BAYES CLASSIFIERS AND BOOSTING

Machine Learning – Lab 3

NAÏVE BAYES CLASSIFIER (I)

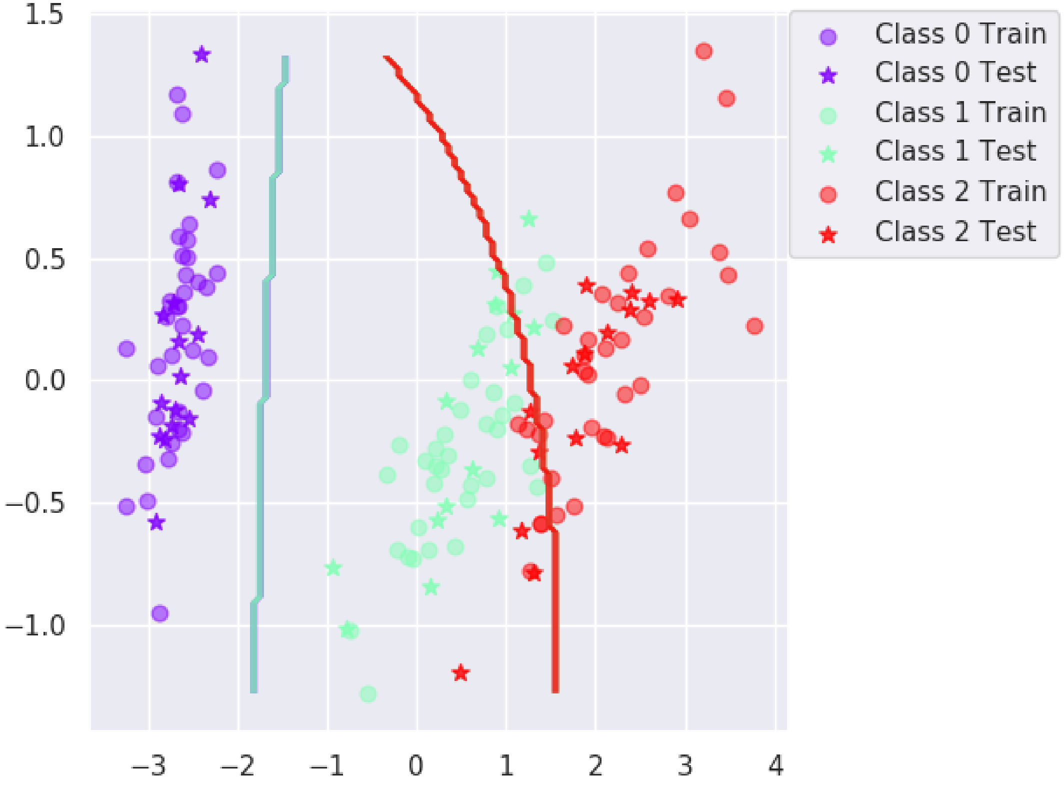
- Naïve **assumption**
  - Features conditionally **independent** given the class
  - $\Pr(\boldsymbol{x}|y) = \prod_{i=1}^{D} \Pr(xi|y)$ (D = dimensionality)
- Advantages
  - Learn D one-dimensional distributions instead of one D-dimensional distribution
  - Simplify calculations (e.g. inverse matrix for multivariate Gaussian distributions)
- Disadvantages
  - Assumption usually not true → can cause high bias

NAÏVE BAYES CLASSIFIER (II)

- Iris dataset
  - 2 correlated features
  - 3 classes

- Accuracy: $\mu = 89$, $\sigma = 4.16$
  - Only 2 features and relatively **simple** decision boundary, but…

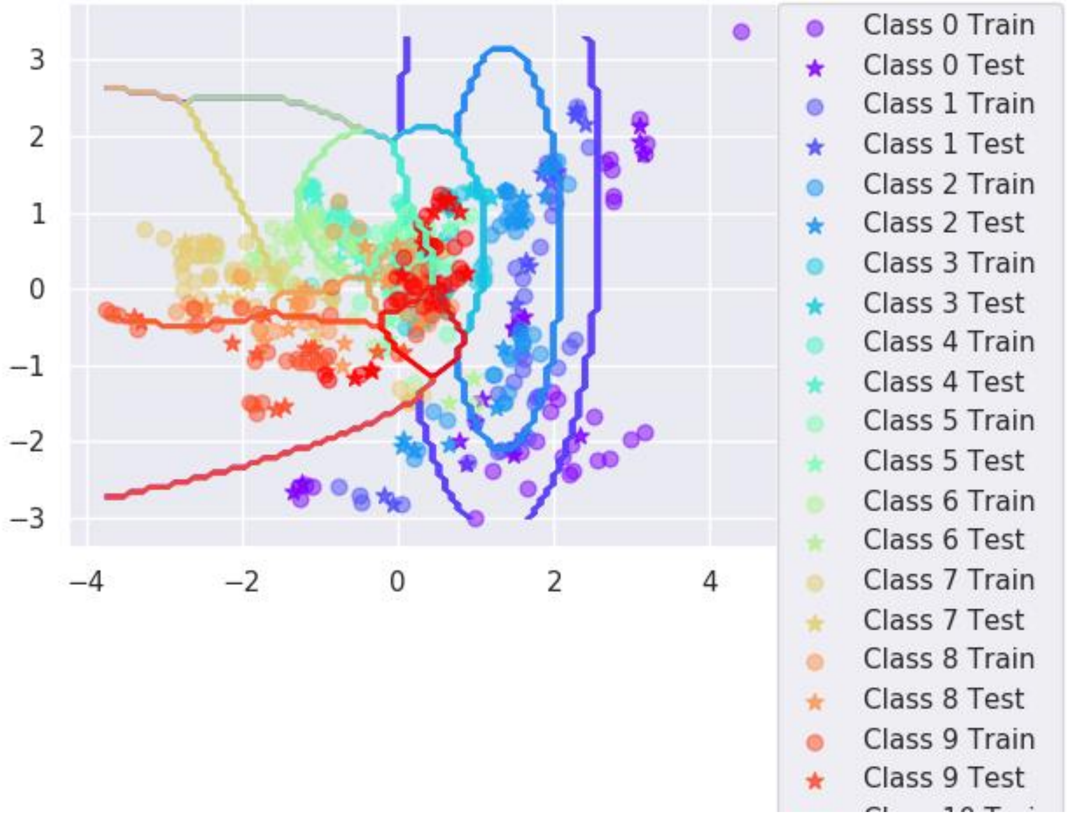
- **Not good** decision boundary between class 1 and class 2



NAÏVE BAYES CLASSIFIER (III)

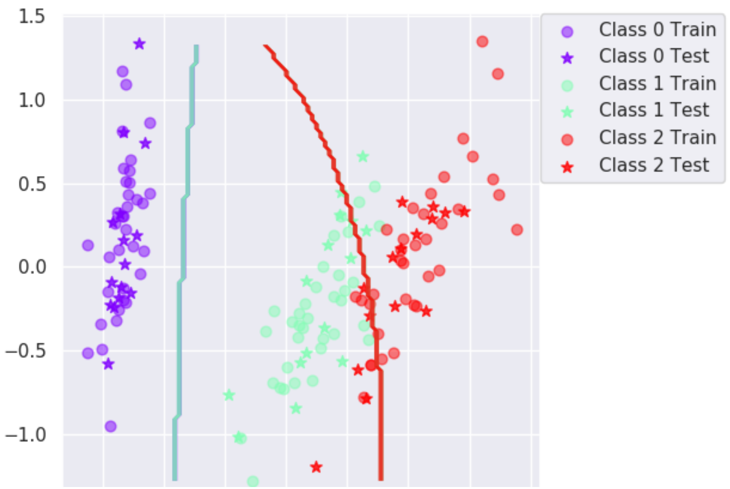
- Vowel dataset
  - 10 correlated features
  - 11 classes

- Accuracy: μ = 64.7, σ = 4.03
  - Much worse, more **difficult**

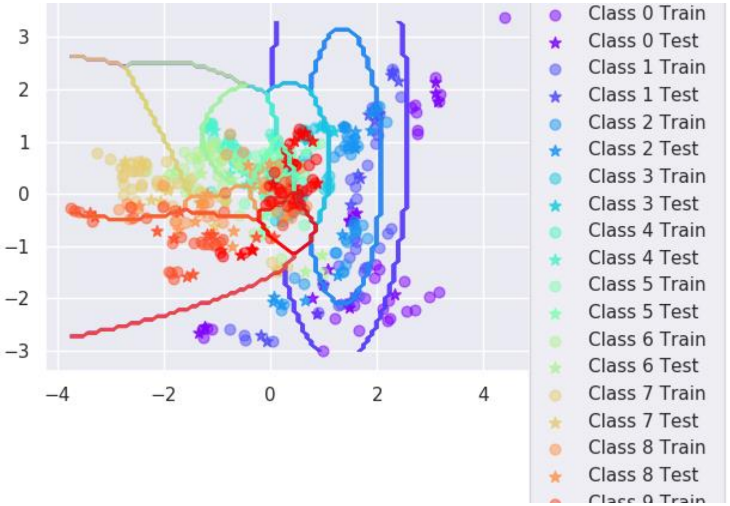


HOW TO IMPROVE

- Change classifier
  - **SVM** (extended for more than 2 classes, one-versus-one or one-versus-all)



- Boosting



BOOSTING (I)

- Combine **weak** classifiers (high bias) in one **strong** classifier (low bias)
- Naïve Bayes classifiers weak due to the **unrealistic assumption**
- Our boosted Naïve Bayes classifier
  - **Adaboost** algorithm
  - 10 Naïve Bayes classifiers (T=10)

BOOSTING (II)

- Iris dataset
  - 2 correlated features
  - 3 classes
- Accuracy: $\mu = 94.7$, $\sigma = 2.91$
  - Without boosting: $\mu = 89$, $\sigma = 4.16$
- **Improved**

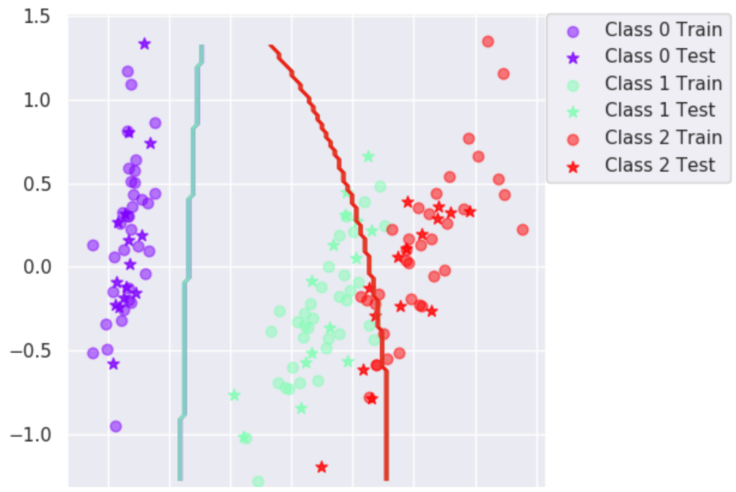
- Vowel dataset
  - 10 correlated features
  - 11 classes
- Accuracy: $\mu = 92.4$, $\sigma = 3.71$
  - Without boosting: $\mu = 64.7$, $\sigma = 4.03$
- **Improved dramatically**

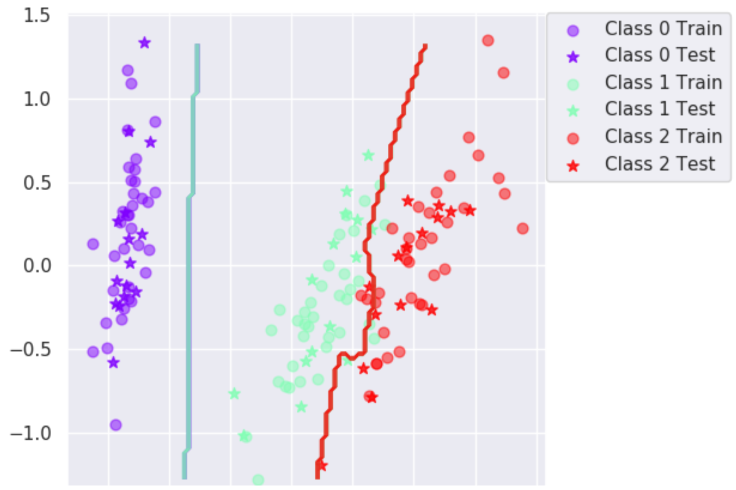
BOOSTING (III)

- Iris dataset
  - 2 correlated features
  - 3 classes

- Accuracy: μ = 92.4, σ = 3.71
  - Without boosting: μ = 64.7, σ = 4.03

- **More complex** decision boundary, fits data better

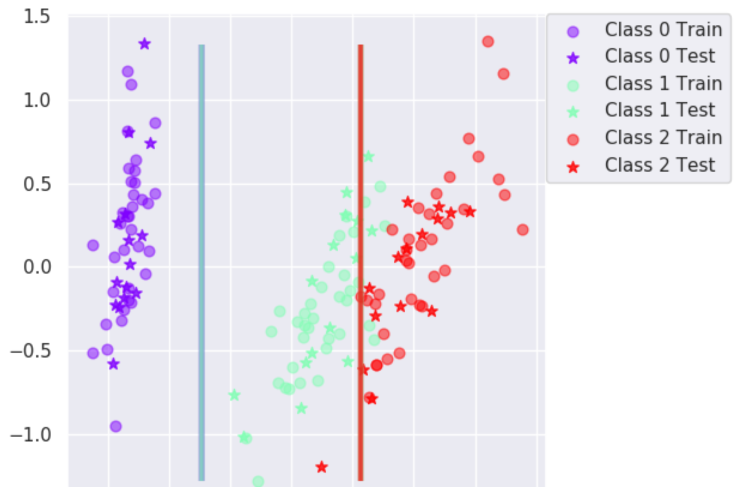


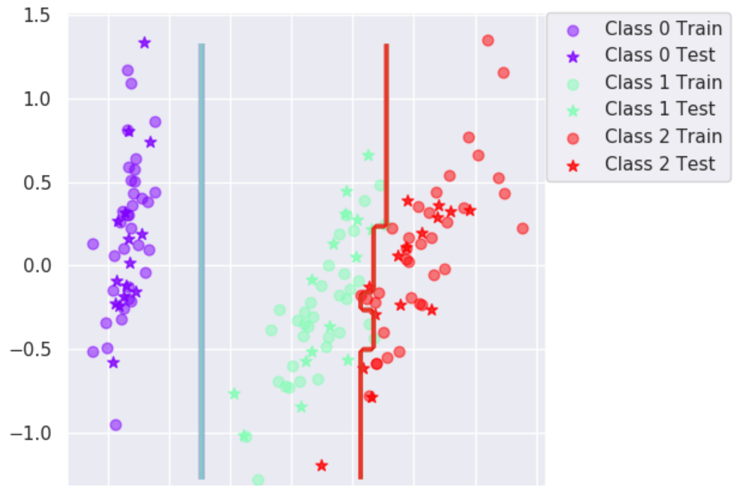


DECISION TREES (I)

- Iris dataset
  - 2 correlated features
  - 3 classes

- Accuracy
  - Without boosting: $\mu = 92.4$, $\sigma = 3.71$
  - With boosting: $\mu = 94.6$, $\sigma = 3.65$

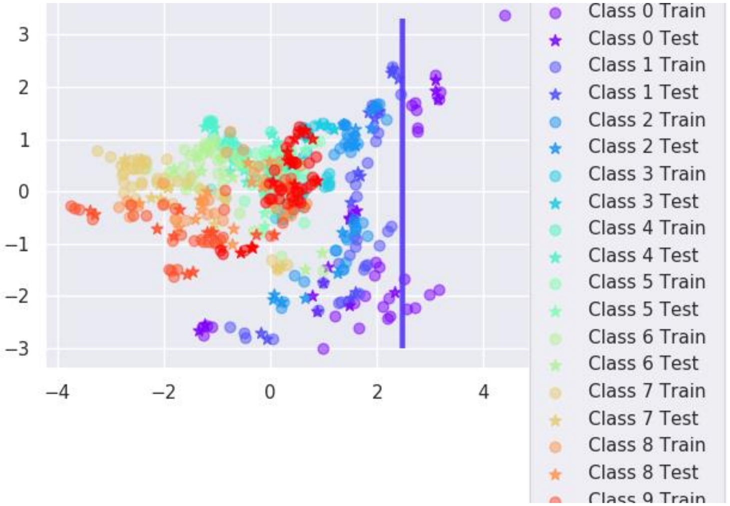


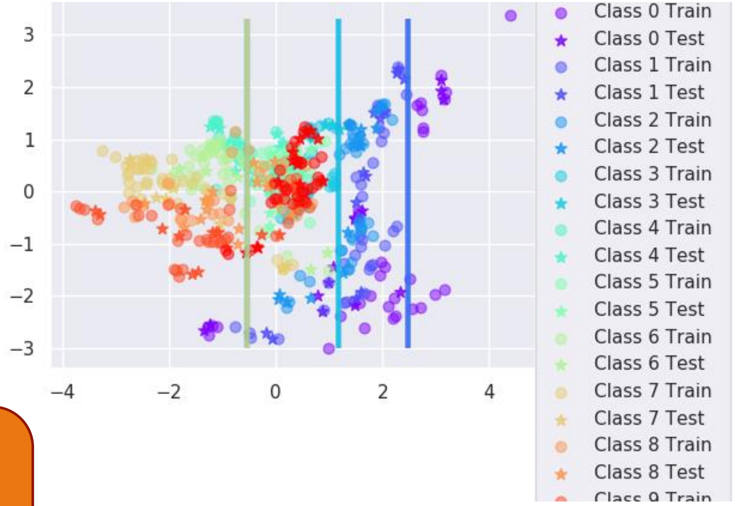


DECISION TREES (II)

- Vowel dataset
  - 10 correlated features
  - 11 classes

- Accuracy
  - Without boosting: $\mu = 64.1$, $\sigma = 4$
  - With boosting: $\mu = 86.9$, $\sigma = 3.07$





The plots are 2D but actually the dataset has 10 features, so decision boundaries are not visualized well.

NAÏVE BAYES VS DECISION TREE

- Outliers
  - Naïve Bayes is robust (based on probabilities)
  - DT is robust (nodes are determined based on purity indexes of partitions, not single points)
  - NO boosting (weights corresponding to outliers increase and outliers become more important)
- Irrelevant features
  - Naïve Bayes does not ignore them (probabilities of all features are always computed)
  - DT can ignore them when pruning (no node uses them)
- Predictive power
  - Naïve Bayes has more expressiveness than DT (DT has decision boundaries parallel to axes)
  - Boosting increases expressiveness
- Mixed types of data
  - Naïve Bayes is more expressive for continuous attributes than DT
  - Boosting increases accuracy
- Scalability
  - DT is much faster than Naïve Bayes (Naïve Bayes has to compute probabilities)
  - Boosting can be time consuming depending on T