Model Overfitting

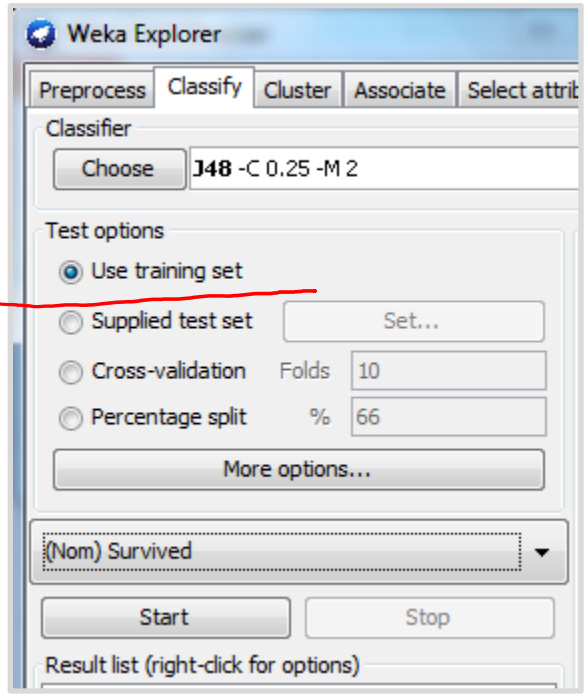
School of Information Studies
Syracuse University

Model Generalization

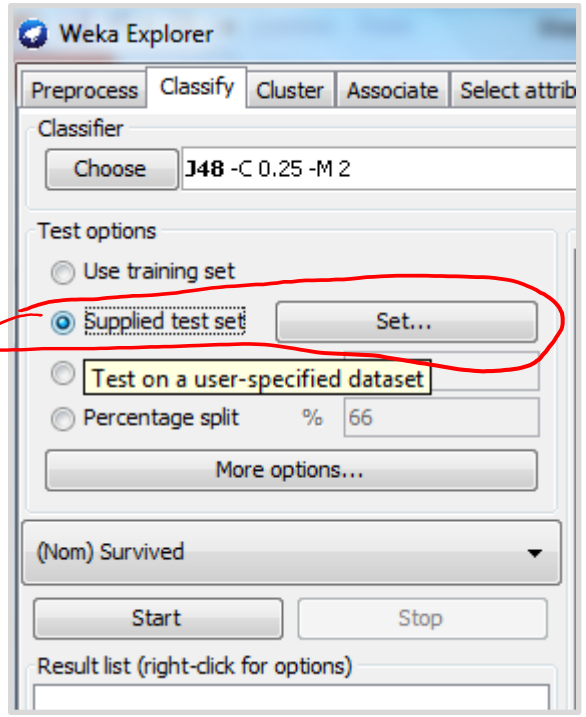
Two fundamental concepts:

- ***Training error**: train a model (e.g., a decision tree model) on the training data set, then test the model on the same training set. The error rate is called "training error," which evaluates how well the model **fits** the training data.
- ***Test error**: test the model on a test data set that is different from the training set. The error rate is called "test error," which evaluates how well the model **generalizes** to unseen data.

Training Error vs. Test Error



Weka: the evaluation option to obtain **training error**



Weka: the evaluation option to obtain **test error**

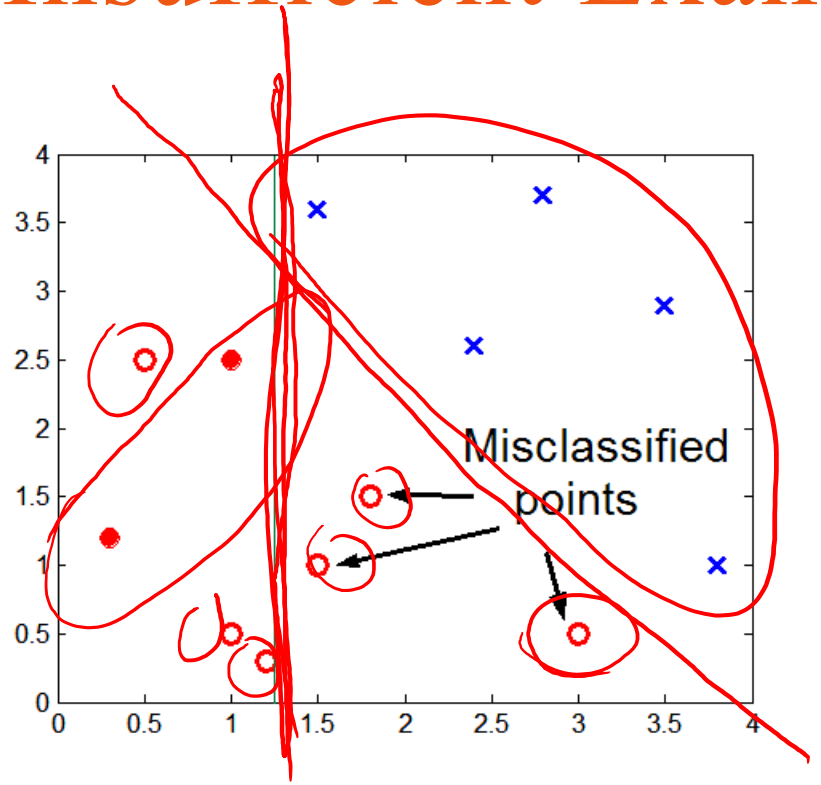
Model Overfitting

Overfitting means a model fits the training data very well, but generalizes to unseen data poorly.

How do I know if my model is overfitting?

- Your model is overfitting if its **training error is small** (fits well with training data), but **the test error is large** (generalizes poorly to unseen data).
- Did it happen to your Naïve Bayes model?

Overfitting Due to Insufficient Examples



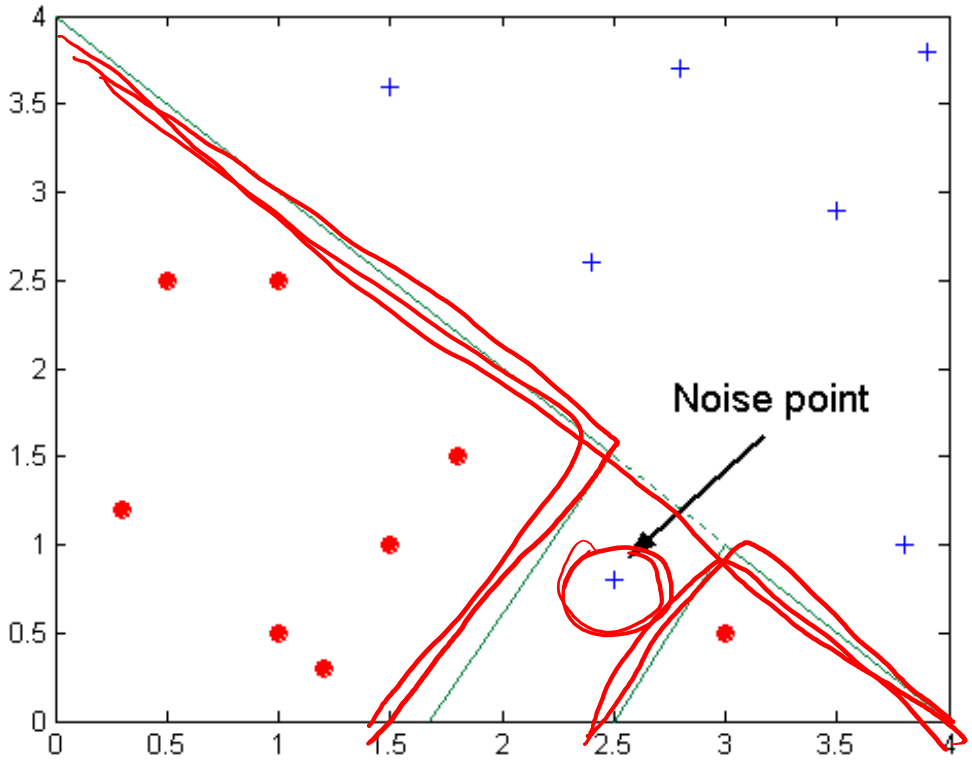
Blue crosses and solid red dots are training data.

Red circles are test data.

The green vertical line is the decision boundary.

Lack of data points in the lower half of the diagram makes it difficult to predict correctly the class labels in that region.

Overfitting Due to Noise



The decision boundary (supposedly a straight line) is distorted by the noise point. The overfitted decision boundary is the solid blue lines.

Occam's Razor

Given two models of similar generalization errors, the simpler model is preferred over the more complex model.

For complex models, there is a greater chance that it was overfitted accidentally by errors in data.

Therefore, model complexity should be considered when evaluating a model.