Lesson 6 Bayes' Rule

Outline

- Supervised vs. Unsupervised Learning
- Classification: a two-step process
- Introduction to Bayes’ Rule
- Using Bayes’ Rule for Classification

Supervised vs. Unsupervised Learning

- Supervised learning (e.g. classification)
  - Supervision: The data (observations, measurements, etc.) are accompanied by **class labels** indicating the class of the observations
  - New data is classified based on the model
- Unsupervised learning (e.g. clustering)
  - The class labels of data are unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Applications of Classification

- Classification is a data mining task that assigns data samples to target categories or classes. The goal of classification is to accurately predict the target class for each data samples.
- Typical applications
  - Credit/loan approval:
  - Medical diagnosis: if a tumor is cancerous or benign
  - Fraud detection: if a transaction is fraudulent

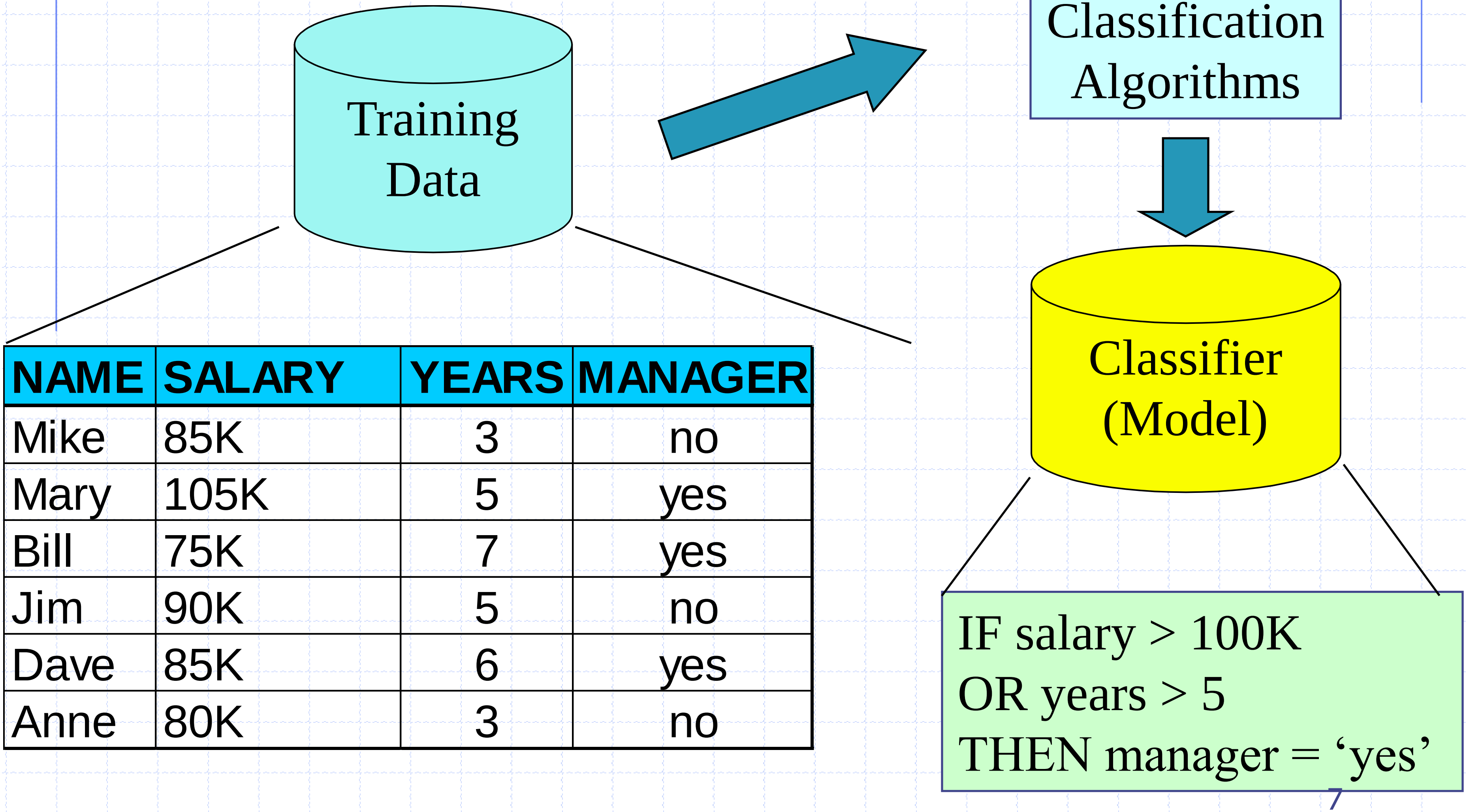
Classification - A Two-Step Process (1)

- Model construction: describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class attribute
  - The set of tuples used for model construction is the training set
  - The model is represented as classification rules, decision trees, mathematical formulae, etc.

Classification - A Two-Step Process (2)

- Model usage: classify future or unknown tuples
  - Estimate accuracy of the model
    - The known label of 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
    - Test set is independent of training set
  - If the accuracy is acceptable, use the model to classify data tuples with unknown class labels. Otherwise, build another model and repeat.

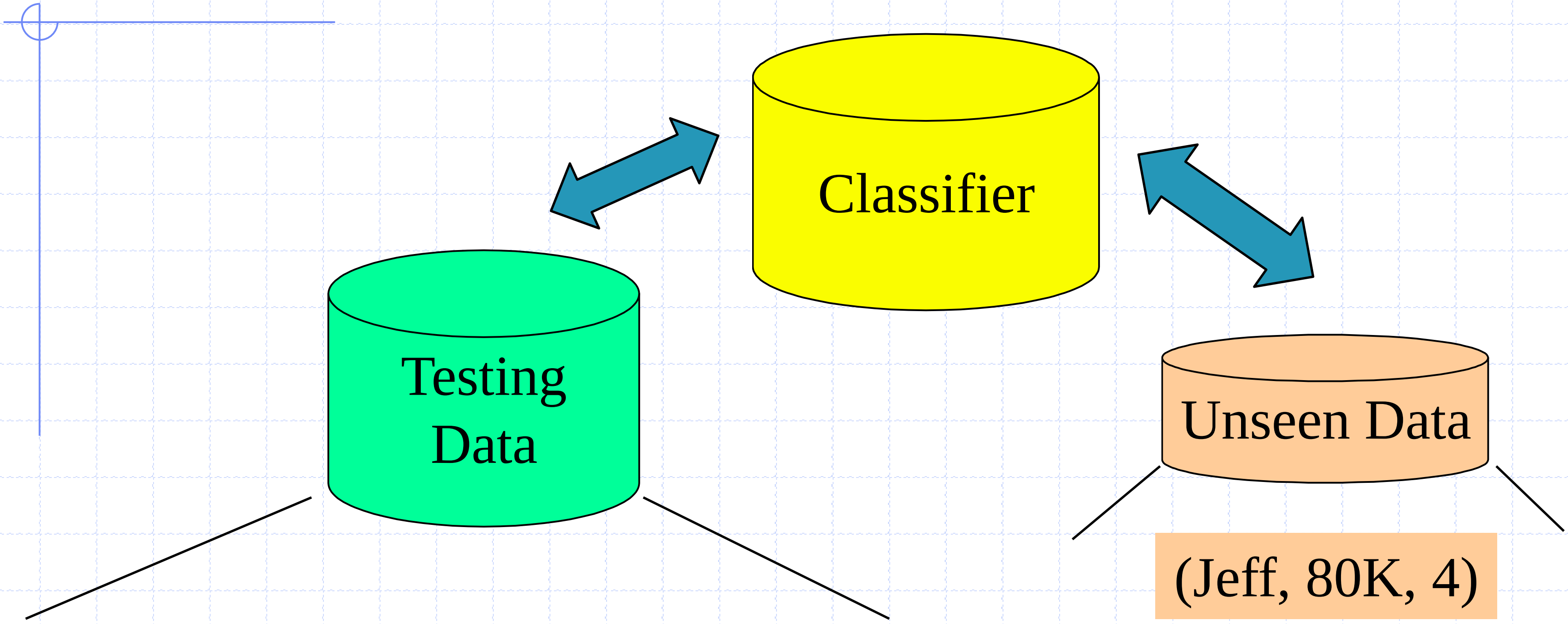
Process (1): Model Construction



NAME	SALARY	YEARS	MANAGER
Mike	85K	3	no
Mary	105K	5	yes
Bill	75K	7	yes
Jim	90K	5	no
Dave	85K	6	yes
Anne	80K	3	no

IF salary > 100K
OR years > 5
THEN manager = ‘yes’

Process (2): Model Usage



(Jeff, 80K, 4)

NAME	SALARY	YEARS	MANAGER
Tom	90K	2	no
Merlisa	75K	4	no
George	95K	6	yes
Joseph	120K	5	yes

Manager?

No

Training Dataset

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

Bayesian Classification

- A statistical classifier: performs probabilistic prediction, i.e., predicts class membership probabilities.
- Foundation: Based on Bayes' Theorem.
- Performance: A simple Bayesian classifier, naïve Bayesian classifier, has comparable performance with decision tree and selected neural network classifiers.

Review of Probabilities

- **Probability** is the measure of the likelihood that an event will occur. Probability is quantified as a number between 0 and 1 (where 0 indicates impossibility and 1 indicates certainty).
- **Conditional probability** measures the probability of an event given that another event has occurred. If the event of interest is X and the event H is known or assumed to have occurred, "the conditional probability of X given H", or "the probability of X under the condition H", is usually written as $P(X|H)$.

Examples

- Simple example of probability
  - Event A: Die 1 lands on 3
  - Event B: Die 2 lands on 1
  - Event C: The two dice sum to 8

- Exercise: Compute P(A), P(B), P(C), P(A|C) and P(C|A)
  - The prior probability of A, P(A), is 1/6. The prior probability of B, P(B), is 1/6. The prior probability of C, P(C), is 5/36.
  - The conditional probability of event C given that A occurs, P(C|A), is 1/6
  - Another example: P(A|C), is 1/5

Bayes' Theorem

- Bayes’ rule (or Bayes’ theorem) of conditional probability:

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}$$

- Bayes’ Rule shows the relation between one conditional probability and its inverse.
- The posterior probability P(H | X) is equal to the conditional probability of event X given H multiplied by the prior probability of H, all divided by the prior probability of X.

Example of Bayes' Rule

Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on the day of Marie's wedding?

Define:

- Event A_1: It rains.
- Event A_2: It does not rain.
- Event B: The weatherman predicts rain.

Question: What do we need to compute to solve the problem?

Answer: $P(A_1|B)$

Example of Bayes' Rule

Solution:

$P(B|A_1) = 0.9$

$P(A_1) = 5/365$

$P(B) = P(B|A_1)*P(A_1) + P(B|A_2)*P(A_2)$

$= 0.9*5/365 + 0.1*360/365$

{By Math Review Lecture:

$P(B) = P(B \mid A) * P(A) + P(B \mid -A) * P(-A)$}

Using Bayes Rule:

$P(A_1|B) = P(B|A_1) * P(A_1)/P(B)$

$= 0.9*(5/365)/(0.9*5/365 + 0.1*360/365)$

$= 0.111$

Towards Naïve Bayesian Classifier

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-(attribute value) vector **X** = (x_1, x_2, …, x_n), where, for each *i*, x_i is an attribute value of A_i.
- Suppose there are *m* classes C_1, C_2, …, C_m.
- Classification is to derive the maximum posterior probability. In other words, compute P(C_1|**X**), P(C_2|**X**), …, P(C_m|**X**) and predict **X** belongs to the class with the highest probability.

Towards Naïve Bayesian Classifier

- $P(C_i|\mathbf{X})$ can be computed using Bayes' theorem

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Buys_computer dataset example: Given that we know properties of a (new) customer, need to compute the probability that the customer will buy a computer (C_1: buys_computer = yes) and the probability that the customer will not buy a computer (C_2 : *buys_computer = no*).
  - In other words, need to compute $P(C_1|X)$ and $P(C_2|X)$

Towards Naïve Bayesian Classifier

- By Bayes' Rule:

$$P(C_1|X) = P(X|C_1) * P(C_1) / P(X)$$
$$P(C_2|X) = P(X|C_2) * P(C_2) / P(X)$$

- And, assign the class with higher probability to X (or we predict the class label of X will be the class with higher probability).

- Notice P(X) is the same, so we compute and compare only the numerators.

Derivation of Naïve Bayes Classifier

- $P(C_i)$ can be easily estimated from the training dataset
  - $P(C_1)$ = (# yes tuples) / (total # tuples)
  - $P(C_2)$ = (# no tuples) / (total # tuples)
- Computation of $P(X|C_i)$ is not easy.
- It can be simplified with class-conditional independence assumption (a naïve assumption).
- class-conditional independence assumption: attributes are conditionally independent (i.e., no dependence relation between attributes).

Derivation of Naïve Bayes Classifier

- Based on this assumption, we compute $P(x_k | C_i)$ for each attribute x_i, and multiply them all to obtain $P(X | C_i)$ (this is possible because we assumed all attributes are independent of each other).

- If A_k is categorical, $P(x_k | C_i)$ is the # of tuples in C_i having value x_k for A_k divided by $|C_{i, D}|$ (# of tuples of C_i in D)

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k | C_i) = P(x_1 | C_i) \times P(x_2 | C_i) \times \ldots \times P(x_n | C_i)$$

(Example follows in the next slides)

Worked Example

Class:
C1:buys_computer = 'yes'
C2:buys_computer = 'no'

Data sample
X = (age >40,
Income = high,
Student = no
Credit_rating = excellent)

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

Worked Example

- Class prior probabilities are:

  P(C_1) = P(buys_computer = “yes”) = 9/14 = 0.643

  P(C_2) = P(buys_computer = “no”) = 5/14= 0.357

- Next, we compute *P*(*X*|*C*$_i$) for each class

  P(age = “>40” | buys_computer = “yes”) = 3/9 = 0.333

  P(income = “high” | buys_computer = “yes”) = 2/9 = 0.222

  P(student = “no” | buys_computer = “yes) = 3/9 = 0.333

  P(credit_rating = “excellent” | buys_computer = “yes”) = 3/9 = 0.333

  P(X|buys_computer = “yes”) = 0.333 x 0.222 x 0.333 x 0.333 = 0.008

  P(age = “>40” | buys_computer = “no”) = 2/5 = 0.4

  P(income = “high” | buys_computer = “no”) = 2/5 = 0.4

  P(student = “no” | buys_computer = “no”) = 4/5 = 0.8

  P(credit_rating = “excellent” | buys_computer = “no”) = 3/5 = 0.6

  P(X|buys_computer = “no”) = 0.4 x 0.4 x 0.8 x 0.6 = 0.077

Worked Example

$P(X|C_i)$:

P(X|buys_computer = "yes") = 0.008

P(X|buys_computer = "no") = 0.077

$P(X|C_i)*P(C_i)$:

$P(X|C_1)*P(C_1)$

= P(X|buys_computer = "yes") * P(buys_computer = "yes")

= 0.008 * 0.643

= 0.005

$P(X|C_2)*P(C_2)$

=P(X|buys_computer = "no") * P(buys_computer = "no")

= 0.077 * 0.357

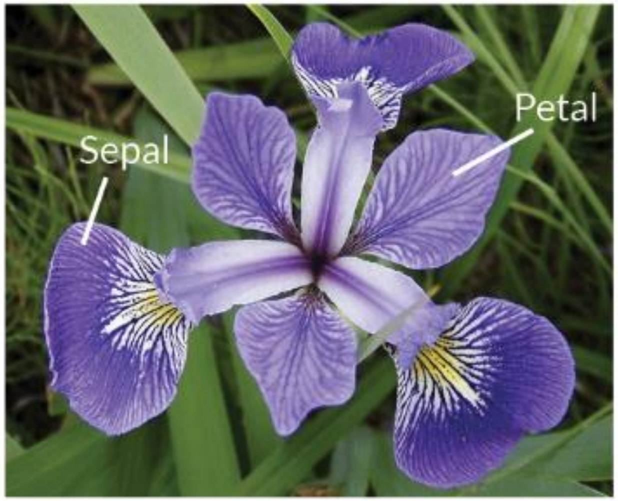
= 0.027

Therefore, the model predicts that X belongs to class ("buys_computer = no")

Naive Bayes For Numeric Attributes

- The algorithm is designed to work primarily with categorical attributes. However, in many practical situations, variables or attributes are seldom categorical. For example, income variable from earlier example might be numeric.
- One common solution is to discretize numeric variables - discretize the continuous variables into a few categories. However doing so is sometimes subjective. For instance, in categorizing income, someone may select 100k as the cutoff at which income can be considered as "High", whereas another person may choose to select 150k. This subjectivity causes obvious loss of information. But it can still be used as a quick way to get going before applying Naive Bayes classification.

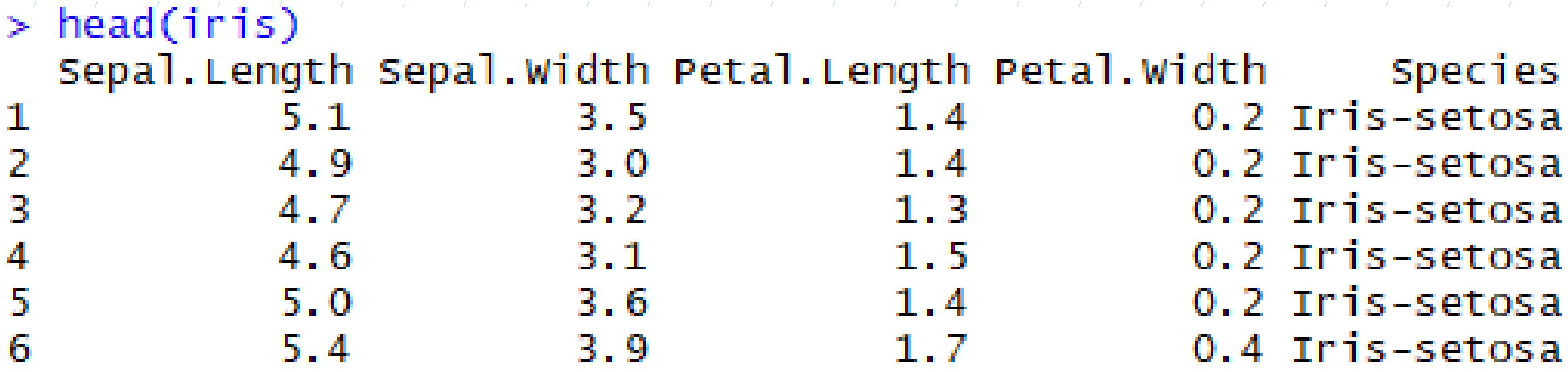
Iris Example and R code



Iris Versicolor

Iris Setosa

Iris Virginica



```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width     Species
1          5.1         3.5          1.4         0.2 Iris-setosa
2          4.9         3.0          1.4         0.2 Iris-setosa
3          4.7         3.2          1.3         0.2 Iris-setosa
4          4.6         3.1          1.5         0.2 Iris-setosa
5          5.0         3.6          1.4         0.2 Iris-setosa
6          5.4         3.9          1.7         0.4 Iris-setosa
```

Iris Example and R code

```
#install.packages("naivebayes")
library("naivebayes")

#loads the data set
data(iris)

nb <- naive_bayes(Species ~ ., data = iris)

#using the naive bayes' model for prediction on the training data
predict(nb, iris[,-5])

#classification result
table(predict(nb, iris[,-5]), iris[,5])
```

Weka: Data Mining Software in Java

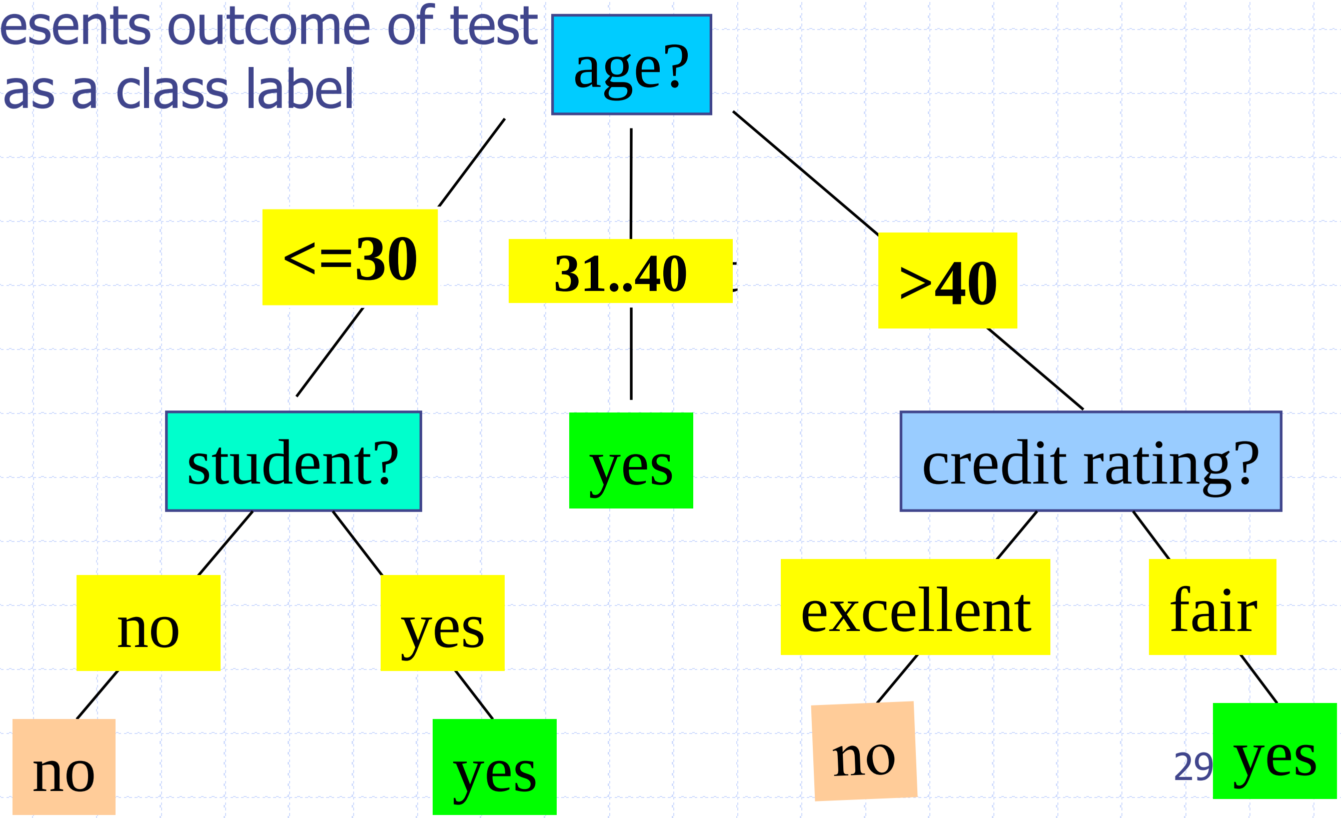
- Open source software.
- Weka contains a collection of tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions.
- Weka supports several standard data mining tasks: data preprocessing, clustering, classification, regression etc.
- Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature.
- Demo of Weka

Summary

- Advantages
  - Easy to implement
  - Good results obtained in most of the cases
- Disadvantages
  - Assumption: class conditional independence, therefore loss of accuracy
  - Practically, dependencies exist among variables:
    - patient_profile: age, height, weight, etc.
    - Dependencies among these cannot be modeled by Naïve Bayesian Classifie
    - How to deal with these dependencies? Optional: Bayesian Belief Networks

Leading to Next Topic

- We will look at another popular classifier – decision trees tomorrow.
- A model is built as a decision tree
- Internal node represents a test on an attribute
- Branch represents outcome of test
- Leaf node has a class label


age?
<=30
31..40
>40
student?
yes
credit rating?
no
yes
excellent
fair
no
yes
no
yes