Classification Methods

- Classification is the process of categorizing the data into some known class-labels.
- It is a supervised process since the class labels are known in advance
- Some major classification algorithms are as follows:
  - Decision Tree classifier
  - Nearest Neighbor Classifier
  - Naïve Bayes Classifier
  - Artificial Neural Network (ANN) Based Classifier
  - Support Vector Machine (SVM)
  - Ensemble Based Classifiers

Measures for Performance Evaluation

Accuracy, sensitivity and specificity

Confusion matrix

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Predicted Class</td></tr>
<tr><td>Yes</td><td>No</td></tr>
<tr><td rowspan="2">Actual Class</td><td>Yes</td><td>TP</td><td>FN</td></tr>
<tr><td>No</td><td>FP</td><td>TN</td></tr>
</table>

The true positives (TP) and true negatives (TN) are the correct classifications

A false positive (FP) is when a ‘no’ sample of a class is incorrectly classified as a ‘yes’ sample

A false negative (FN) is when a ‘yes’ sample of a class is classified as ‘no’ sample.

Measures for Performance Evaluation

We define the following measures:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

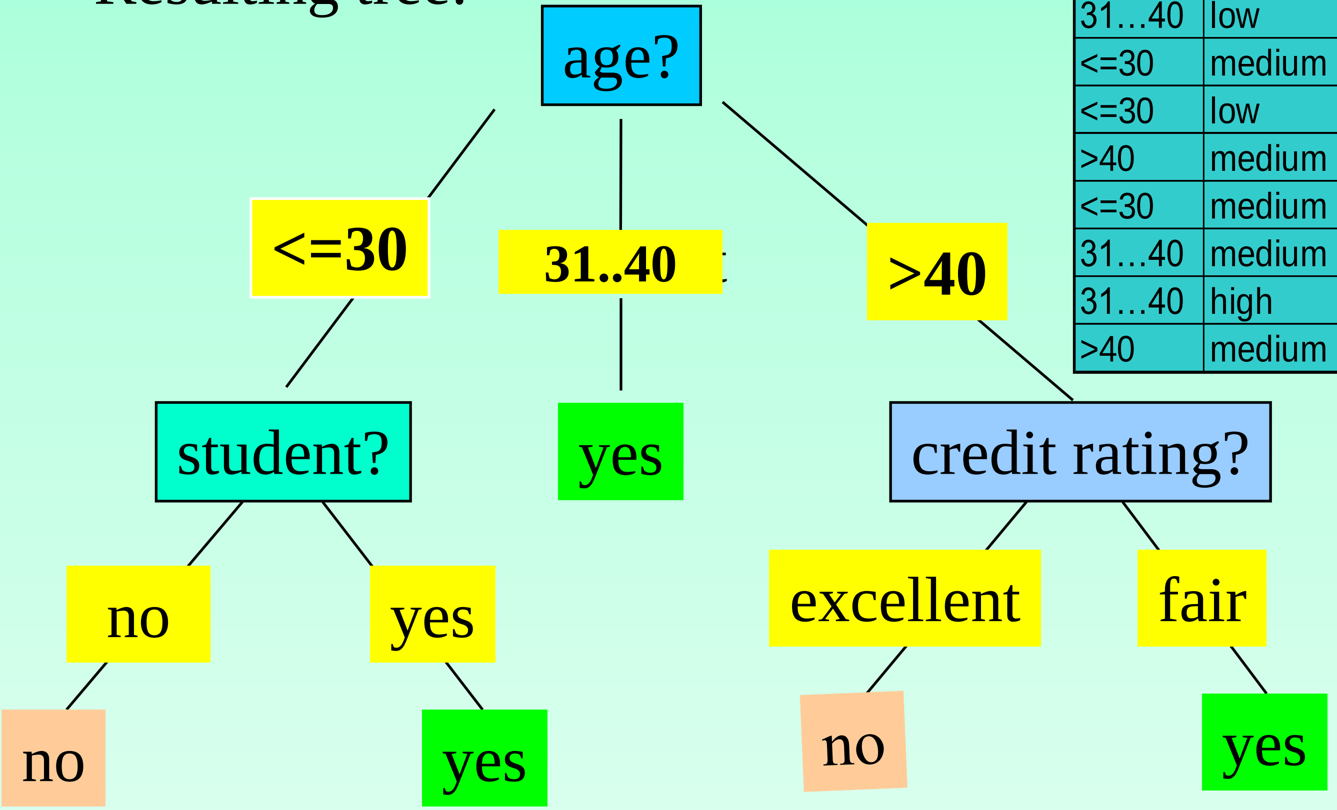
Decision Tree

- A decision tree is a hierarchical structure consisting of nodes and directed edges
- The non leaf node contain attribute test conditions
- Each leaf node is assigned a class label
- Each recursive step of the tree growing process must select an attribute test condition to divide the records into smaller subsets
- Measures used to select the best split are entropy, gain, etc.

Decision Tree Induction: An Example

- Training data set: Buys_computer
- The data set follows an example of Quinlan's ID3 (Playing Tennis)
- Resulting tree:

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



Attribute Selection Measure: Information Gain

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

Attribute Selection: Information Gain

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

$\frac{5}{14}I(2,3)$ means "age <=30" has 5 out of 14 samples, with 2 yes'es and 3 no's. Hence

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

Similarly,

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$