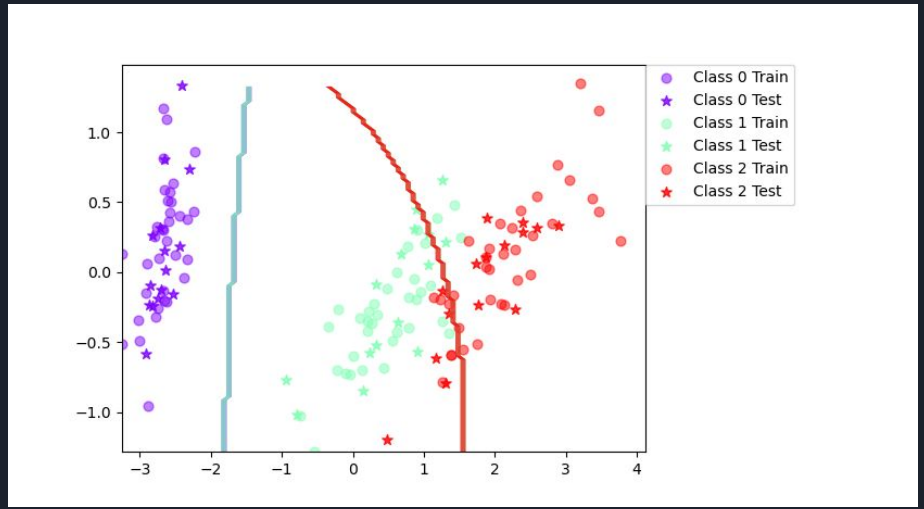
Lab 3

Erik Persson & Simon Larspers Qvist

Assignment 3 - Questions

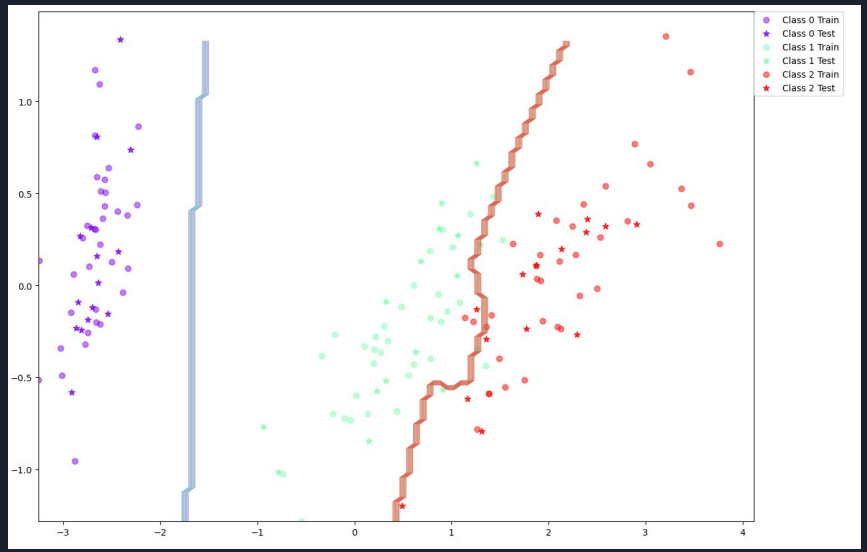
- It can be reasonable to assume independence when we know that all of, or most of, the features are independent of each other.
  - Naïve-Bayes can work pretty well with dependent data anyway, some features may cancel each other out.

- We can choose a more complex model, e.g Non-Naïve Bayes. We could use PCA to modify the data to be easier to separate by features.



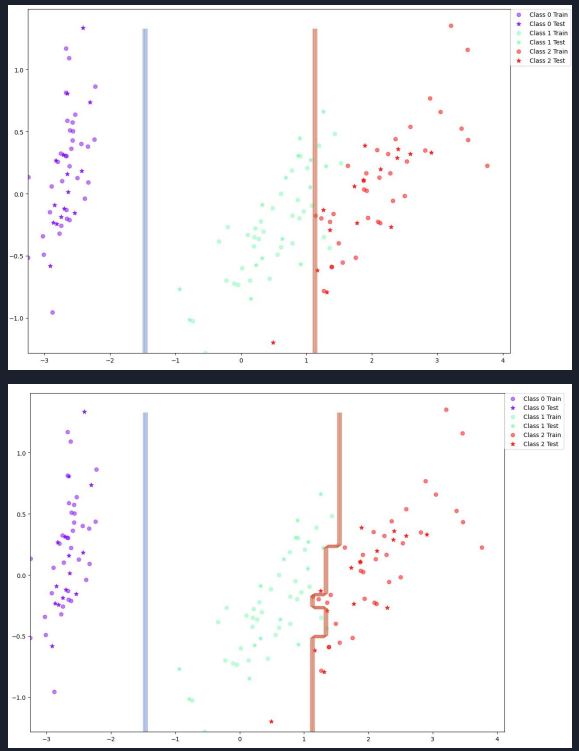
Assignment 5 - Questions

- There is an improvement. This is because we use multiple weak classifiers and combine their results. Better balance between bias and variance.
- The boundary of the boosted version is more complex. It follows a more reasonable direction in our opinion.
- Yes. With boosting we can improve the results of our simple model.
  But it's not a guarantee that it will be much better.



Assignment 6 - Questions

- Yes, there is an improvement, just like before.
- The boundary of the boosted version is more complex.
- As before, with boosting we can improve the results of our simple model.
  But it's not a guarantee that it will be much better.



Assignment 7 - Questions

Outliers: N-Bayes more robust, gives more of a generalization, while D-tree risk overfitting. (Boosting could make outliers have more of an affect)

Irrelevant features: None of them are sensitive to irrelevant features. (Ignored) Boosting should not change this.

Predictive power: D-trees gave us a slightly better result, both with and without boosting. Boosted D-tree gave the best result.

Mixed types: Both can handle this well. Just different way to do it.

Scalability: N-Bayes scales well over high dimensional data thanks to feature independence assumption and requires relatively low amount of data to train. D-Trees can grow quite complex.

We would choose a boosted algorithm, they are always better (but a little slower). D-Trees perform better here, we choose D-Trees for this reason, it is also faster.