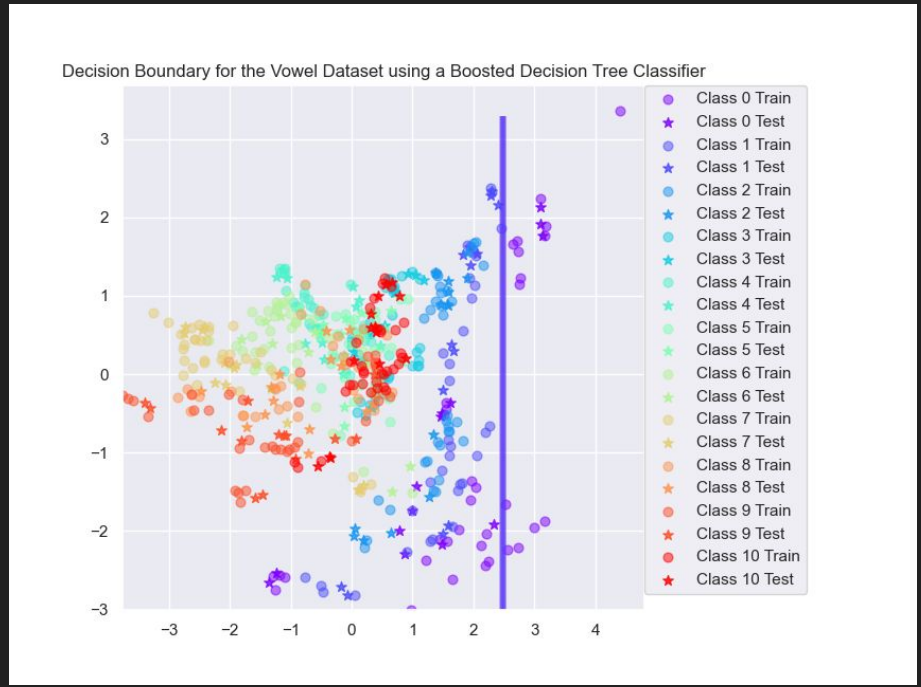
Standard deviation: 3.67

ASSIGNMENT 6



Mean classification accuracy: 64.1

Standard deviation: 4



Mean classification accuracy 86.5

Standard deviation 2.96

ASSIGNMENT 6

- The accuracy for the Iris dataset improved from **92.4** to **94.6** when boosting was used with the decision trees.
- The accuracy for the Vowel dataset improved from **64.1** to **86.5** when boosting was used with the decision trees.
- For the iris dataset, the decision boundary is more complex after boosting.
- For the vowel dataset, some of the decision boundary lines disappear possibly due to overfitting making the boundaries too complex.
- Boosting improves the performance of weak models, but if the problem requires a more complex decision tree, boosting may result in overfitting or an unsatisfactory classification.

ASSIGNMENT 7

Question: If you had to pick a classifier, Naive Bayes or a decision tree or the boosted versions of these, which one would you pick? Motivate from the following criteria:

1) Outliers 2) Irrelevant inputs 3) Predictive power
4) Mixed types of data 5) Scalability

1. **Outliers**: If the dataset contains outliers, a **decision tree** or a **boosted decision tree** might be better than Naive Bayes since decision trees are less sensitive to outliers due to partitioning the feature space into smaller regions.

2. **Irrelevant Inputs**: A **decision tree** or **boosted decision tree** can remove irrelevant features during the training process. Since Naive Bayes assumes all features are independent, it mightn’t be able to accurately identify and remove irrelevant features.

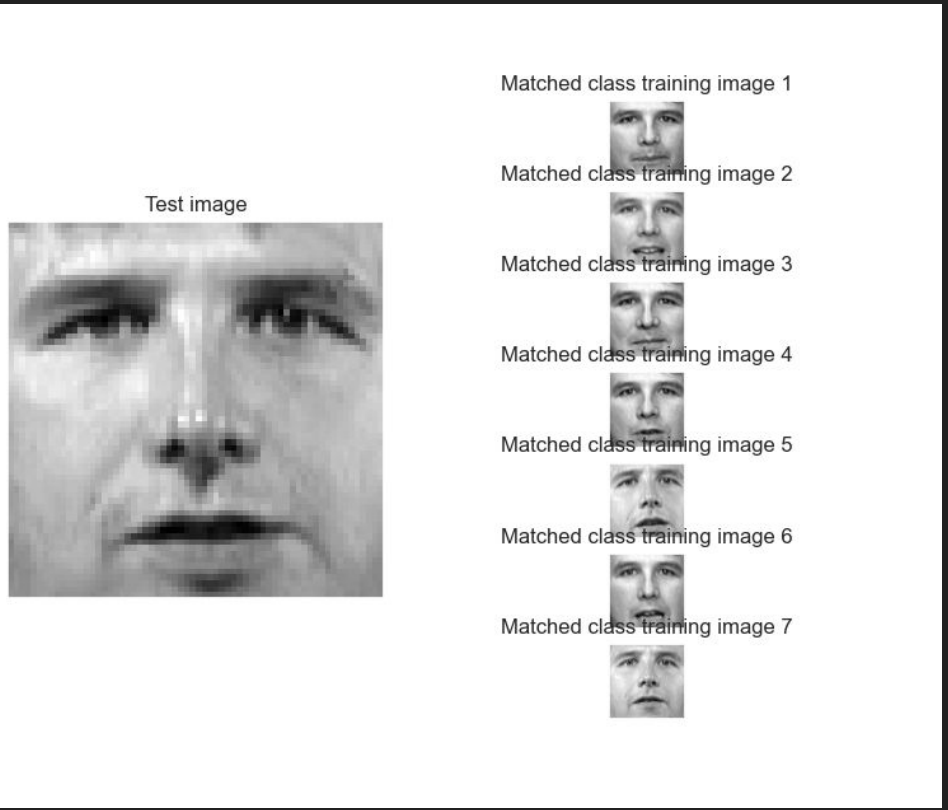
ASSIGNMENT 7

3. **Predictive Power**: **Decision trees** and **boosted decision trees** can still have high predictive power when the dataset is very complex, however Naive Bayes might not be able to capture complex relationships accurately.

4. **Mixed Types of Data**: **Naive Bayes** is suitable for datasets with mixed data types. Decision trees are also useable, but require preprocessing to convert all features to binary features.

5. **Scalability**: **Naive Bayes** is computationally efficient and can handle very large datasets. Decision trees can handle large datasets too, but can become very computationally expensive if there is a large number of features. Boosted decision trees help with this

VOLUNTARY ASSIGNMENT



Mean classification accuracy: 71.4
Standard deviation: 6.17

THANK YOU FOR LISTENING!