

Data Mining Fundamentals

Chapter 8. Classification: Basic Concepts

Chapter 8. Classification: Basic Concepts

Classification: Basic Concepts



Decision Tree Induction

Bayes Classification Methods

Model Evaluation and Selection

Techniques to Improve Classification Accuracy: Ensemble Methods

Summary

Supervised vs. Unsupervised Learning

□ Supervised learning (classification)

- Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
- New data is classified based on the training set

□ Unsupervised learning (clustering)

- The class labels of training data is unknown
- Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Prediction Problems: Classification vs. Numeric Prediction

□ Classification

- predicts categorical class labels (discrete or nominal)
- classifies data (constructs a model) based on the training set and the values (**class labels**) in a classifying attribute and uses it in classifying new data

□ Numeric Prediction

- models continuous-valued functions, i.e., predicts unknown or missing values

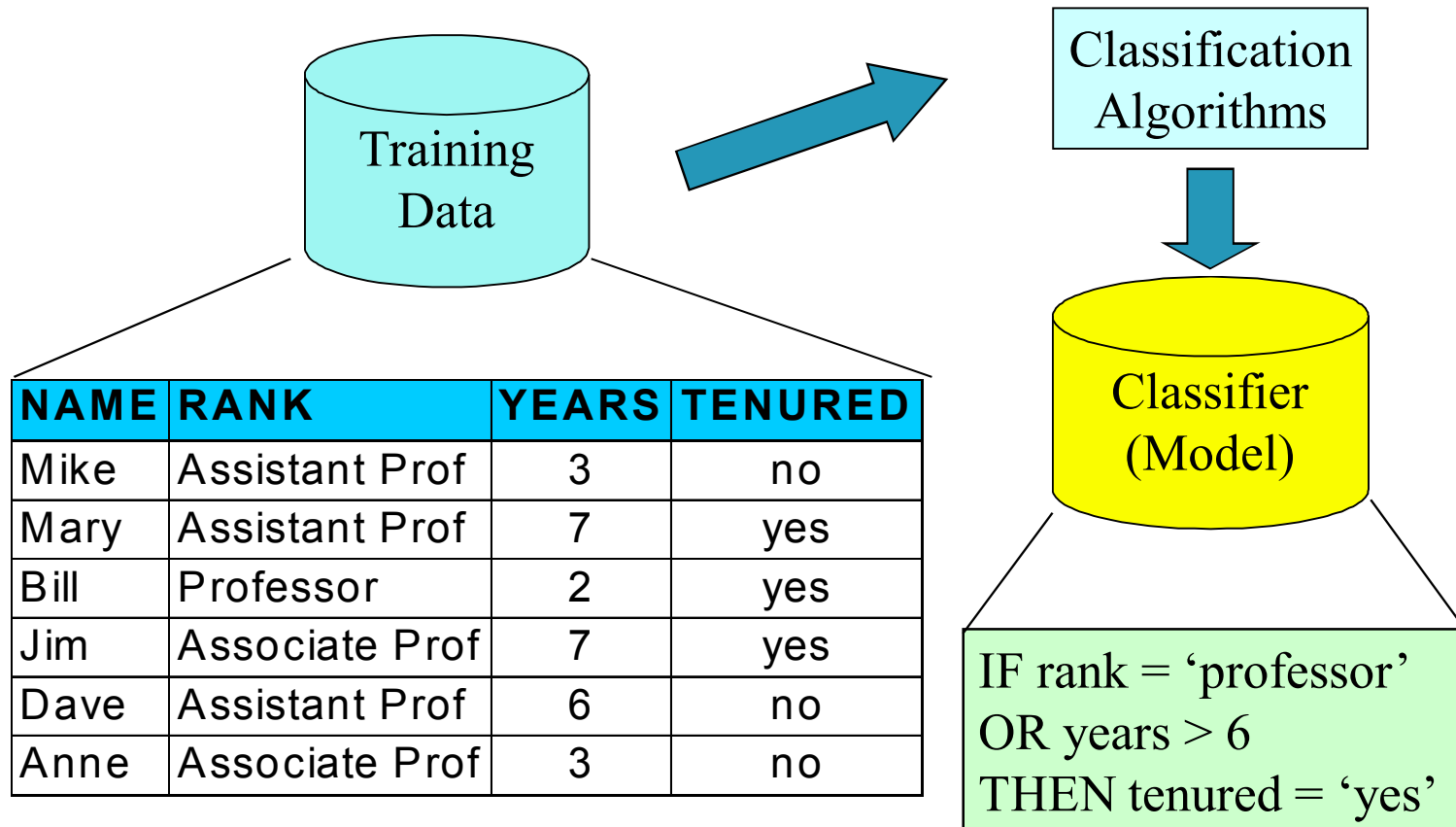
□ Typical applications

- Credit/loan approval:
- Medical diagnosis: if a tumor is cancerous or benign
- Fraud detection: if a transaction is fraudulent
- Web page categorization: which category it is

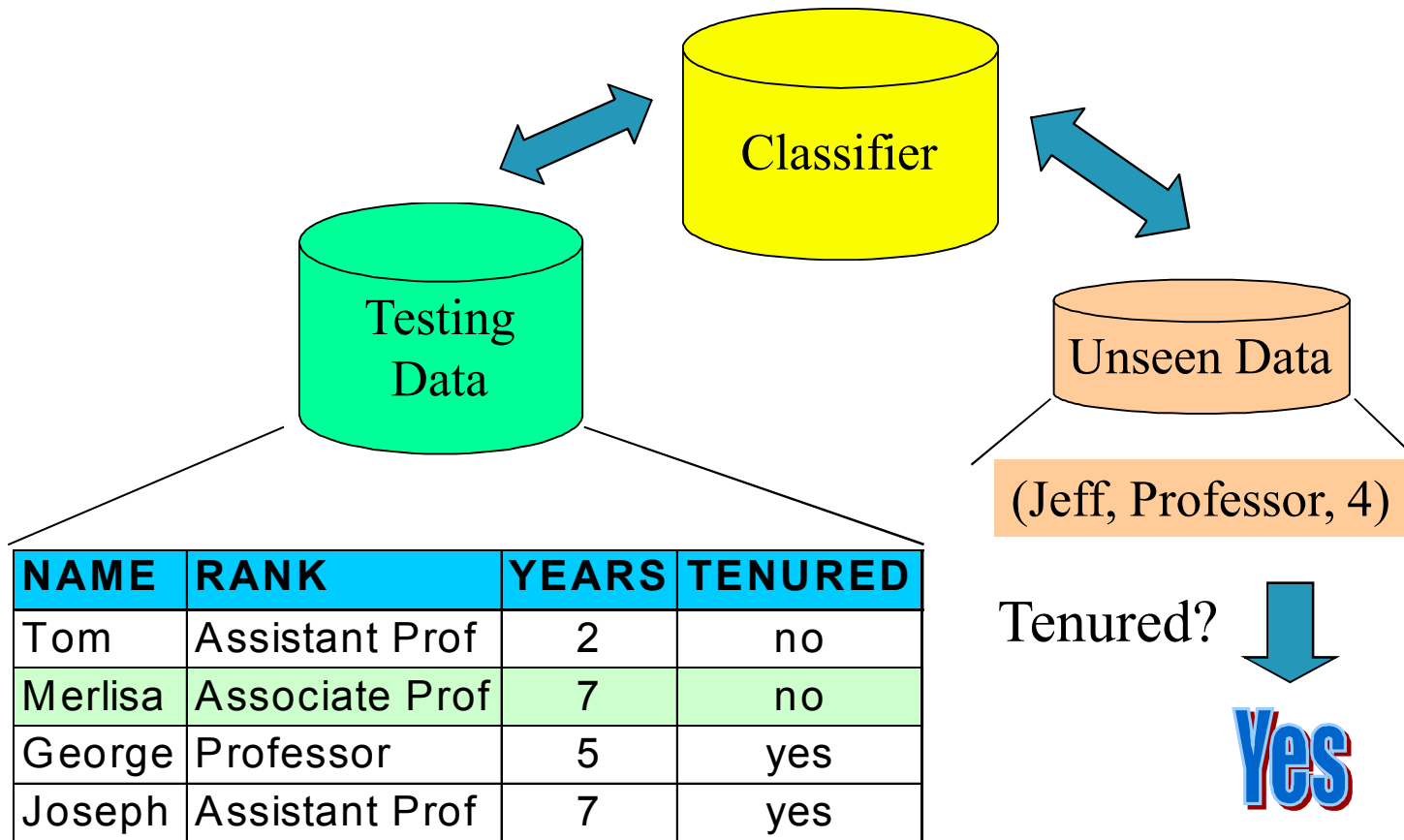
Classification—A Two-Step Process

- ❑ **Model construction:** describing a set of predetermined classes
 - ❑ Each tuple/sample is assumed to belong to a predefined class, as determined by the **class label** attribute
 - ❑ The set of tuples used for model construction is **training set**
 - ❑ Model: represented as classification rules, decision trees, or mathematical formulae
- ❑ **Model usage:** for classifying future or unknown objects
 - ❑ Estimate accuracy of the model
 - ❑ The known label of test sample is compared with the classified result from the model
 - ❑ **Accuracy:** % of test set samples that are correctly classified by the model
 - ❑ Test set is independent of training set (otherwise overfitting)
 - ❑ If the accuracy is acceptable, use the model to classify new data
- ❑ Note: If *the test set* is used to select/refine models, it is called **validation (test) set** or development test set


Process (1): Model Construction



Process (2): Using the Model in Prediction

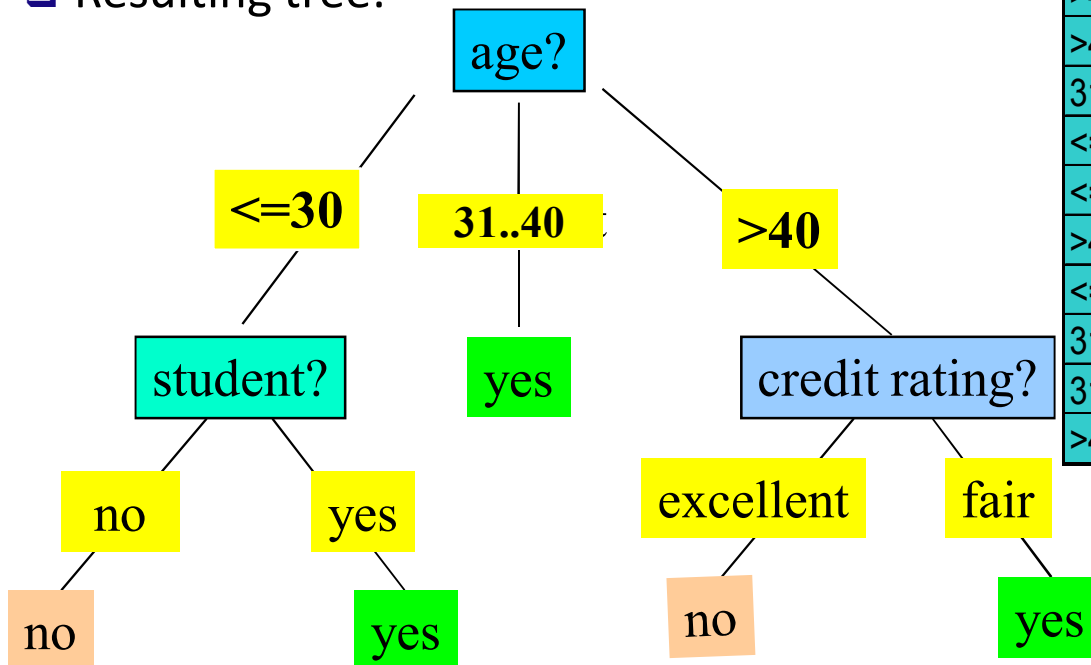


Chapter 8. Classification: Basic Concepts

- ❑ Classification: Basic Concepts
- ❑ Decision Tree Induction 
- ❑ Bayes Classification Methods
- ❑ Model Evaluation and Selection
- ❑ Techniques to Improve Classification Accuracy: Ensemble Methods
- ❑ Summary

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

Algorithm for Decision Tree Induction

- ❑ Basic algorithm (a greedy algorithm)
 - ❑ Tree is constructed in a **top-down recursive divide-and-conquer manner**
 - ❑ At start, all the training examples are at the root
 - ❑ Attributes are categorical (if continuous-valued, they are discretized in advance)
 - ❑ Examples are partitioned recursively based on selected attributes
 - ❑ Test attributes are selected on the basis of a heuristic or statistical measure (e.g., **information gain**)
- ❑ Conditions for stopping partitioning
 - ❑ All samples for a given node belong to the same class
 - ❑ There are no remaining attributes for further partitioning—**majority voting** is employed for classifying the leaf
 - ❑ There are no samples left

Brief Review of Entropy

□ Entropy (Information Theory)

- A measure of uncertainty associated with a random number
- Calculation: For a discrete random variable Y taking m distinct values $\{y_1, y_2, \dots, y_m\}$

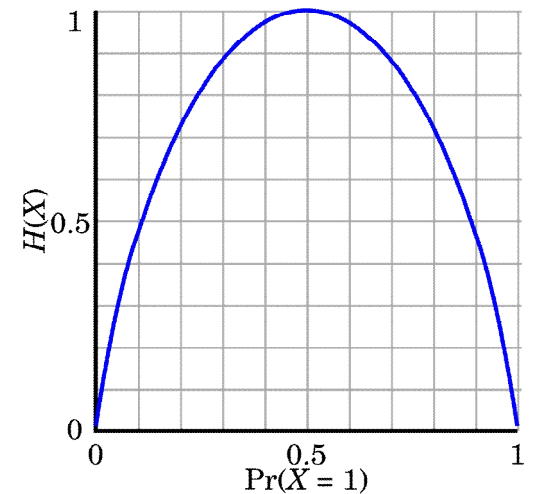
$$H(Y) = - \sum_{i=1}^m p_i \log(p_i) \quad \text{where } p_i = P(Y = y_i)$$

□ Interpretation

- Higher entropy \rightarrow higher uncertainty
- Lower entropy \rightarrow lower uncertainty

□ Conditional entropy

$$H(Y|X) = \sum_x p(x) H(Y|X = x)$$



Attribute Selection Measure: Information Gain (ID3/C4.5)

- ❑ 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,

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit_rating) = 0.048$$

Computing Information-Gain for Continuous-Valued Attributes

- ❑ Let attribute A be a continuous-valued attribute
- ❑ Must determine the *best split point* for A
 - ❑ Sort the value A in increasing order
 - ❑ Typically, the midpoint between each pair of adjacent values is considered as a possible *split point*
 - ❑ $(a_i + a_{i+1})/2$ is the midpoint between the values of a_i and a_{i+1}
 - ❑ The point with the *minimum expected information requirement* for A is selected as the split-point for A
- ❑ Split:
 - ❑ D1 is the set of tuples in D satisfying $A \leq \text{split-point}$, and D2 is the set of tuples in D satisfying $A > \text{split-point}$