

Artificial Intelligence for Medicine II

Spring 2025

Lecture 63: Supervised Learning Overfitting and Underfitting

(Many slides adapted from Bing Liu, Han, Kamber & Pei; Tan, Steinbach, Kumar and the web)

Underfitting and overfitting

- Underfitting and overfitting are common issues in machine learning that can affect the performance of models:

Underfitting: This occurs when a model is too simple to capture the underlying patterns in the data. An underfit model performs poorly on both the training data and new, unseen data. It fails to learn the relationships within the data, leading to high errors across the board

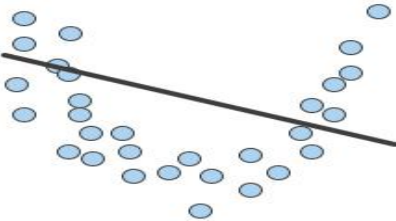
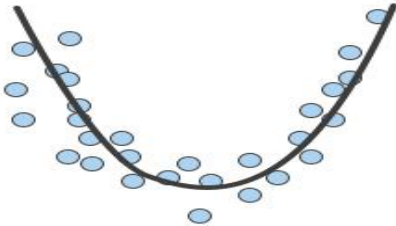

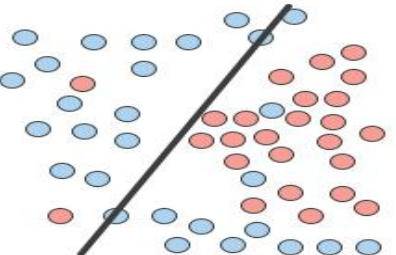
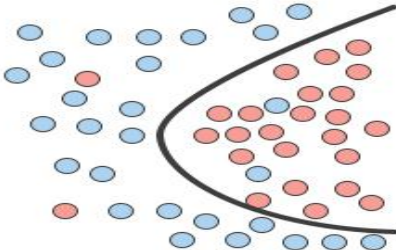
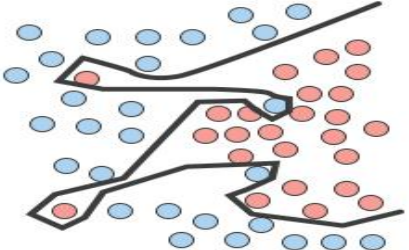
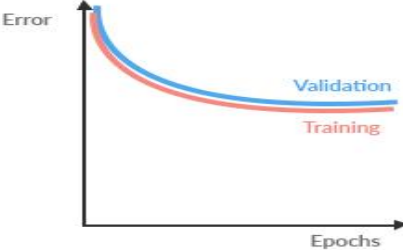
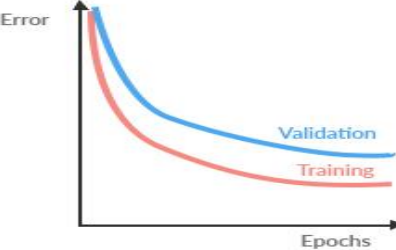
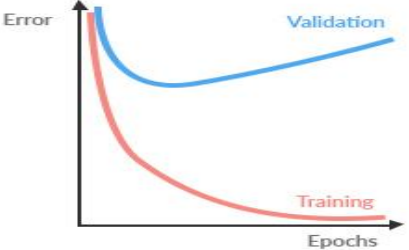
For example, using a linear regression model to predict a complex, non-linear relationship will likely result in underfitting.

Overfitting: This happens when a model is too complex and learns the noise and details of the training data rather than the actual patterns. An overfit model performs very well on the training data but poorly on new, unseen data because it fails to generalize

For instance, a deep neural network with too many layers might memorize the training data, leading to overfitting.

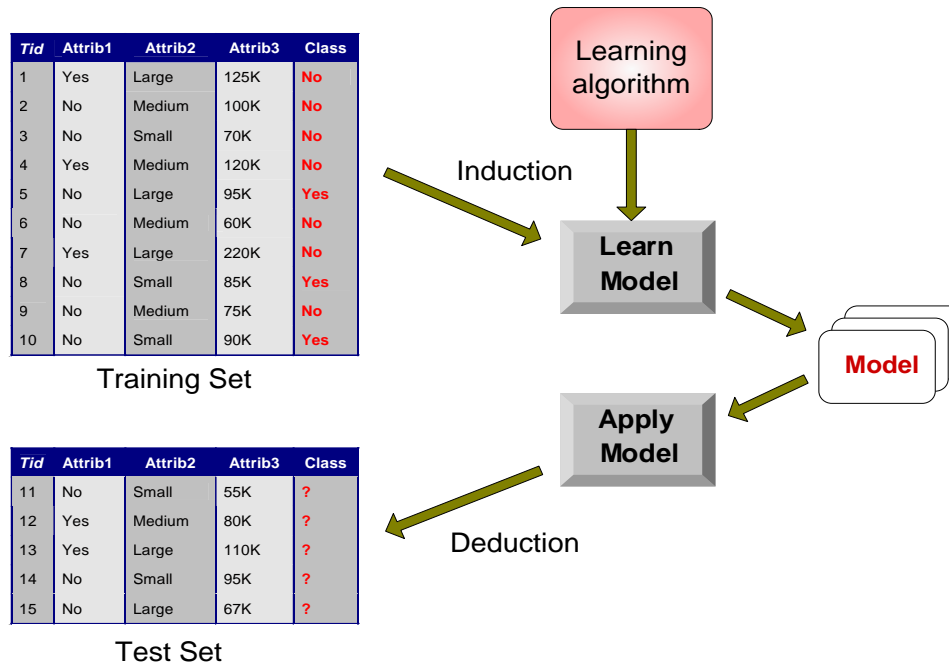
- To achieve optimal model performance, it's crucial to find a balance between underfitting and overfitting. This involves selecting the right model complexity and using techniques like cross-validation, regularization, and pruning

Underfitting and overfitting

| | Underfitting | Just right | Overfitting |
|-----------------------------|--|---|---|
| Symptoms | <ul style="list-style-type: none"> • High training error • Training error close to test error • High bias | <ul style="list-style-type: none"> • Training error slightly lower than test error | <ul style="list-style-type: none"> • Very low training error • Training error much lower than test error • High variance |
| Regression illustration |  |  |  |
| Classification illustration |  |  |  |
| Deep learning illustration |  |  |  |
| Possible remedies | <ul style="list-style-type: none"> • Complexify model • Add more features • train longer | | <ul style="list-style-type: none"> • Perform regularization • Get more data |

Classification Errors

- **Training errors:** Errors committed on the training set
- **Test errors:** Errors committed on the test set
- **Generalization errors:** Expected error of a model over random selection of records from same distribution



Example Data Set

Two class problem:

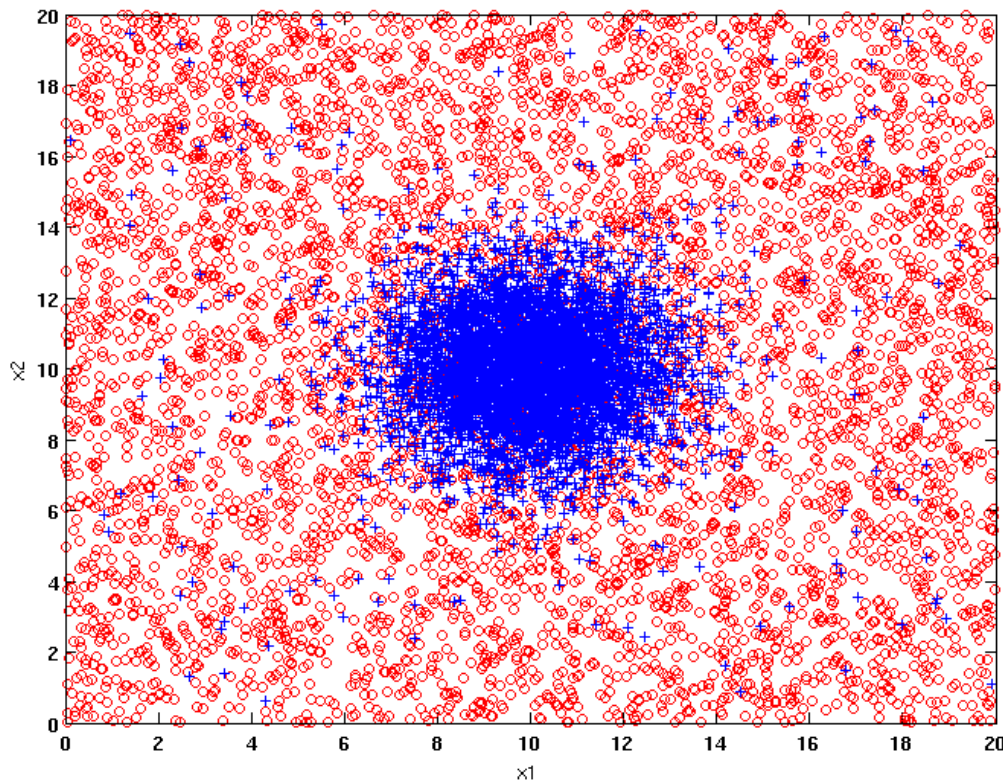
+ : 5400 instances

- 5000 instances generated from a Gaussian centered at (10,10)
- 400 noisy instances added

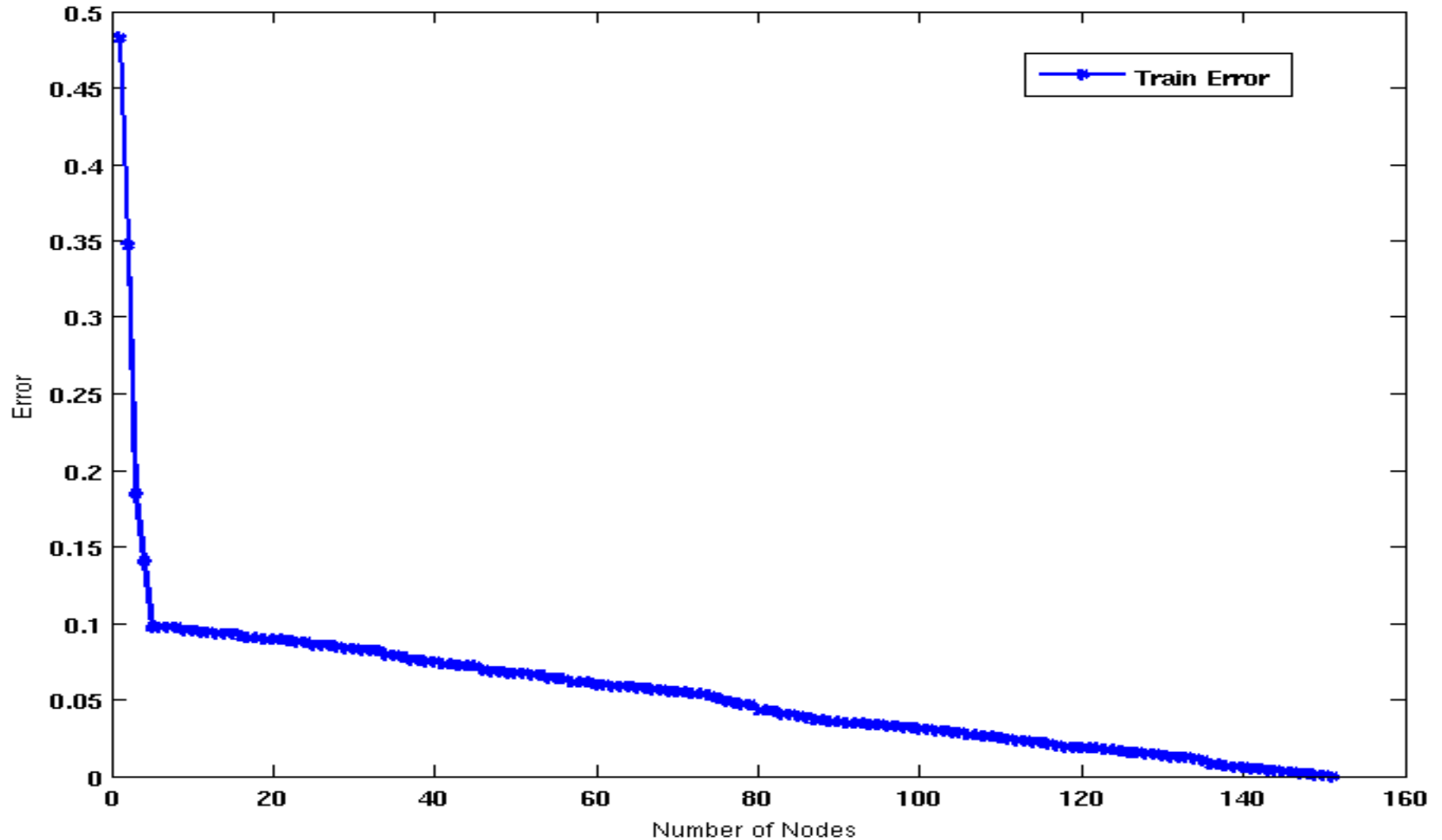
o : 5400 instances

- Generated from a uniform distribution

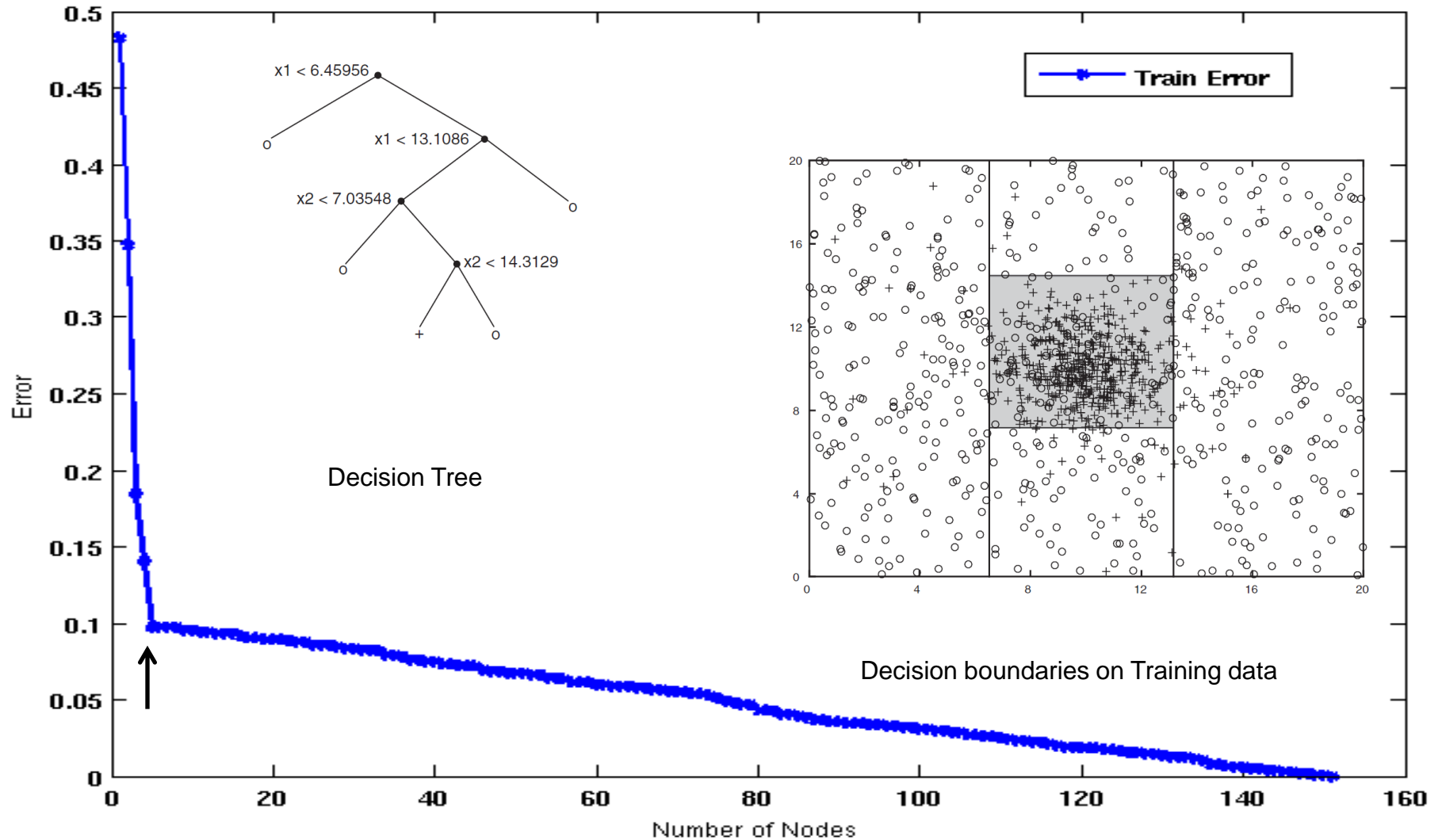
10 % of the data used for training and 90% of the data used for testing



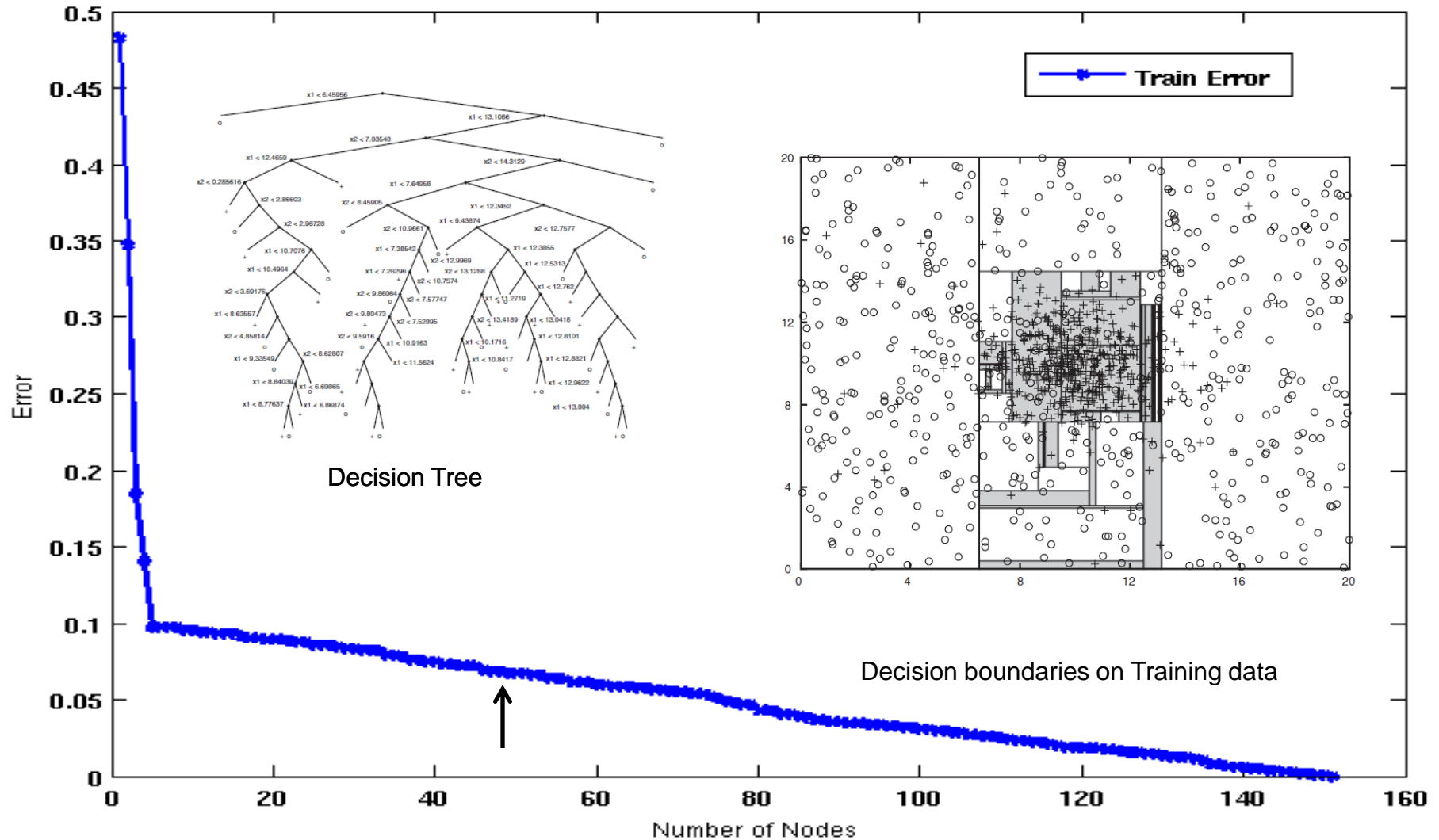
Increasing number of nodes in Decision Trees



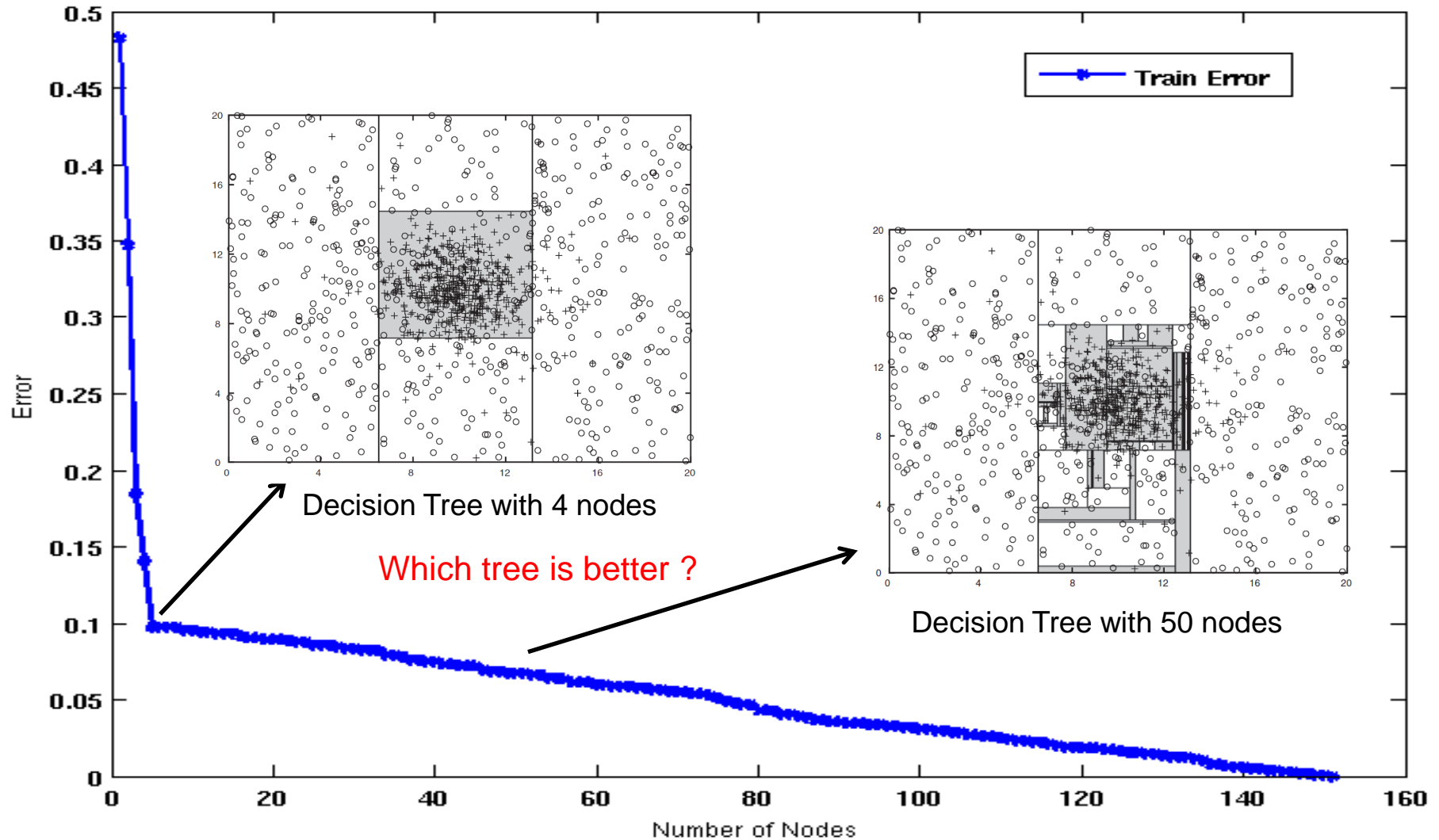
Decision Tree with 4 nodes



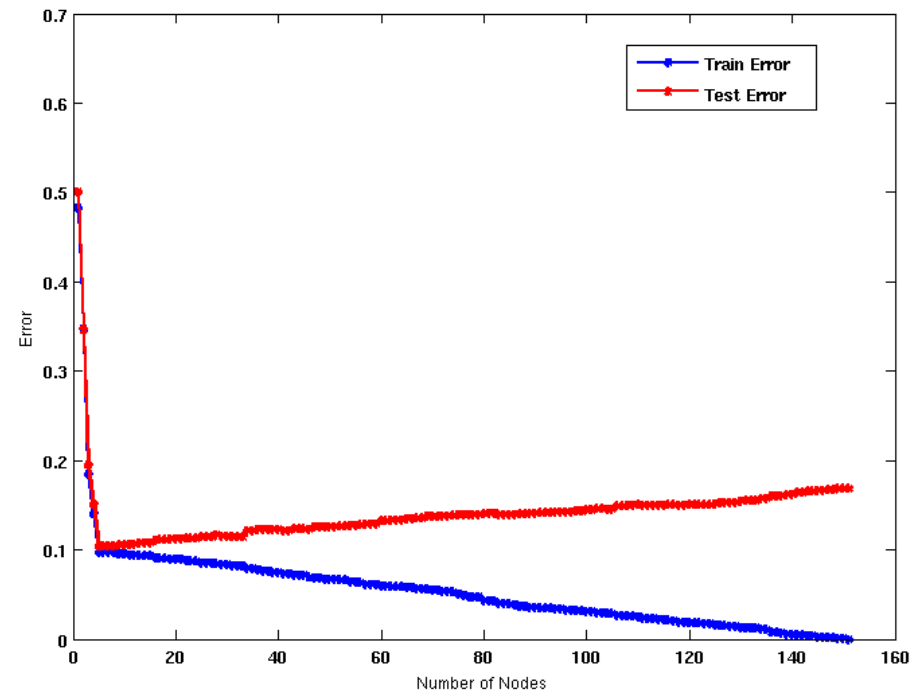
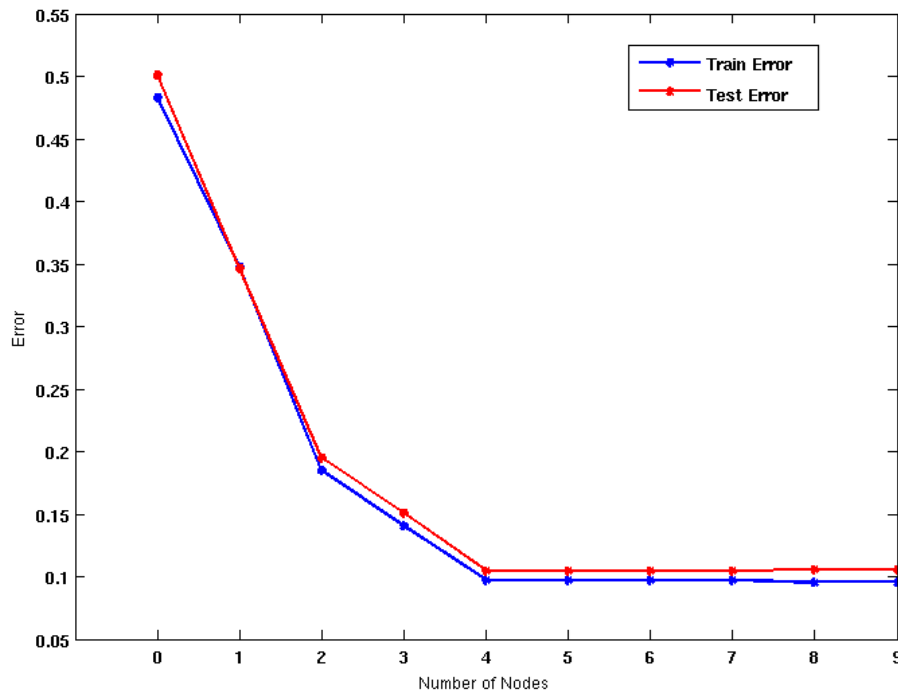
Decision Tree with 50 nodes



Which tree is better?



Model Underfitting and Overfitting

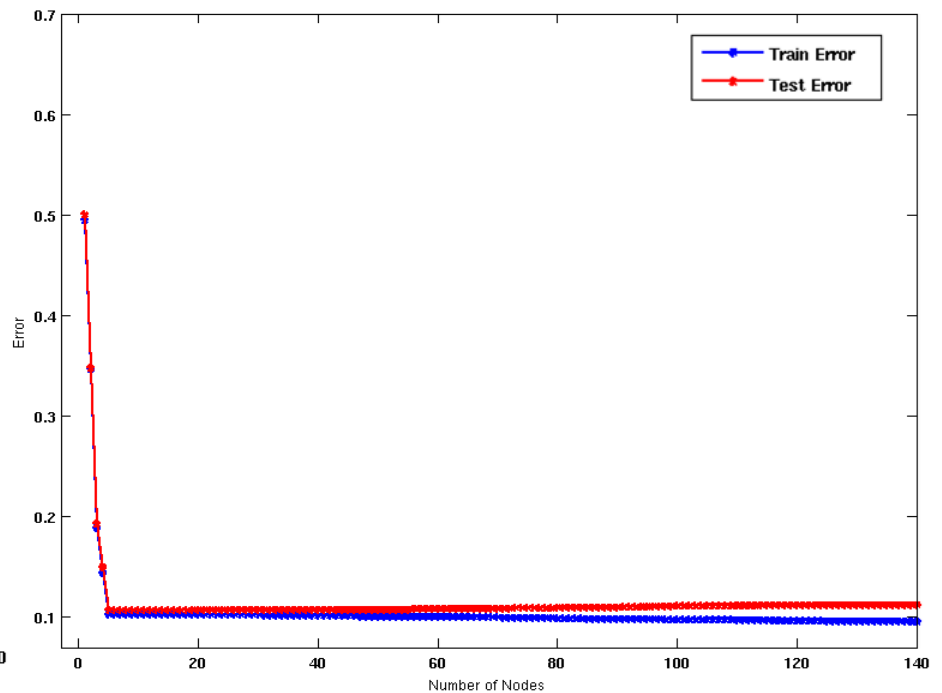
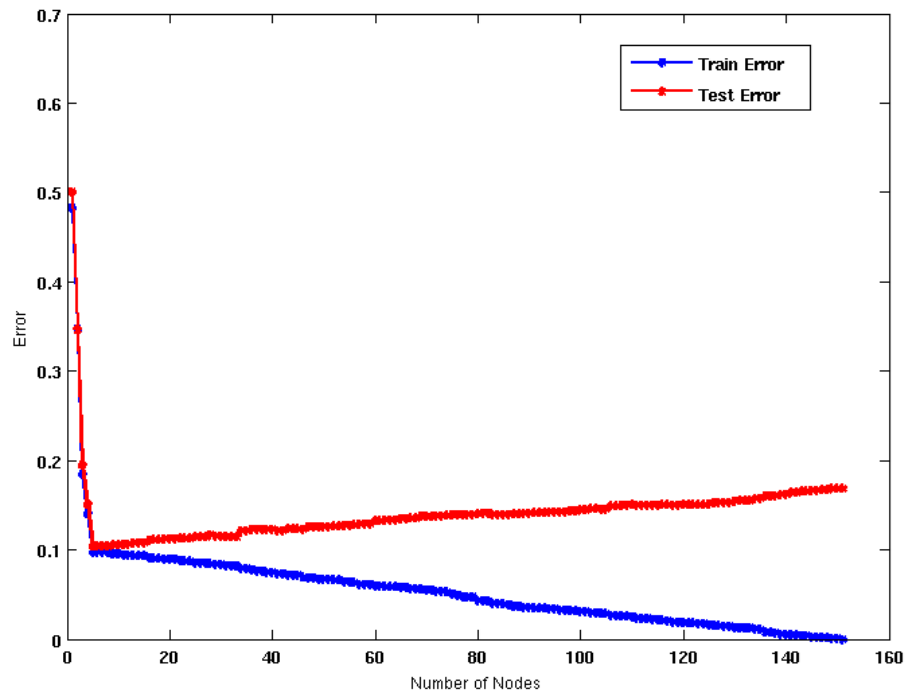


- As the model becomes more and more complex, test errors can start increasing even though training error may be decreasing

Underfitting: when model is too simple, both training and test errors are large

Overfitting: when model is too complex, training error is small but test error is large

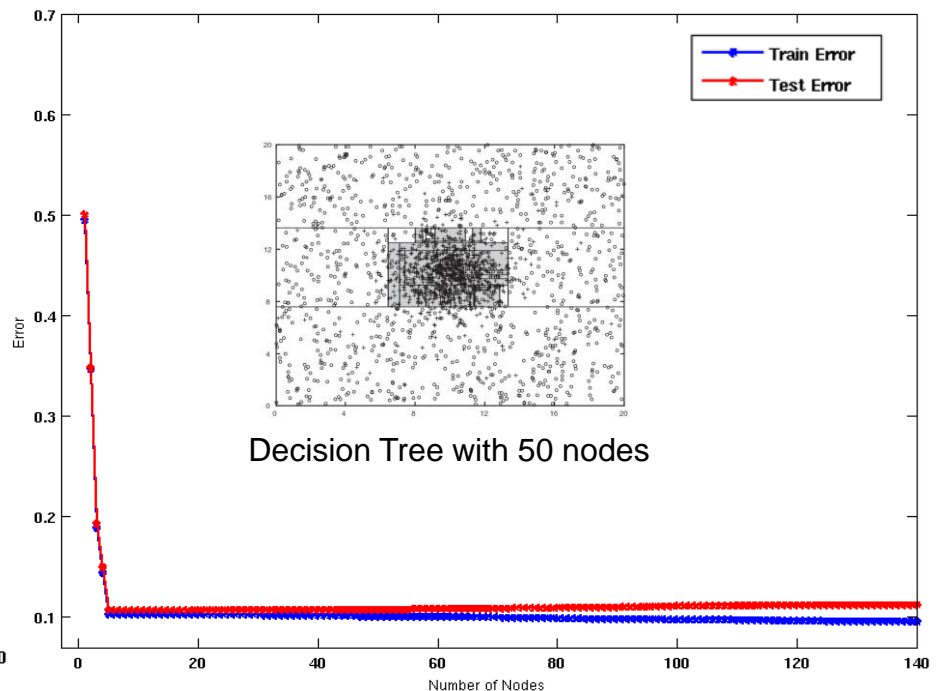
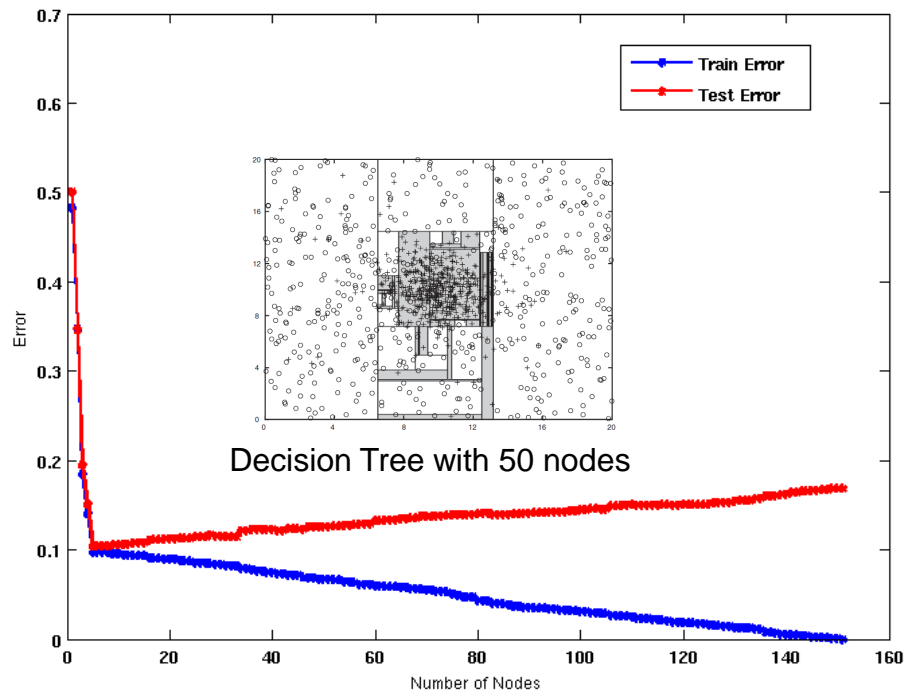
Model Overfitting – Impact of Training Data Size



Using twice the number of data instances

- Increasing the size of training data reduces the difference between training and testing errors at a given size of model

Model Overfitting – Impact of Training Data Size



Using twice the number of data instances

- Increasing the size of training data reduces the difference between training and testing errors at a given size of model

Reasons for Model Overfitting

- Not enough training data
- High model complexity
 - Multiple Comparison Procedure

Notes on Overfitting

- Overfitting results in decision trees that are more complex than necessary
- Training error does not provide a good estimate of how well the tree will perform on previously unseen records
- Need ways for estimating generalization errors

Model Selection

- Performed during model building
- Purpose is to ensure that model is not overly complex (to avoid overfitting)
- Need to estimate generalization error
 - Using Validation Set
 - Incorporating Model Complexity