# Artificial Intelligence & Machine Learning and Pattern Recognition ——Decision Trees 1


Yanghui Rao
Assistant Prof., Ph.D
School of Mobile Information Engineering,
Sun Yat-sen University
raoyangh@mail.sysu.edu.cn

# Classification

- Predict discrete class labels
  - classify objects (construct a model) based on the training set and the class labels in a classifying attribute and then use the rules to classify new objects.
- Typical applications
  - Target marketing (电子商务)
  - Credit approval (银行/金融)
  - Medical diagnosis (健康医疗)
  - Fraud/Intrusion detection (互联网)

# A Two Step Process

- **Model construction**: describing a set of predetermined classes (类别)
  - Each object/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The set of objects/samples used for model construction is training set (训练数据集)
  - The constructed model can be represented as classification rules, decision trees, or mathematical formula (kNN, NB, …)

# A Two Step Process

- **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 rate is the percentage of test set samples that are correctly classified by the model; under-fitting and over-fitting (过拟合)
  - If the *accuracy* (多评测指标) is acceptable, use the model to classify objects whose class labels are not known (用于测试数据)

# Evaluation Metrics

- Accuracy
- Speed
  - time to construct the model (training time)
  - time to use the model (prediction time)
- Robustness
  - handling noise and missing values
- Scalability
  - efficiency in disk-resident databases
- Interpretability

# Evaluation Metrics

# Evaluation Metrics


# Decision Tree

- A flow-chart-like tree structure
- Internal node（中间节点）denotes a splitting test on an attribute
- Branch (分支) represents an outcome of the test (试验)
- Leaf nodes represent class distribution

# Decision 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 |

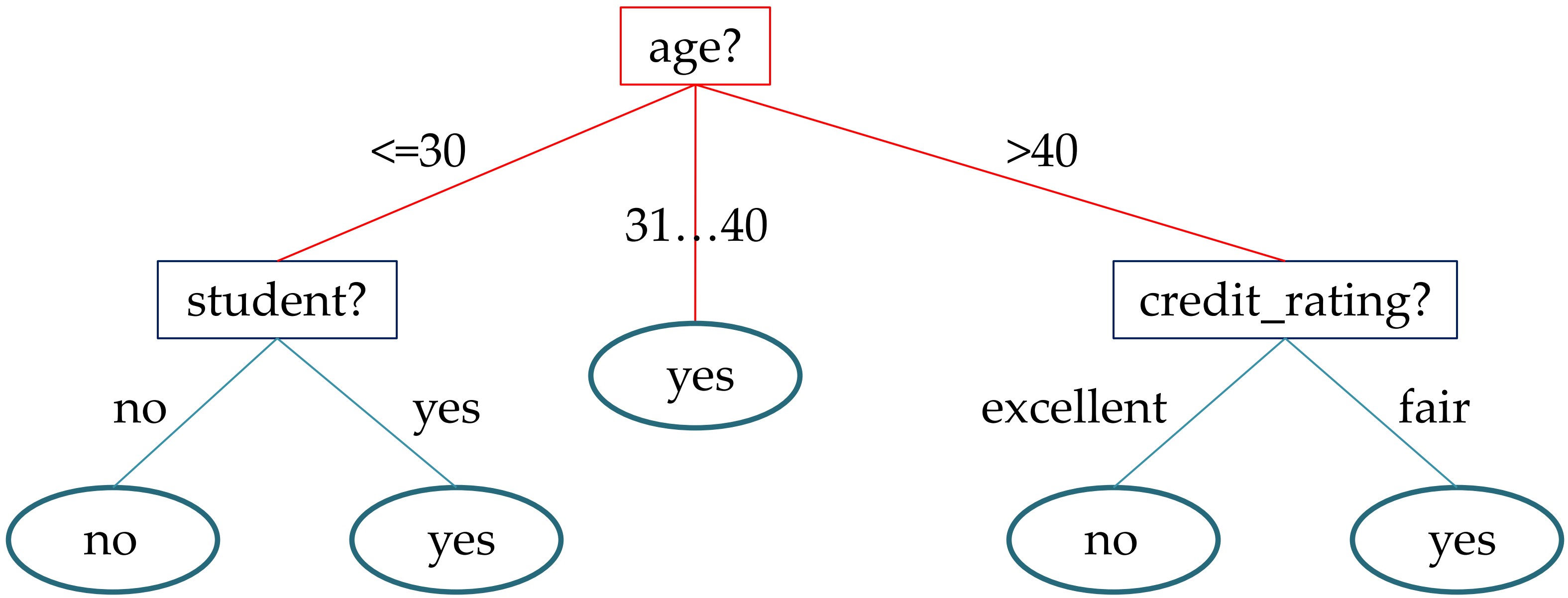
# Decision Tree

| 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 |

# Decision 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 |

# Decision Tree



# Decision Tree

- Decision tree generation: two phases
  - Tree construction (建树)
    - At first, all the training examples are at the root
    - Partition examples recursively (迭代地) based on selected attributes
  - Tree pruning (剪枝)
    - Identify and remove branches that reflect noise or outliers
- Usage of decision trees: Classifying an unknown sample

# Algorithm for Decision Tree

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down (自顶向下) recursive divide-and-conquer manner
  - At first, all training samples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - (训练) samples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic (启发式) or statistical measure (e.g., **information gain**, **Gini index**)

# Algorithm for Decision Tree

- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes (无属性) for further partitioning - majority voting is employed for classifying the leaf
  - There are no samples left (无训练数据)

# Algorithm for Decision Tree

- How to determine the "*importance*" of each attribute?

| 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 |

# Information theory

- Suppose you are reporting the results of rolling an 8-sided die. How many bits are needed?

# Information theory

- Suppose you are reporting the results of rolling an 8-sided die. How many bits are needed?

$$3bits = \log_2 8 = -\sum_{i=1}^{8} \frac{1}{8} \log_2 \frac{1}{8} = -\sum_{i=1}^{8} p(i) \log_2 p(i) = H(X)$$

# Information theory

- Suppose you are reporting the results of rolling an 8-sided die. How many bits are needed?

$$3bits = \log_2 8 = -\sum_{i=1}^{8} \frac{1}{8} \log_2 \frac{1}{8} = -\sum_{i=1}^{8} p(i) \log_2 p(i) = H(X)$$

- If we wish to send the result of rolling an eight-sided die, the most efficient way is to simply encode the result as a 3 digit binary message: 000 - 111

# Information theory

- Entropy (熵)
  - represent the expectation of uncertainty for a random variable (用来衡量离散变量的不确定性，如抛硬币、掷骰子)

$$H(X) = -\sum_{x\in X} p(x)\log_2 p(x)$$

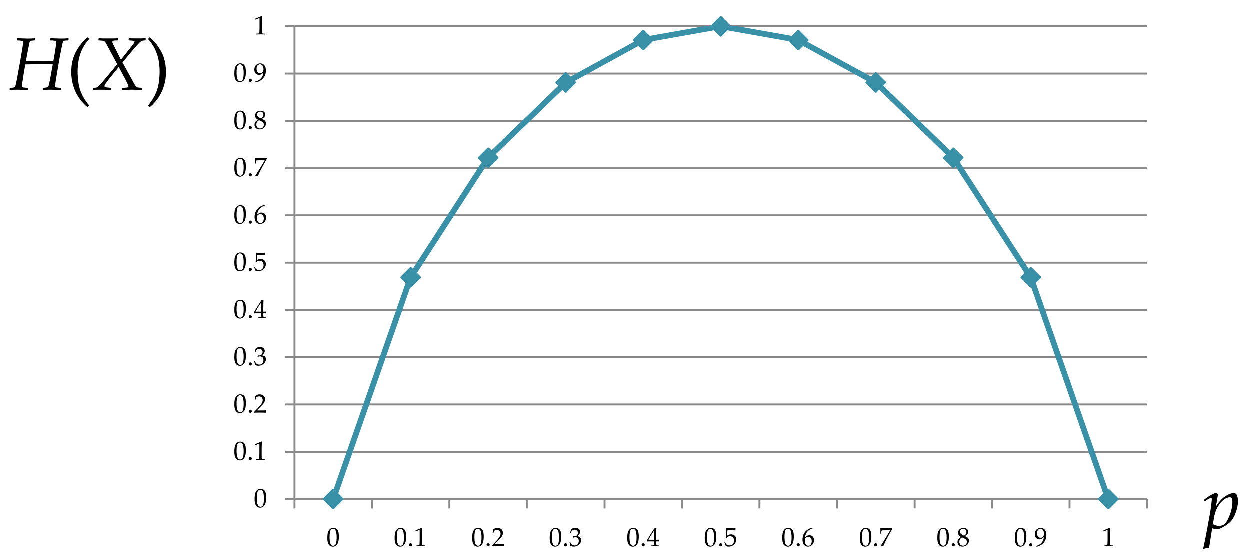
$$= \sum_{x\in X} p(x)\log_2 \frac{1}{p(x)}$$

$$= E\left( \log_2 \frac{1}{p(X)} \right)$$

# Information theory

- P($X$=1) = $p$, P($X$=0) = 1-$p$
  - 假设抛一枚硬币，正面朝上的概率为$p$，反面朝上的概率为1-$p$，则抛这枚硬币所得结果的不确定性（熵值）是$p$的下述函数：

$$H(X) = -p\log_2 p - (1-p)\log_2(1-p)$$



# Information theory

- Conditional/joint entropy

条件熵：
$$H(Y|X)=\sum_{x\in X}p(x)H(Y|X=x)$$
$$=\sum_{x\in X}p(x)\left[-\sum_{y\in Y}p(y|x)\log_2 p(y|x)\right]$$
$$=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log_2 p(y|x)$$
$$=-\sum_{x\in X}\sum_{y\in Y}p(x)p(y|x)\log_2 p(y|x)$$

联合熵：
$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log_2 p(x,y)$$

# Information theory

$$H(X,Y) = -E_{p(x,y)} \log_2 p(x,y)$$

$$= -E_{p(x,y)} \left( \log_2 ( p(x) p(y \mid x) ) \right)$$

$$= -E_{p(x,y)} \left( \log_2 p(x) + \log_2 p(y \mid x) \right)$$

$$= -E_{p(x)} \log_2 p(x) - E_{p(x,y)} \log_2 p(y \mid x)$$

$$= H(X) + H(Y \mid X)$$

两个离散变量$X$和$Y$的联合熵（即，联合出现的不确定性）
= $X$的熵 + 给定$X$，出现$Y$的条件熵
= $X$的不确定性 + 给定$X$，出现$Y$的不确定性

# Information theory

- Mutual information (互信息)

因为：$H(X,Y)=H(X)+H(Y\mid X)=H(Y)+H(X\mid Y)$

所以：$H(Y)-H(Y\mid X)=H(X)-H(X\mid Y)=I(X;Y)$

两个离散变量$X$和$Y$的互信息$I(X;Y)$
衡量的是这两个变量之间的相关度

一个连续变量$X$的不确定性，用方差$Var(X)$来度量
一个离散变量$X$的不确定性，用熵$H(X)$来度量
两个连续变量$X$和$Y$的相关度，用协方差或相关系数来度量
两个离散变量$X$和$Y$的相关度，用互信息$I(X;Y)$来度量

# Information Gain (ID3)

- Class label: buy_computer="yes/no"
- 用字母$D$表示类标签，字母$A$表示每个属性
- $H(D)$=0.940

14个训练样本中，9个买了电脑

$$H(D) = -\frac{9}{14}\log_2\frac{9}{14} - (1-\frac{9}{14})\log_2(1-\frac{9}{14})$$

- $H(D|A="age")$=0.694

$$H(D|A="age") = \frac{5}{14}\times\left(-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}\right)$$

$$+\frac{4}{14}\times\left(-\frac{4}{4}\log_2\frac{4}{4}-\frac{0}{4}\log_2\frac{0}{4}\right)+\frac{5}{14}\times\left(-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}\right)$$

# Information Gain (ID3)

- Compute the mutual information between $D$ (类标签) and each attribute $A$ (每个属性)
- $H(D)$=0.940
- $H(D|A=\text{"age"})$=0.694

$$g(D,A)=I(D;A)=H(D)-H(D|A)$$

- $g(D,A=\text{"age"})$=0.246
- $g(D,A=\text{"income"})$=?
- $g(D,A=\text{"student"})$=?
- $g(D,A=\text{"credit\_rating"})$=?

# Information Gain (ID3)

- Compute the mutual information between $D$ (类标签) and each attribute $A$ (每个属性)
- $H(D)$=0.940
- $H(D|A$="$age$")=0.694

$$g(D,A) = I(D;A) = H(D) - H(D|A)$$

- $g(D,A$="$age$")=0.246
- $g(D,A$="$income$")=0.029
- $g(D,A$="$student$")=0.151
- $g(D,A$="$credit\_rating$")=0.048