Outline

- Classification basics
- Decision tree classifiers
- Overfitting
- Lazy vs Eager Learners
- k-Nearest Neighbors (or learning from your neighbors)
- Evaluation of classifiers

Training vs generalization errors

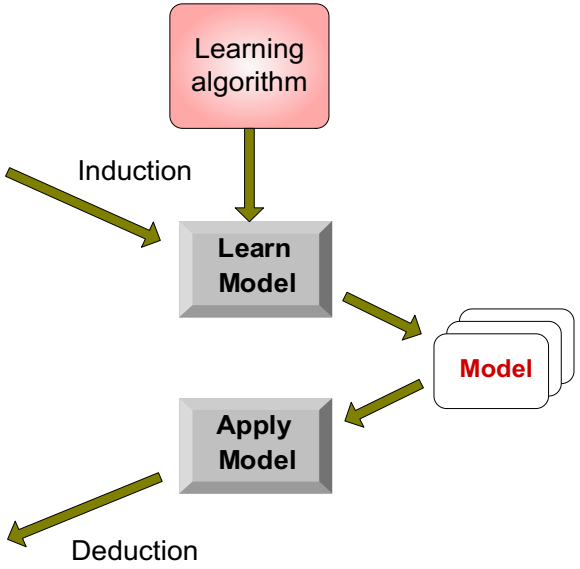
- The errors of a classifier are divided into
  - Training errors (or resubstitution error or apparent error):
    - errors committed in the training set
  - Generalization errors:
    - the expected error of the model on previously unseen examples
- A good classifier must
  1. Fit the training data &
  2. Accurately classify records never seen before
- i.e., a good model ➔ low training error & low generalization error

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Model overfitting

- Model overfitting
  - A model that fits the training data well (low training error) but has a poor generalization power (high generalization error)
- Overfitting: Consider an hypothesis *h*
  - $error_{train}(h)$: the error of *h* in the training set
  - $error_{D}(h)$: the error of *h* in the entire distribution *D* of data (i.e., including instances beyond the training set)
  - Hypothesis *h* overfits training data if there is an alternative hypothesis *h'* in *H* such that:

$$error_{train}(h) < error_{train}(h')$$

and

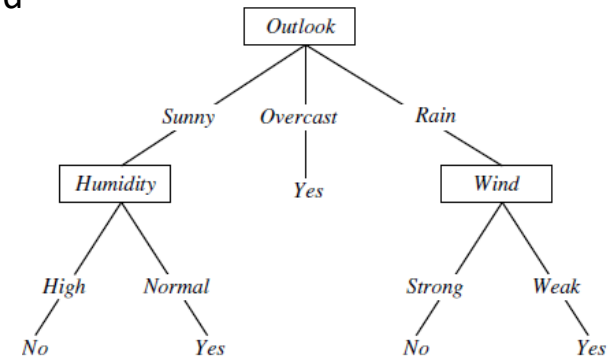
$$error_{\mathcal{D}}(h) > error_{\mathcal{D}}(h')$$

Decision trees overfitting

- An induced decision tree may overfit the training data
  - Too many branches, some may reflect anomalies due to noise or outliers
  - Very good performance in the training (already seen) samples
  - Poor accuracy for unseen samples
- Example
  - Let us add a *noisy/outlier* training example (D_{15}) to the training set
  - How the earlier tree (built upon training examples D_1-D_{14}) would be effected?

Training set

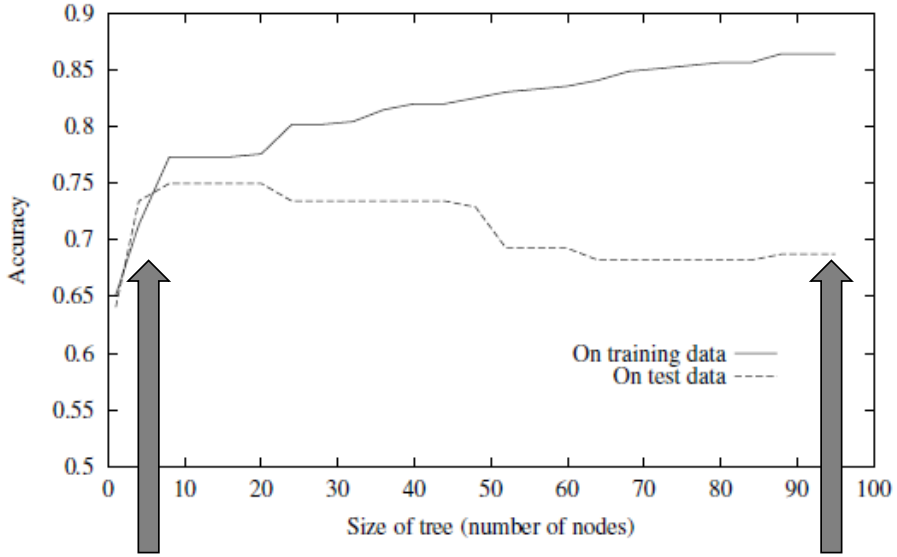
Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No
D15	Sunny	Hot	Normal	Strong	No



Underfitting & Overfitting

- The training error can be decreased by increasing the model complexity
- But, a complex DT, tailored to the training data model, will also have a high generalization error

How the error in both training and test data evolves with the tree complexity



The model has yet to learn the true structure from the training data.

Model underfitting

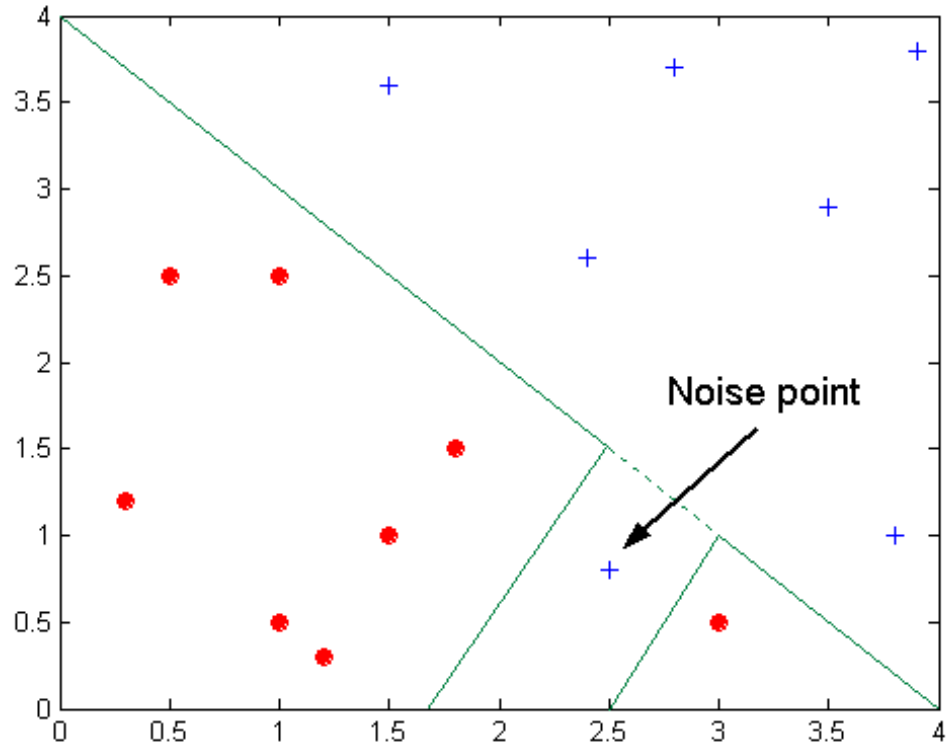
Model overfitting

The model overspecializes to the training data

Potential causes of model overfitting

- Overfitting due to presence of noise
- Overfitting due to lack of representative samples

Overfitting due to presence of noise



The decision boundary is distorted by the noise point.

Overfitting due to presence of noise – an example

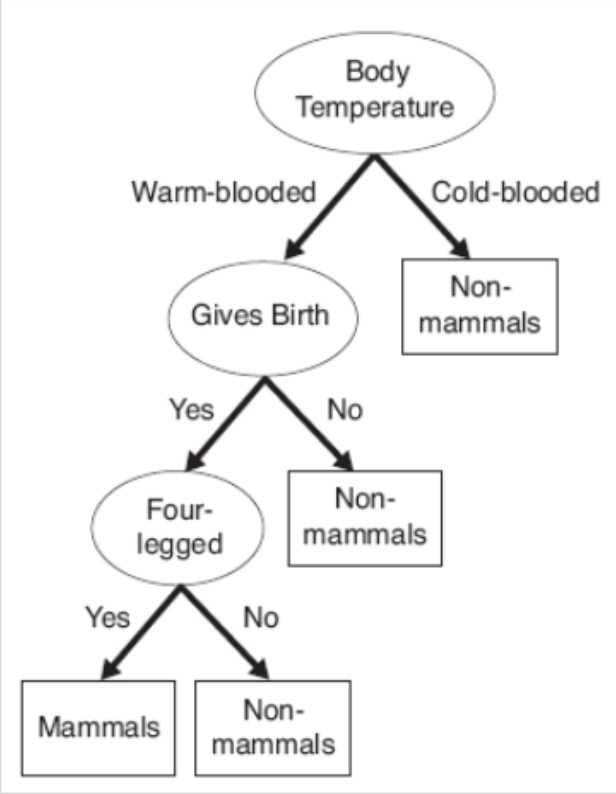
Training set

(stands for missclassified instances)*

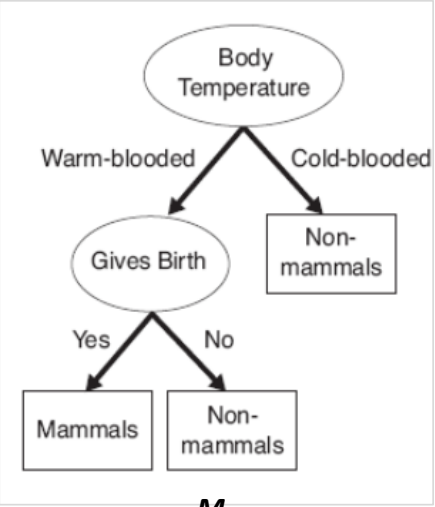
Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
porcupine	warm-blooded	yes	yes	yes	yes
cat	warm-blooded	yes	yes	no	yes
bat	warm-blooded	yes	no	yes	no*
whale	warm-blooded	yes	no	no	no*
salamander	cold-blooded	no	yes	yes	no
komodo dragon	cold-blooded	no	yes	no	no
python	cold-blooded	no	no	yes	no
salmon	cold-blooded	no	no	no	no
eagle	warm-blooded	no	no	no	no
guppy	cold-blooded	yes	no	no	no

Test set

Name	Body Temperature	Gives Birth	Four-legged	Hibernates	Class Label
human	warm-blooded	yes	no	no	yes
pigeon	warm-blooded	no	no	no	no
elephant	warm-blooded	yes	yes	no	yes
leopard shark	cold-blooded	yes	no	no	no
turtle	cold-blooded	no	yes	no	no
penguin	cold-blooded	no	no	no	no
eel	cold-blooded	no	no	no	no
dolphin	warm-blooded	yes	no	no	yes
spiny anteater	warm-blooded	no	yes	yes	yes
gila monster	cold-blooded	no	yes	yes	no

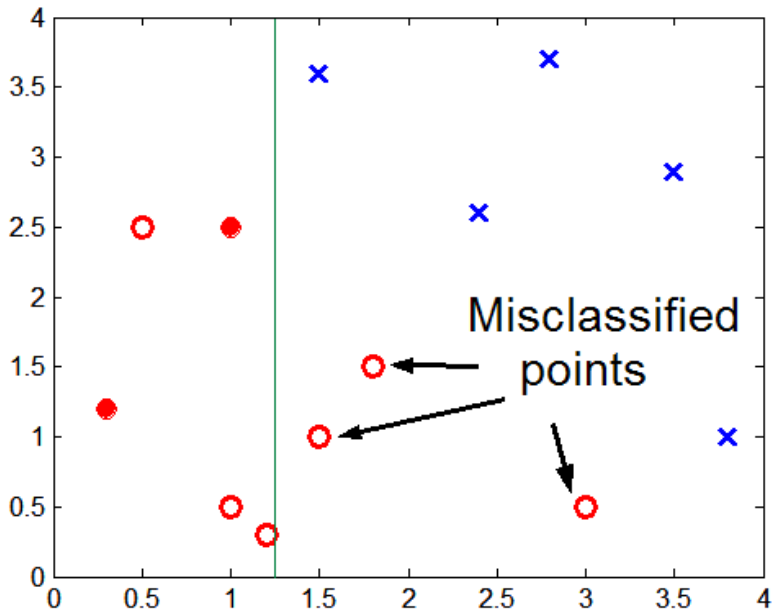


M_1
Training error: 0
Test error: 30%



M_2
Training error: 20%
Test error: 10%

Overfitting due to lack of representative samples



- Lack of data points in the lower half of the diagram makes it difficult to predict correctly the class labels of that region
  - Insufficient number of training records in the region causes the decision tree to predict the test examples using other training records that are irrelevant to the classification task

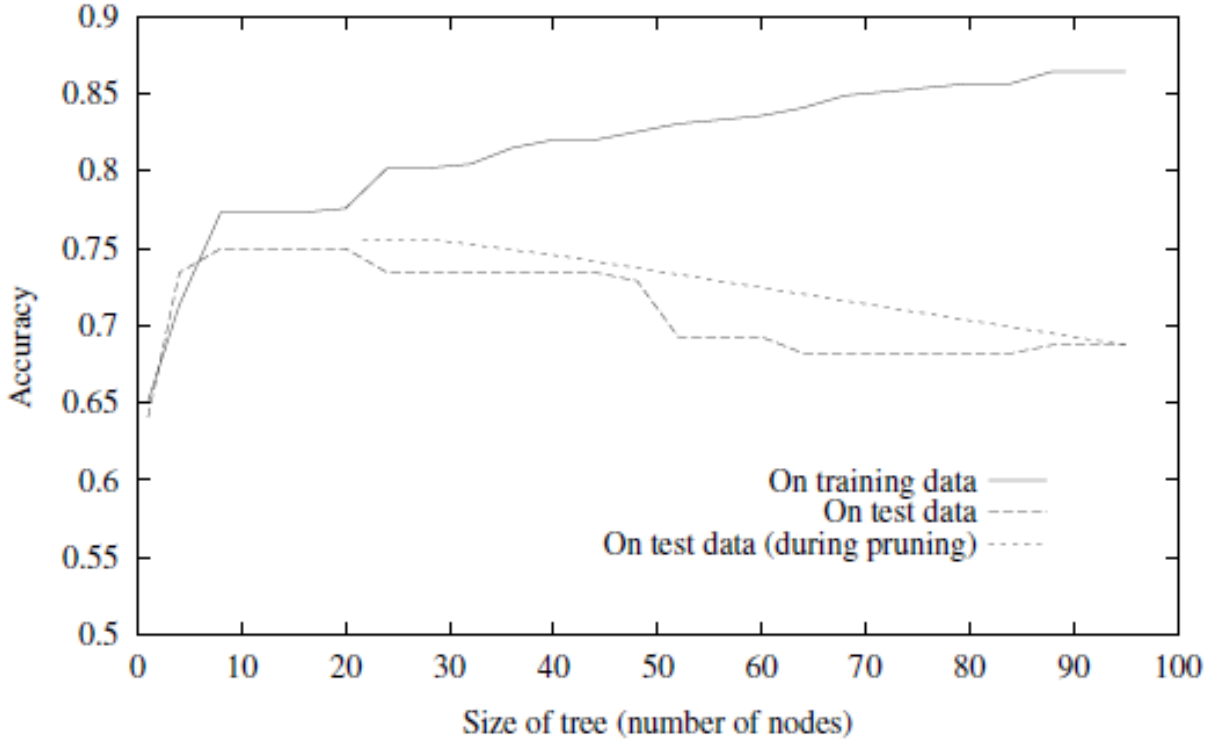
Avoiding overfitting in decision trees

- Overfitting results in decision trees that are more complex than necessary
- The training error no longer provides a good estimate of how well the tree will perform on previously unseen records

  → Generalization error is very important
- Two approaches to avoid overfitting in decision trees
  - Pre-pruning: Halt tree construction early—do not split a node if this would result in the goodness measure falling below a threshold
    - Difficult to choose an appropriate threshold
  - Post-pruning: Remove decision nodes from a “fully grown” tree—get a sequence of progressively pruned trees
    - Use a set of data different from the training data to decide whether pruning node is effective

Effect of pruning

- How the error in both training and test data evolves with the tree complexity; with and without pruning



Outline

- Classification basics
- Decision tree classifiers
- Overfitting
- Lazy vs Eager Learners
- k-Nearest Neighbors (or learning from your neighbors)
- Evaluation of classifiers

Lazy vs Eager learners

- Eager learners
  - Construct a classification model (based on a training set)
  - Learned models are ready and eager to classify previously unseen instances
  - e.g., decision trees
- Lazy learners
  - Simply store training data and wait until a previously unknown instance arrives
  - No model is constructed.
  - known also as instance based learners, because they store the training set
  - e.g., k-NN classifier

Eager learners

- Do lot of work on training data
- Do less work on classifying new instances

Lazy learners

- Do less work on training data
- Do more work on classifying new instances

Outline

- Classification basics
- Decision tree classifiers
- Overfitting
- Lazy vs Eager Learners
- k-Nearest Neighbors (or learning from your neighbors)
- Evaluation of classifiers