Chapter 6. Classification and Prediction

- What is classification? What is prediction?
- Issues regarding classification and prediction
- Classification by decision tree induction
- Bayesian classification
- Rule-based classification
- Associative classification
- Lazy learners (or learning from your neighbors)
- Prediction
- Accuracy and error measures
- Ensemble methods
- Model selection
- Summary

Example: Image Classification

Classification vs. Prediction

- Classification
  - predicts *categorical* class labels (discrete or nominal)
  - constructs a model by learning the training set (having the class labels) and classifies new data by using the model
- Prediction
  - models a continuous-valued functions and predicts unknown or missing values by using the model
- Typical applications
  - Credit approval
  - Target marketing
  - Medical diagnosis
  - Fraud detection

Classification—A Two-Step Process

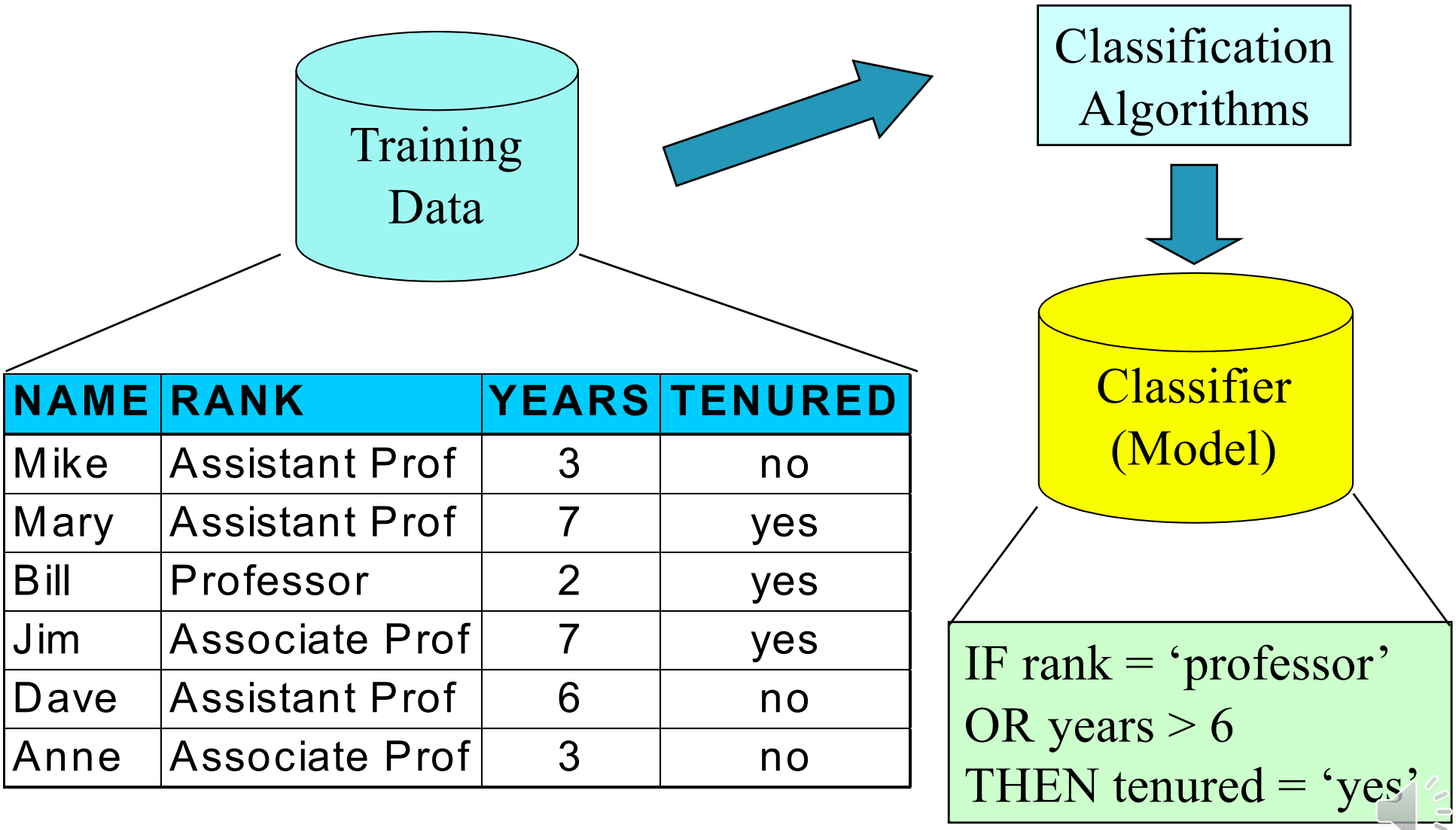
- Model construction
  - Goal: to describe a set of predetermined classes by using a training data
  - Training data
    - A set of tuples/samples used for model construction
    - Each sample/tuple: <attr-1, attr-2, …., attri-n, class label>
    - Each tuple/sample is assumed to belong to a predefined class
  - Model
    - Explains how the attributes of tuples/samples determine the class label
    - Represented as classification rules, decision trees, networks, or mathematical formulae

- Model usage
  - Goal: to classify the future or unknown samples by using the model

Classification—A Two-Step Process

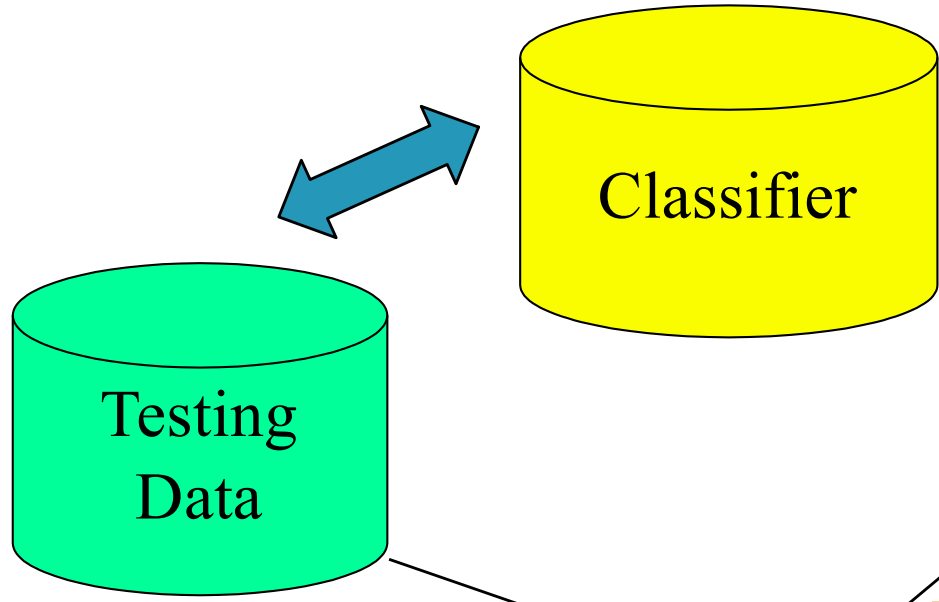
- Accuracy evaluation
  - Goal: to evaluate the accuracy of the model by using a test data
  - Test data
    - A set of tuples/samples used for accuracy evaluation
    - Each sample/tuple: <attr-1, attr-2, …., attri-n, class label>
    - Each tuple/sample has a predefined class
  - The known label of a test sample is compared with the classified result from the model
  - Accuracy rate is the percentage of test set samples that are correctly classified by the model
  - The test set should be independent of the training set; otherwise over-fitting will occur
- If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known

Process (1): Model Construction



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Process (2): Using the Model in Prediction



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

(Jeff, Professor, 4)

Tenured?

Yes

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

Chapter 6. Classification and Prediction

- What is classification? What is prediction?
- Issues regarding classification and prediction
- Classification by decision tree induction
- Bayesian classification
- Rule-based classification
- Classification by back propagation
- Support Vector Machines (SVM)
- Associative classification
- Lazy learners (or learning from your neighbors)
- Other classification methods
- Prediction
- Accuracy and error measures
- Ensemble methods
- Model selection
- Summary

Issues: Data Preparation

- Data cleaning
  - Preprocess data in order to reduce noise and handle missing values
- Relevance analysis (feature selection)
  - Remove the irrelevant or redundant attributes
- Data transformation
  - Generalize and/or normalize data

Issues: Evaluating Classification Methods

- Accuracy
  - classifier accuracy: predicting class label
  - predictor accuracy: guessing value of predicted attributes
- Speed
  - time to construct the model (training time)
  - time to use the model (classification/prediction time)
- Robustness: handling noise and missing values
- Scalability: efficiency in disk-resident databases
- Interpretability
  - understanding and insight provided by the model
- Other measures, e.g., goodness of rules, such as decision tree size or compactness of classification rules

Chapter 6. Classification and Prediction

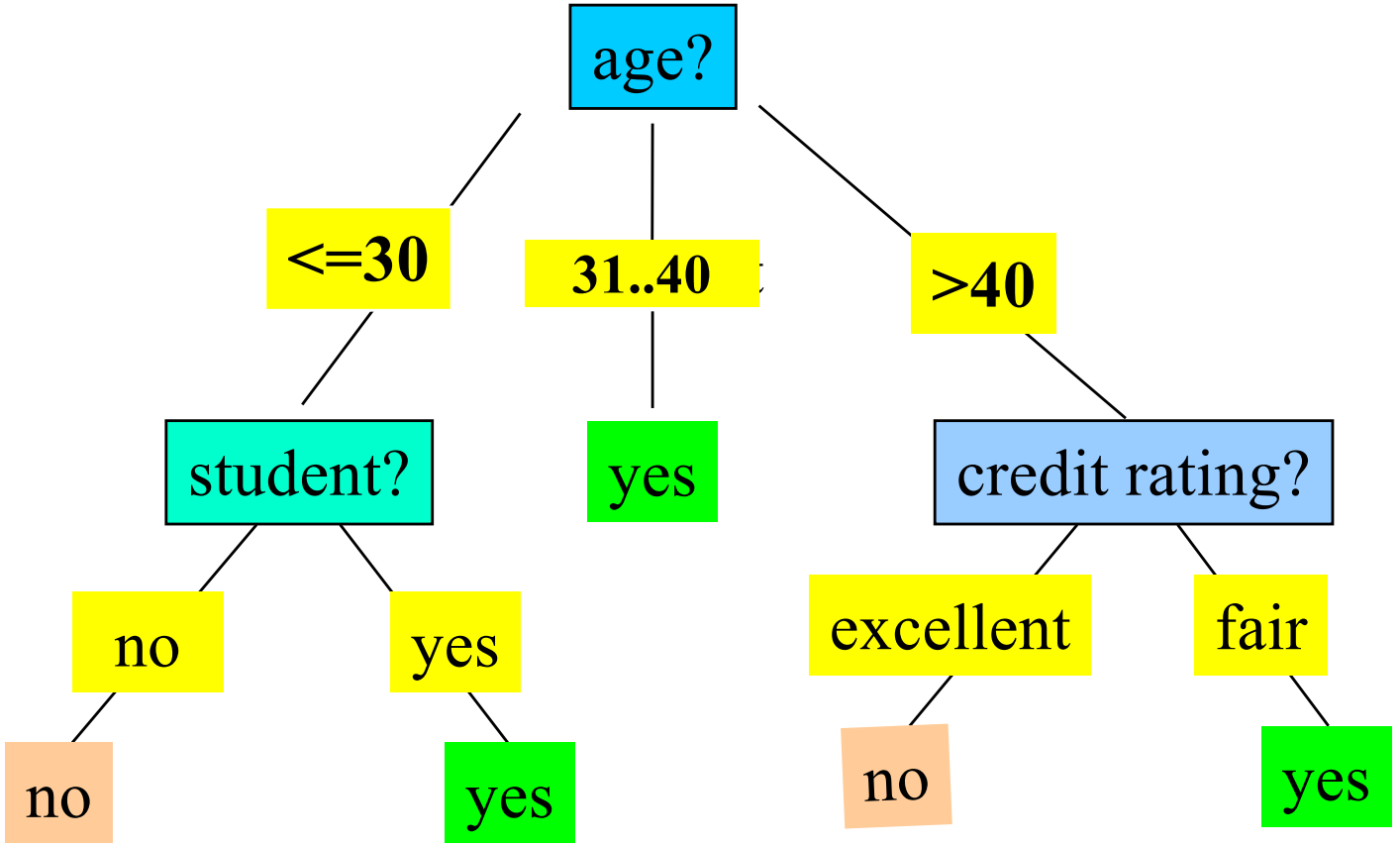
- What is classification? What is prediction?
- Issues regarding classification and prediction
- Classification by decision tree induction
- Bayesian classification
- Rule-based classification
- Classification by back propagation
- Support Vector Machines (SVM)
- Associative classification
- Lazy learners (or learning from your neighbors)
- Other classification methods
- Prediction
- Accuracy and error measures
- Ensemble methods
- Model selection
- Summary

Decision Tree Induction: Training Dataset

Class label

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

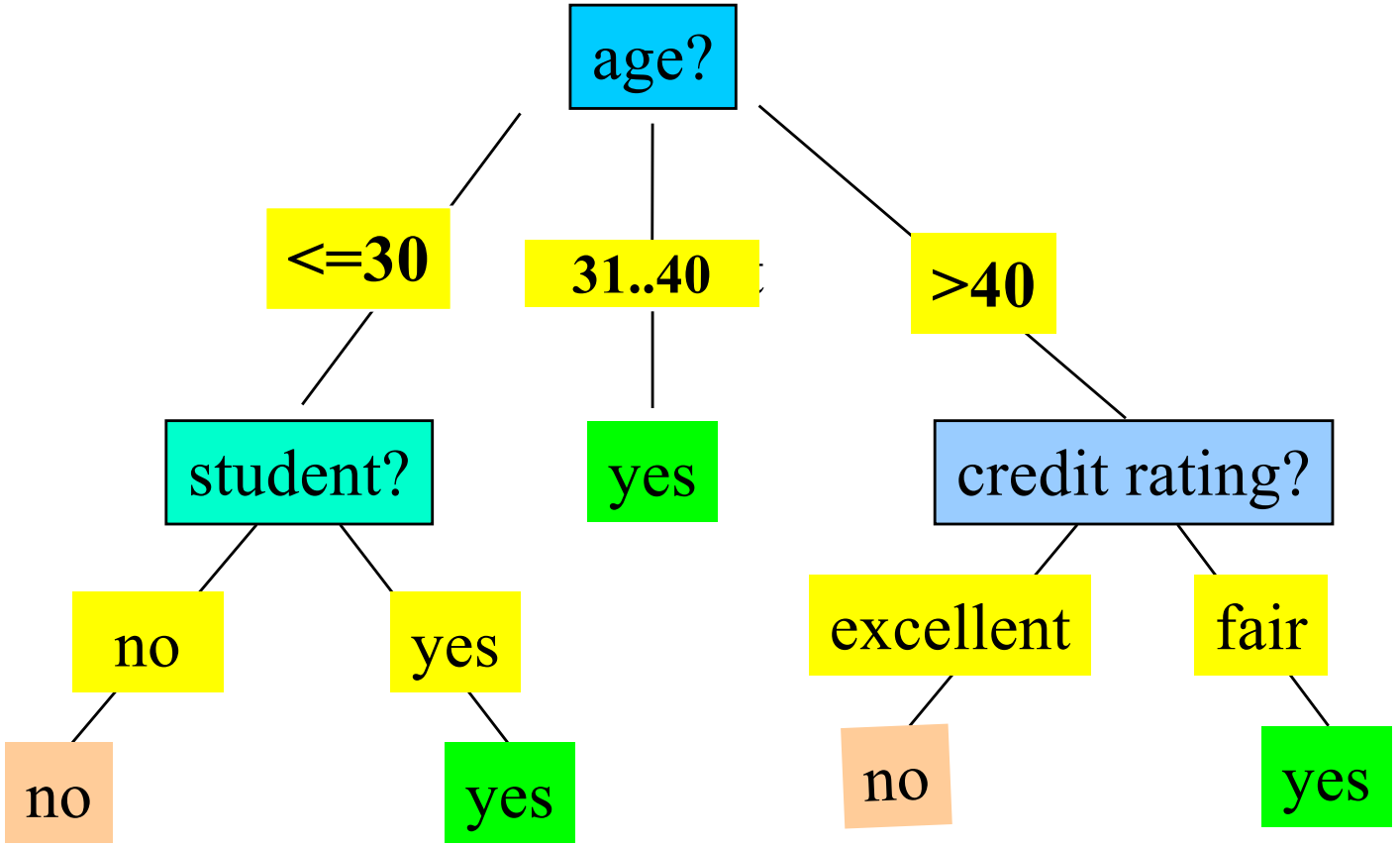
Output: A Decision Tree for "*buys_computer*"



Algorithm for Decision Tree Induction

- Basic algorithm
  - A greedy algorithm that constructs a decision tree in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are assumed to be categorical
    - If continuous-valued, they are discretized in advance
  - Examples are partitioned recursively based on the selected test attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

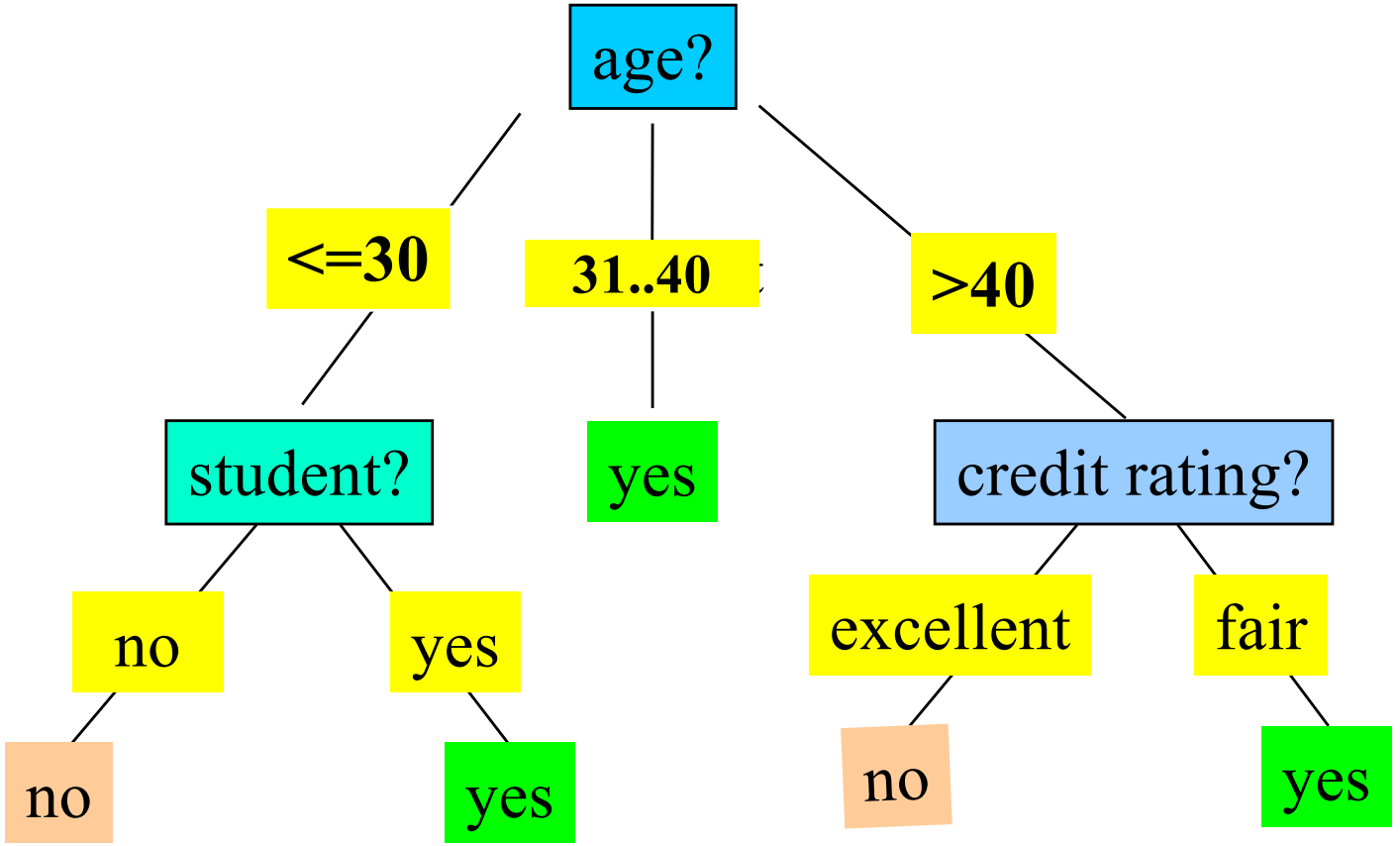
Output: A Decision Tree for "*buys_computer*"



Algorithm for Decision Tree Induction

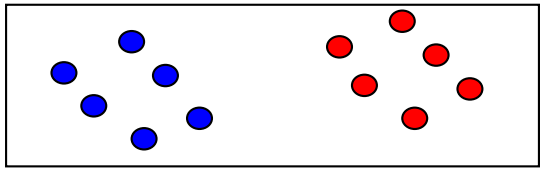
- Conditions for stopping the partitioning process
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
  - There are no samples left

Output: A Decision Tree for "*buys_computer*"

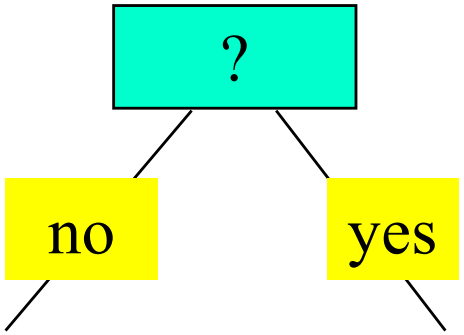


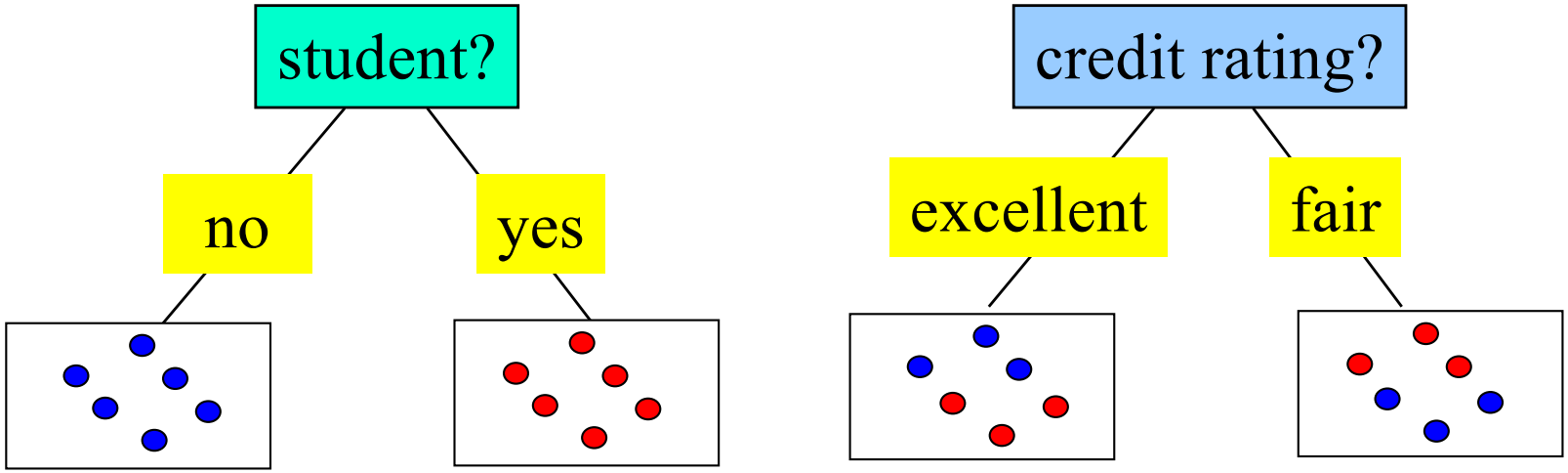
Test Attribute Selection

- Which is better as a test attribute?
  - Partitions a group into more homogeneous ones



Buy-computer: yes
Buy-computer: no





Attribute Selection Measure: Information Gain

- Select the test attribute having the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,D}|/|D|$
- Entropy (expected information) to classify a tuple in D:
  - $$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$
  - The more heterogeneous the higher