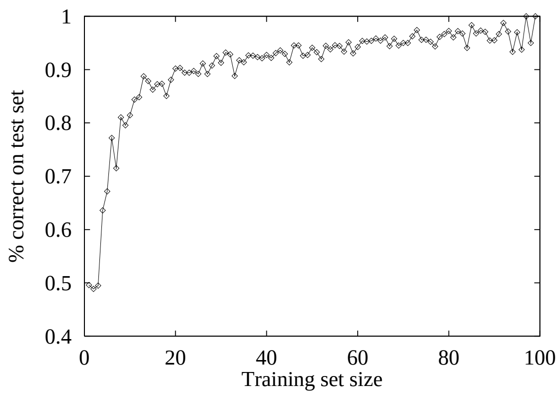
Performance Assessment

- Data sets:
  - Training set:
    - Data used for learning a decision tree.
  - Validation set:
    - Data (unseen examples) available to us for evaluating the generalization performance of the decision tree.
    - Cross-validation
  - Test set:
    - Data (unseen examples) only available during the *real test* of the generalization performance.

  One analogy: The training set is the set of assignment problems, the validation set is the mock examination, and the test set is the real examination. It is the result of the real examination (i.e. test set performance) that counts, not the mock examination result.

- Generalization performance generally improves with training set size:



Measure of Attribute "Importance"

- Intuitively, we want to measure how effective an attribute is in splitting examples into subsets each of which is as "homogeneous" as possible. (A set of examples is "homogeneous" if all the examples belong to the same class.)
- Information-theoretic measure
- Information content:
  - Suppose a random variable has m possible values, $v_i, 1 \leq i \leq m$, with corresponding probabilities of occurrence, $P(v_i), 1 \leq i \leq m$.
  - Information content (in bits) of event corresponding to value v_i:

$$-\log_2 P(v_i)$$

  - Average information content:

$$I(P(v_1), \ldots, P(v_m)) = \sum_{i=1}^{m} -P(v_i) \log_2 P(v_i)$$

  - Examples:
    - For two equally likely outcomes:

$$I(\frac{1}{2}, \frac{1}{2}) = 1 bit$$

    - For two outcomes with probabilities 0.99 and 0.01:

$$I(\frac{99}{100}, \frac{1}{100}) = 0.08 bit$$

Information Gain

- Average information content *before* any attribute test:

$$I(\frac{p}{p+n},\frac{n}{p+n}) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

  assuming that the training set E has p positive examples and n negative examples.

- Average information content *after* a test based on attribute A:

$$Remainder(A) = \sum_{i=1}^{s}\frac{p_i+n_i}{p+n}I(\frac{p_i}{p_i+n_i},\frac{n_i}{p_i+n_i})$$

  assuming that A divides the training set E into subsets $E_1,\ldots,E_s$ and each subset E_i has p_i positive examples and n_i negative examples.

- Information gain:

$$Gain(A) = I(\frac{p}{p+n},\frac{n}{p+n}) - Remainder(A)$$

- The decision tree learning algorithm chooses the attribute that results in the largest information gain.

$$\begin{aligned}\max_A\{Gain(A)\} &= \max_A\left\{I(\frac{p}{p+n},\frac{n}{p+n}) - Remainder(A)\right\}\\ &= \min_A\{Remainder(A)\}\end{aligned}$$

Play/Not-Play Example Revisited

- Training set:

Attributes				Class
Outlook	Temperature	Humidity	Windy	
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

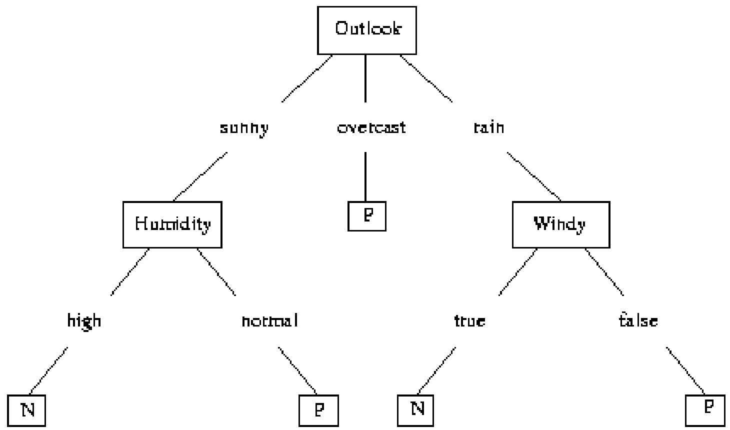
- Choosing the best attribute to split the training set by computing $Remainder(A)$ for each attribute A:
  - Outlook:

$$\frac{5}{14}I(\frac{2}{5},\frac{3}{5}) + \frac{4}{14}I(\frac{4}{4},\frac{0}{4}) + \frac{5}{14}I(\frac{3}{5},\frac{2}{5})$$

  - Similar computation for other attributes (Temperature, Humidity, Windy).

Play/Not-Play Example (cont'd)

- A simple decision tree constructed:



- Not all attributes are necessarily needed in the decision tree.
- Each path from the root to a terminal node can be seen as an IF-THEN rule:
  - Rule 1:
    IF Outlook = sunny AND Humidity = high THEN Class = N
  - Rule 2:
    IF Outlook = sunny AND Humidity = normal THEN Class = P
  - ...

Overfitting

- Noisy training data may lead to overly large decision trees that fit training examples well but unseen examples poorly.
- Pruning some nodes (and hence irrelevant attribute tests) may help to improve generalization performance.
- An issue to address is when to stop the tree pruning process. Cross-validation methods may be used to determine the best stopping point which gives the best generalization performance.