Classification

Data Mining:
Data Mining Methods
with **Dr. Qin Lv**

Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

Learning objective: Apply techniques for classification and explain how they work. Evaluate and compare methods.

Supervised vs. Unsupervised

➢ Supervised learning (e.g., classification)

- Predefined classes, training data with groundtruth label
- Classify new data based on training data

➢ Unsupervised learning (e.g., clustering)

- No predefined classes
- Aims to identify potential clusters/patterns

Classification vs. Prediction

➢ Classification

- Categorical class labels
- E.g., fraud detection, disease diagnosis, object recognition

➢ Prediction

- Continuous numerical values
- E.g., stock price, traffic volume, #likes

Classification Process

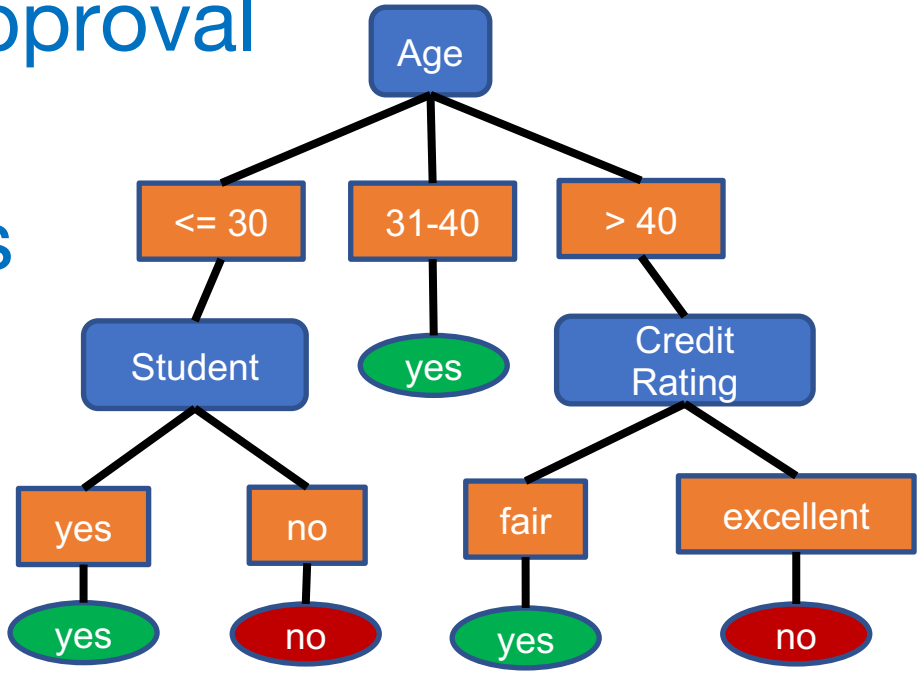
- Step 1: Learning
  - Training data, class labels, model construction
- Step 2: Classification
  - Test data, model evaluation, model selection
- Real-world deployment
  - New data, model adaptation

Evaluation Criteria

- Accuracy: classification vs. prediction
- Speed: model construction, online use
- Interpretability: explain the decision
- Robustness: noises, missing data
- Scalability: large data, incremental data
- …

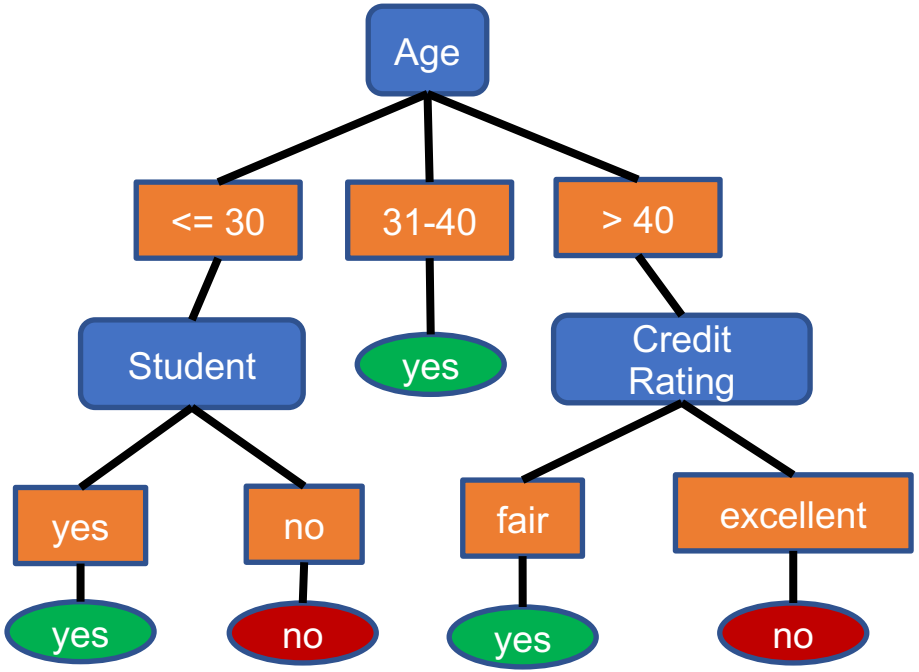
Decision Tree Induction (1)

- Loan application approval
  - Classes: Yes or No
- Applicant attributes
  - ID, age, student
  - Income, credit rating
  - ...



Decision Tree Induction (2)

- ➢ Basic algorithm
  - Attribute selection
  - Attribute split
- ➢ Key properties
  - Top-down, recursive
  - Divide-and-conquer
  - Greedy algorithm



Information Gain (ID3)

- Intuition reduce class entropy (information)
  - D, m classes C_i

    $p_i = |C_{i,D}|/|D|$

    $$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$$
- Classify D using attribute A
  - A: a_1, a_2, …, a_v

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

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

DT Example (1)

- Loan approval
  - 12 applicants
  - 2 classes: yes(7), no(5)

ID	Age	Income	Student	Credit Rating	Loan
1	<= 30	high	no	fair	no
2	<= 30	high	no	excellent	no
3	31-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	excellent	no
6	31-40	low	yes	excellent	yes
7	<= 30	medium	no	fair	no
8	<= 30	low	yes	fair	yes
9	>40	medium	yes	fair	yes
10	<= 30	medium	yes	excellent	yes
11	31-40	high	yes	fair	yes
12	>40	medium	no	excellent	no

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

$$Info(D) = I(7,5) = -\frac{7}{12}log_2(\frac{7}{12}) - \frac{5}{12}log_2(\frac{5}{12}) = 0.980$$

DT Example (2)

ID	Age	Income	Student	Credit Rating	Loan
1	<= 30	high	no	fair	no
2	<= 30	high	no	excellent	no
3	31-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	excellent	no
6	31-40	low	yes	excellent	yes
7	<= 30	medium	no	fair	no
8	<= 30	low	yes	fair	yes
9	>40	medium	yes	fair	yes
10	<= 30	medium	yes	excellent	yes
11	31-40	high	yes	fair	yes
12	>40	medium	no	excellent	no

- Attribute Age:
  - <=30: I(2, 3)
  - 31-40: I(3, 0)
  - > 40: I(2, 2)
- Information gain
  - 0.980 – 0.738 = 0.242

$$Info_{Age}(D) = \frac{5}{12}I(2,3) + \frac{3}{12}I(3,0) + \frac{4}{12}I(2,2) = 0.738$$

Other DT Methods

- Gain Ratio (C4.5) $SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2\left(\frac{|D_j|}{|D|}\right)$

$$gainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

- Gini Index (CART) $Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$

$$Gini_A(D) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2)$$

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

Bayesian Classification

- Bayes' Theorem $$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$
- Statistical classifier
- X: a data sample, class unknown
- H: hypothesis that X belongs to class C
- P(H), P(X): prior probability
- P(X|H), P(H|X): posterior probability

Naïve Bayesian Classifier

- $X = (x_1, x_2, \ldots, x_n)$ (i.e., n attributes)
- m classes: $C_1, C_2, .., C_m$
- Classification: maximal $P(C_i|X)$

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

- Ignore P(X) since it's a constant for all classes
- Naïve assumption: no dependence between attributes
- 0-probability => add 1 to each case (Laplacian correction)

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

NB Example (1)

- Loan: 2 classes
  - P(loan = “yes”) = 7/12
  - P(loan = “no”) = 5/12
- Applicant X
  - Age <= 30
  - Income = medium
  - Student = yes, Credit rating: fair

ID	Age	Income	Student	Credit Rating	Loan
1	<= 30	high	no	fair	no
2	<= 30	high	no	excellent	no
3	31-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	excellent	no
6	31-40	low	yes	excellent	yes
7	<= 30	medium	no	fair	no
8	<= 30	low	yes	fair	yes
9	>40	medium	yes	fair	yes
10	<= 30	medium	yes	excellent	yes
11	31-40	high	yes	fair	yes
12	>40	medium	no	excellent	no

NB Example (2)

- X (loan=yes, loan=no)
- Age <= 30 (2, 3)
- Income = medium (3, 2)
- Student = yes (5, 1)
- Credit rating = fair (5, 2)

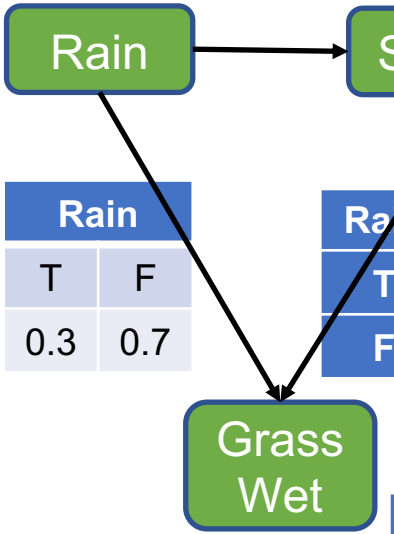
ID	Age	Income	Student	Credit Rating	Loan
1	<= 30	high	no	fair	no
2	<= 30	high	no	excellent	no
3	31-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	excellent	no
6	31-40	low	yes	excellent	yes
7	<= 30	medium	no	fair	no
8	<= 30	low	yes	fair	yes
9	>40	medium	yes	fair	yes
10	<= 30	medium	yes	excellent	yes
11	31-40	high	yes	fair	yes
12	>40	medium	no	excellent	no

NB Example (3)

- P(age <=30 | loan = yes) = 2/7
- P(income = medium | loan = yes) = 3/7
- P(student = yes | loan = yes) = 5/7
- P(credit_rating = fair | loan = yes) = 5/7
- P(X|loan = yes) * P(loan = yes) = 2/7 * 3/7 * 5/7 * 5/7 * 7/12
- Same process for P(X|loan = no) * P(loan = no)

Bayesian Belief Network

- Conditional dependency of variables
  - Probabilistic graphical model
  - Directed acyclic graph
  - Conditional probability table



Rain	
T	F
0.3	0.7

Rain	Sprinkler	
	T	F
T	0.1	0.9
F	0.5	0.5

Rain	Sprinkler	Grass Wet	
		T	F
T	T	0.99	0.01
T	F	0.9	0.1
F	T	0.8	0.2
F	F	0.0	1.0