MODEL OVERFITTING

SYRACUSE UNIVERSITY
School of Information Studies

MODEL COMPLEXITY AND OVERFITTING

Complex models are more likely to overfit than simple models.

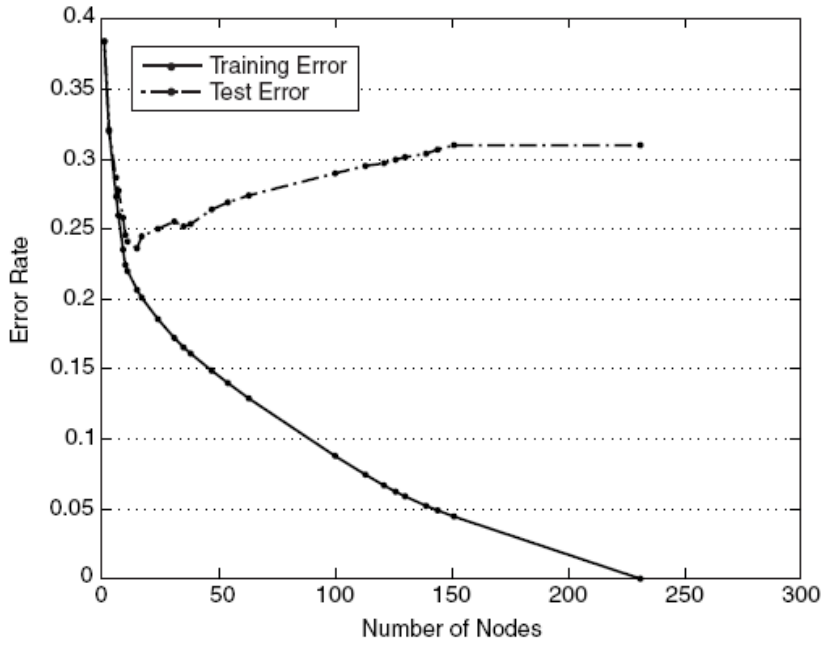
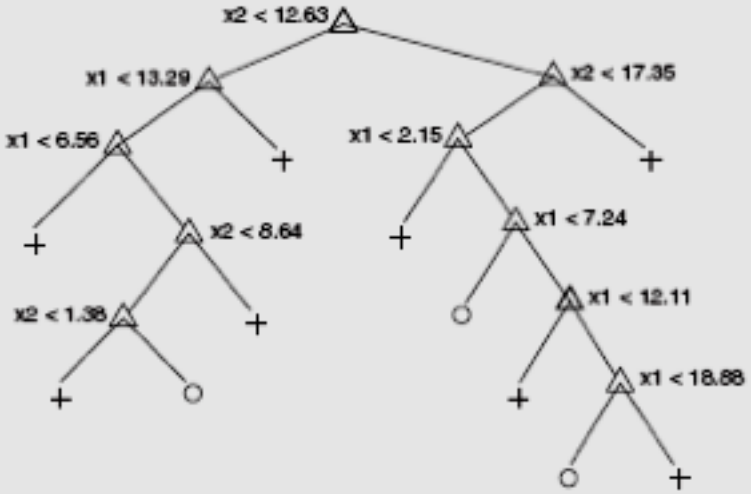


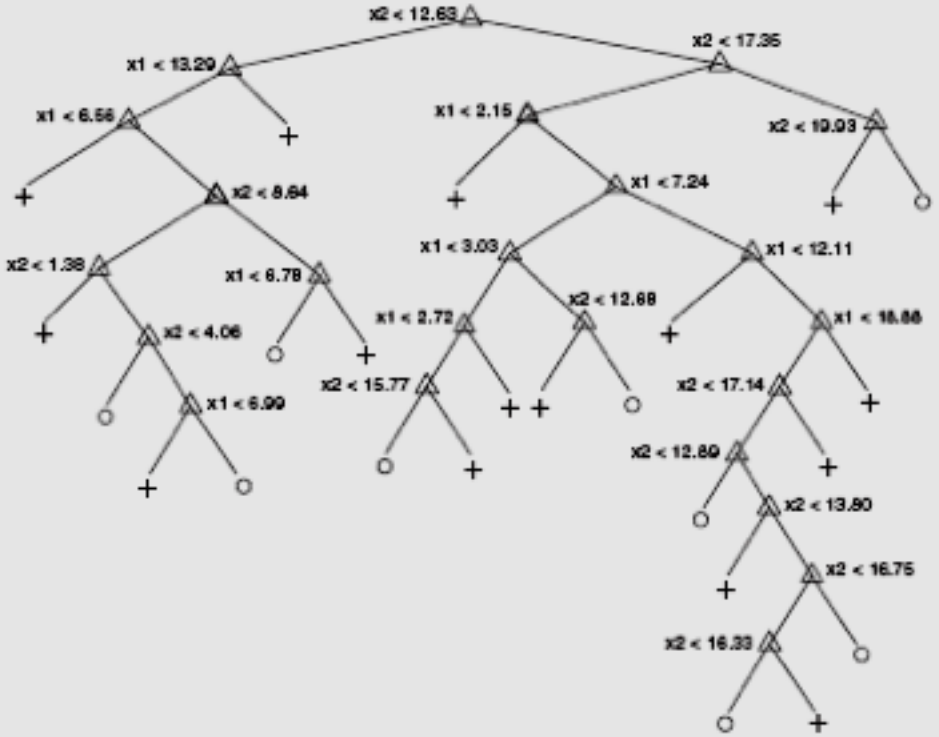
Figure 4.23. Training and test error rates.

For decision tree, number of nodes indicates model complexity.

Higher number of nodes -> higher model complexity -> lower training error and higher test error



(a) Decision tree with 11 leaf nodes.



(b) Decision tree with 24 leaf nodes.

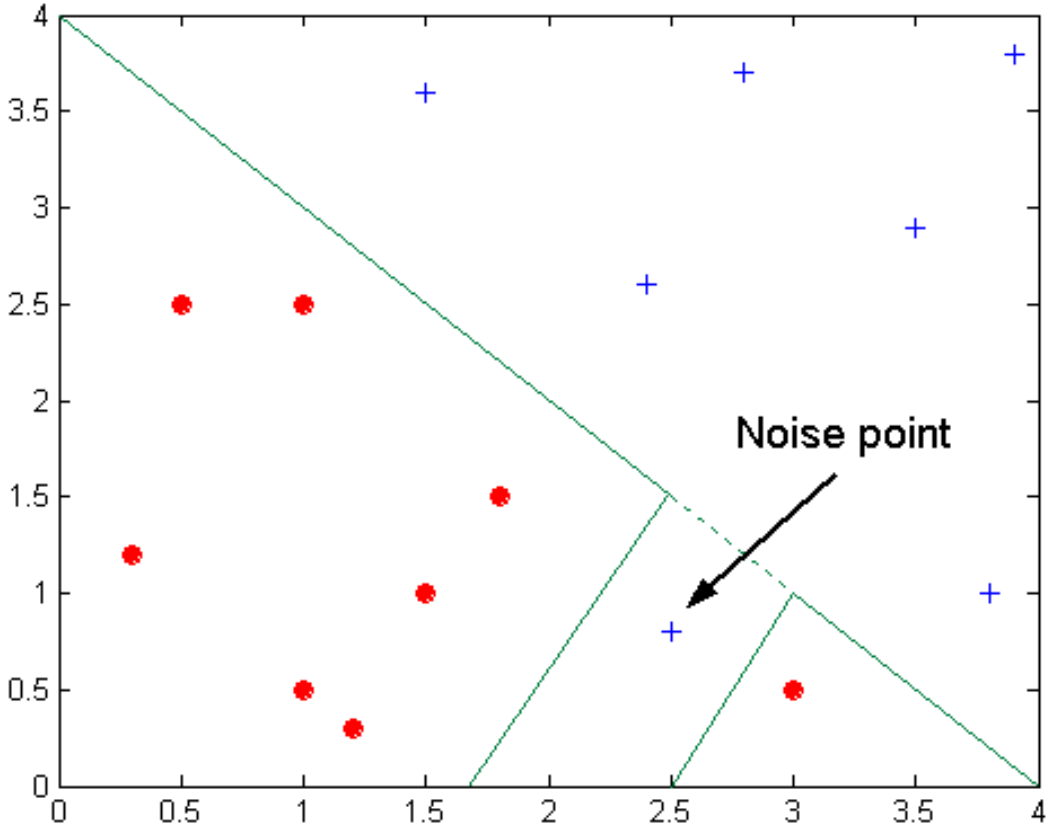
Figure 4.24. Decision trees with different model complexities.

MAIN REASONS FOR MODEL OVERFITTING

Overfitting due to noise

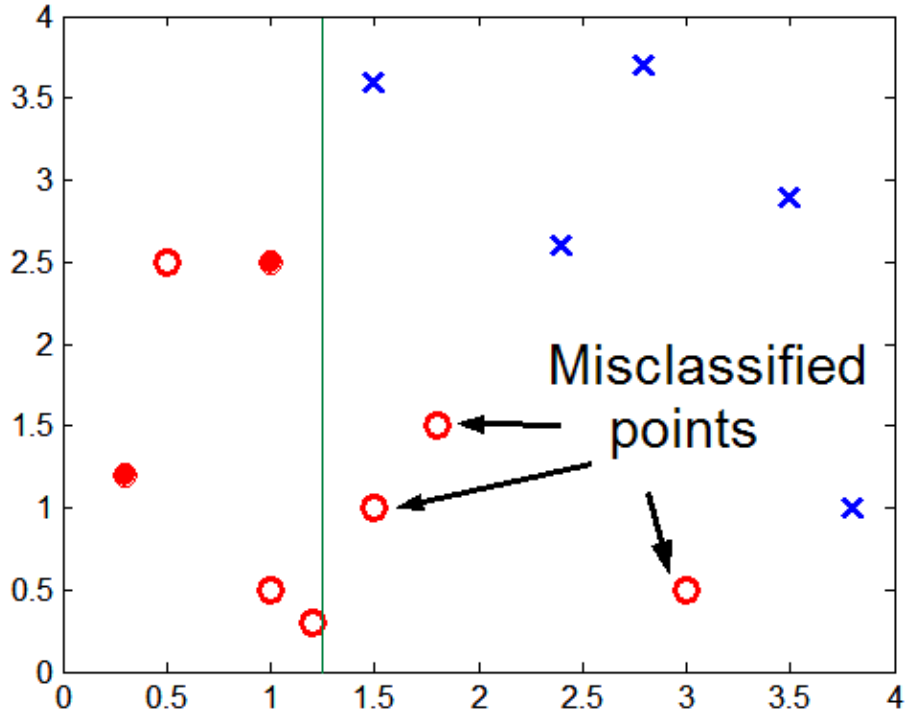
Overfitting due to insufficient samples

OVERFITTING DUE TO NOISE



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

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 created by a simple decision tree (if x > 1.25, label = blue; otherwise, label = red).

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

OCCAM'S RAZOR

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

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

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