University of Colorado Boulder

CSCI 4502/5502

Data Mining

Fall 2019
Lecture 10 (Sep 26)

Announcements

- Homework 3
  - due at 9:30am, Thursday, Sep 26

- No new homework today
  - work on your project proposal

Course Project (40%)

- A self-contained project related to this course's topics
- Team of 3-4 students
  - other group size w/ instructor permission
  - can mix students in different sections
- Pick your own project idea
- Discuss project ideas with the instructor, TA, and other students
- Data & subtasks

Course Project Announcement

- Due at 9:30am, Thursday Oct 3
- Post at moodle project forum
- One announcement per team
  - title
  - team members (name, course section)
  - brief project description
  - dataset(s) to use
  - tool(s) to use -- if known

Proposal Summary Slides

- Due at 9:30am, Th Oct 3; submit at Moodle
- A few slides to summarize your proposal
  - title, team (name, course section)
  - problem, prior work, proposed work
  - data, tools, how to evaluate
- Used for project presentation & discussion
  - will send out a poll to check teams' time availability

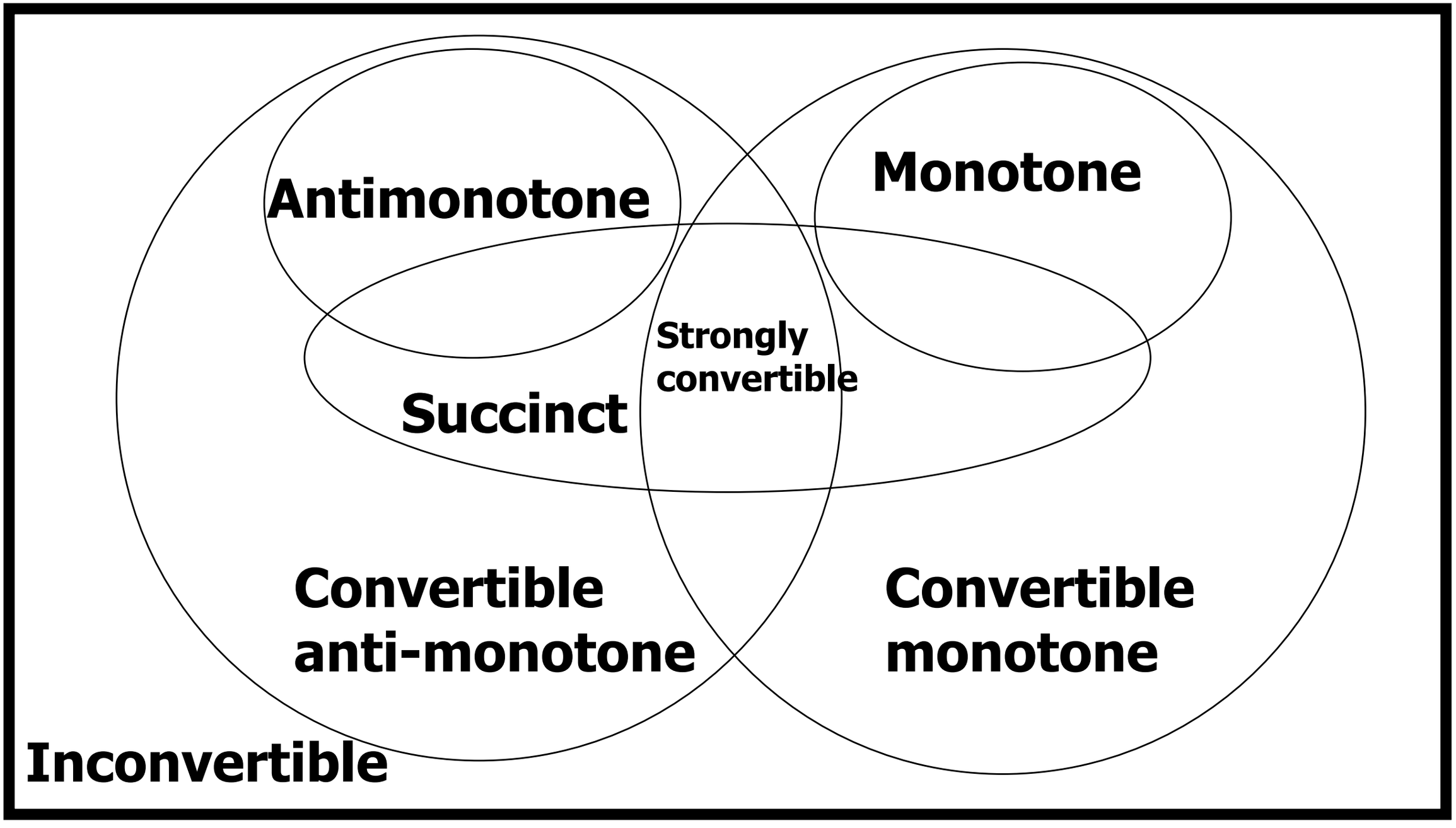
Course Project Proposal

- Due at 9:30am, Th Oct 10; submit at Moodle
- Course project proposal (~3 pages)
  - ACM SIG Proceedings Templates
  - title, team (name, course section)
  - motivation, literature survey
  - proposed work (data, subtasks)
  - how to evaluate, milestones
  - brief summary of project discussion

Review

- Chap 6 & 7: Frequent Pattern Mining
  - correlations
  - road map
  - association rules
    - multi-level: support
    - multi-dimensional: intra, inter, hybrid
    - quantitative: static/dynamic
      discretization, clustering, deviation
  - constraint-based mining

Classification of Constraints



Example: Apriori Algorithm

tid	items
1	1, 3, 4
2	2, 3, 5
3	1, 2, 3, 5
4	2, 5

min_sup = 2

itemset	sup
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

itemset	sup
{1,2}	1
{1,3}	2
{1,5}	1
{2,3}	2
{2,5}	3
{3,5}	2

itemset	sup
{2, 3, 5}	2

Apriori + Constraint

tid	items
1	1, 3, 4
2	2, 3, 5
3	1, 2, 3, 5
4	2, 5

min_sup = 2

constraint:
sum(S.price)<5

itemset	sup
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

itemset	sup
{1,2}	1
{1,3}	2
{1,5}	1
{2,3}	2
{2,5}	3
{3,5}	2

itemset	sup
{2, 3, 5}	2

Apriori + Constraint

tid	items
1	1, 3, 4
2	2, 3, 5
3	1, 2, 3, 5
4	2, 5

min_sup = 2

constraint:
sum(S.price)<5

itemset	sup
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

itemset	sup
{1,2}	1
{1,3}	2
{1,5}	1
{2,3}	2
{2,5}	3
{3,5}	2

itemset	sup
{2, 3, 5}	2

Apriori + Constraint

tid	items
1	1, 3, 4
2	2, 3, 5
3	1, 2, 3, 5
4	2, 5

min_sup = 2

itemset	sup
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

itemset	sup
{1,2}	1
{1,3}	2
{1,5}	1
{2,3}	2
{2,5}	3
{3,5}	2

constraint:
min(S.price) ≤ 1

itemset	sup
{2, 3, 5}	2

Apriori + Constraint

tid	items
1	1, 3, 4
2	2, 3, 5
3	1, 2, 3, 5
4	2, 5

min_sup = 2

constraint:
min(S.price) ≤ 1

itemset	sup
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

itemset	sup
{1,2}	1
{1,3}	2
{1,5}	1
{2,3}	2
{2,5}	3
{3,5}	2

itemset	sup
{2, 3, 5}	2

Summary (1)

- Frequent patterns
  - itemset, subsequence, substruture
  - closed patterns, max-patterns
  - kinds of patterns, completeness, levels & dimensions, types of values, kinds of rules
- Mining frequent itemsets
  - Apriori, FP-growth, vertical data format
  - partition, sampling, hash-tree

Summary (2)

- Correlation analysis
  - support, confidence, correlation
  - lift, X^2, all_conf, max_conf, cosine, Kulc
- Mining various association rules
  - multi-level, multi-dimensional, quantitative
- Constraint-based mining
  - anti-monotonic, monotonic, succinct, convertible, inconvertible

Chapter 8
Classification: Basic Concepts

Chap 8: Classification

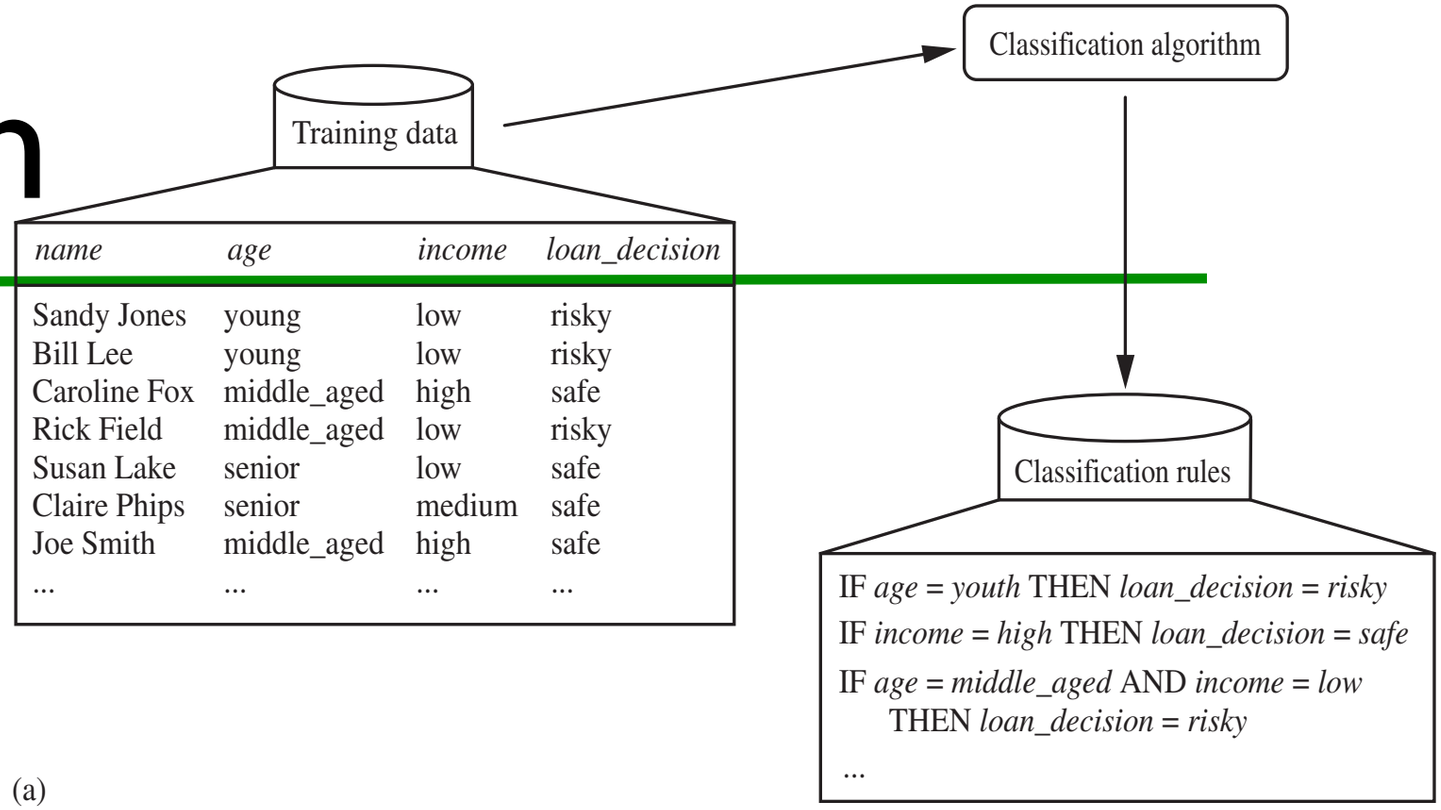
- Basic concepts
- Decision tree induction
- Bayesian classification
- Rule-based classification
- Model evaluation and selection
- Improve classification accuracy
- Summary

Classification vs. Prediction

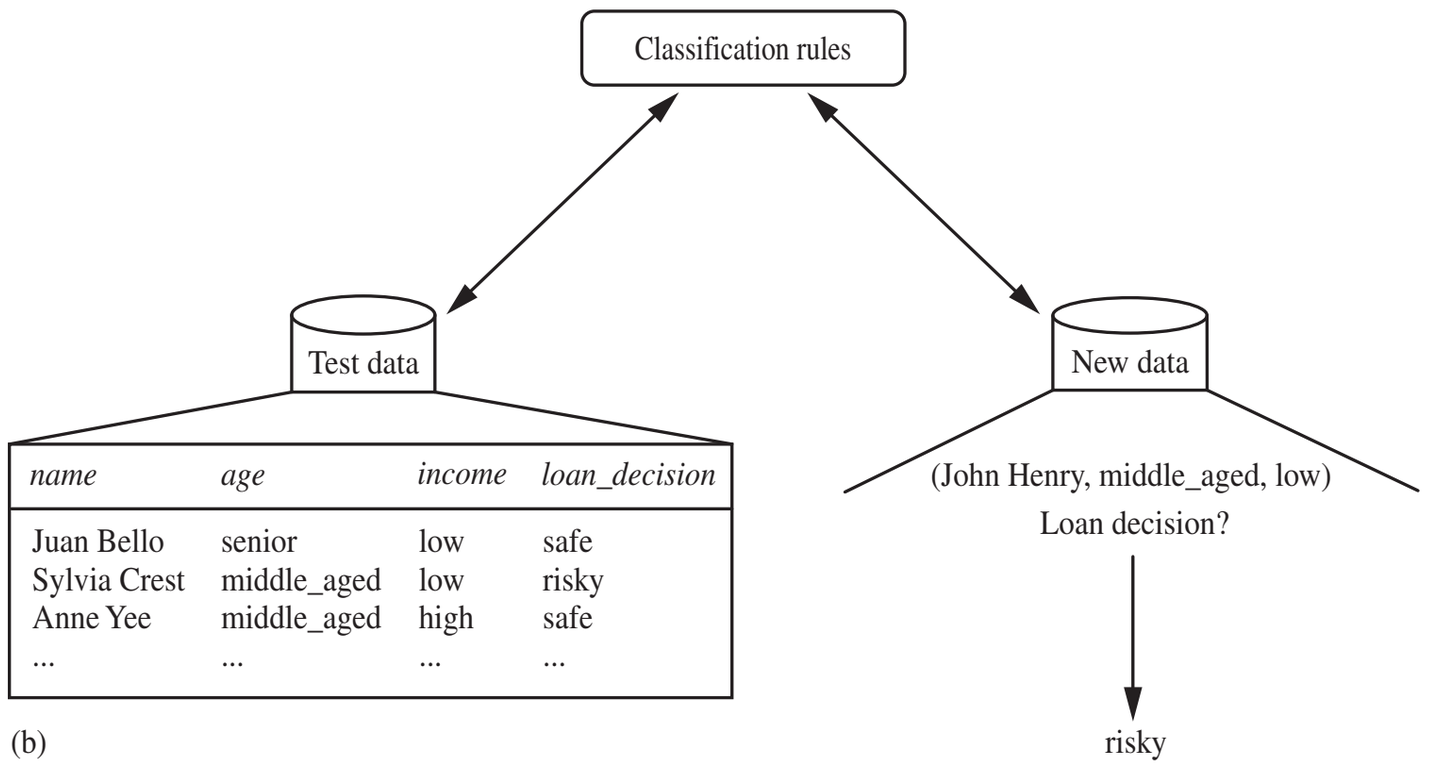
- Classification
  - determines categorical class labels
  - e.g., safe vs. risky; weather condition
- Prediction
  - models continuous-valued functions
- Typical applications
  - loan approval, target marketing, medical diagnosis, fraud detection, etc.

Classification

- Step 1: Learning
  - model construction
  - training set
  - class labels
- Step 2: Classification
  - test set
  - accuracy



(a)



(b)

Supervised vs. Unsupervised

- Supervised learning (classification)
  - supervision: training data accompanied by class labels
  - new data is classified based on training set
- Unsupervised learning (clustering)
  - class labels of training data is unknown
  - aims to establish the existence of classes or clusters in the data

Issues: Evaluation Criteria

- Accuracy: classification vs. prediction
- Speed: time to construct / use the model
- Robustness: handling noise & missing values
- Scalability: large amounts of data
- Interpretability: understanding and insight
- Goodness of rules: e.g., decision tree size, compactness of classification rules

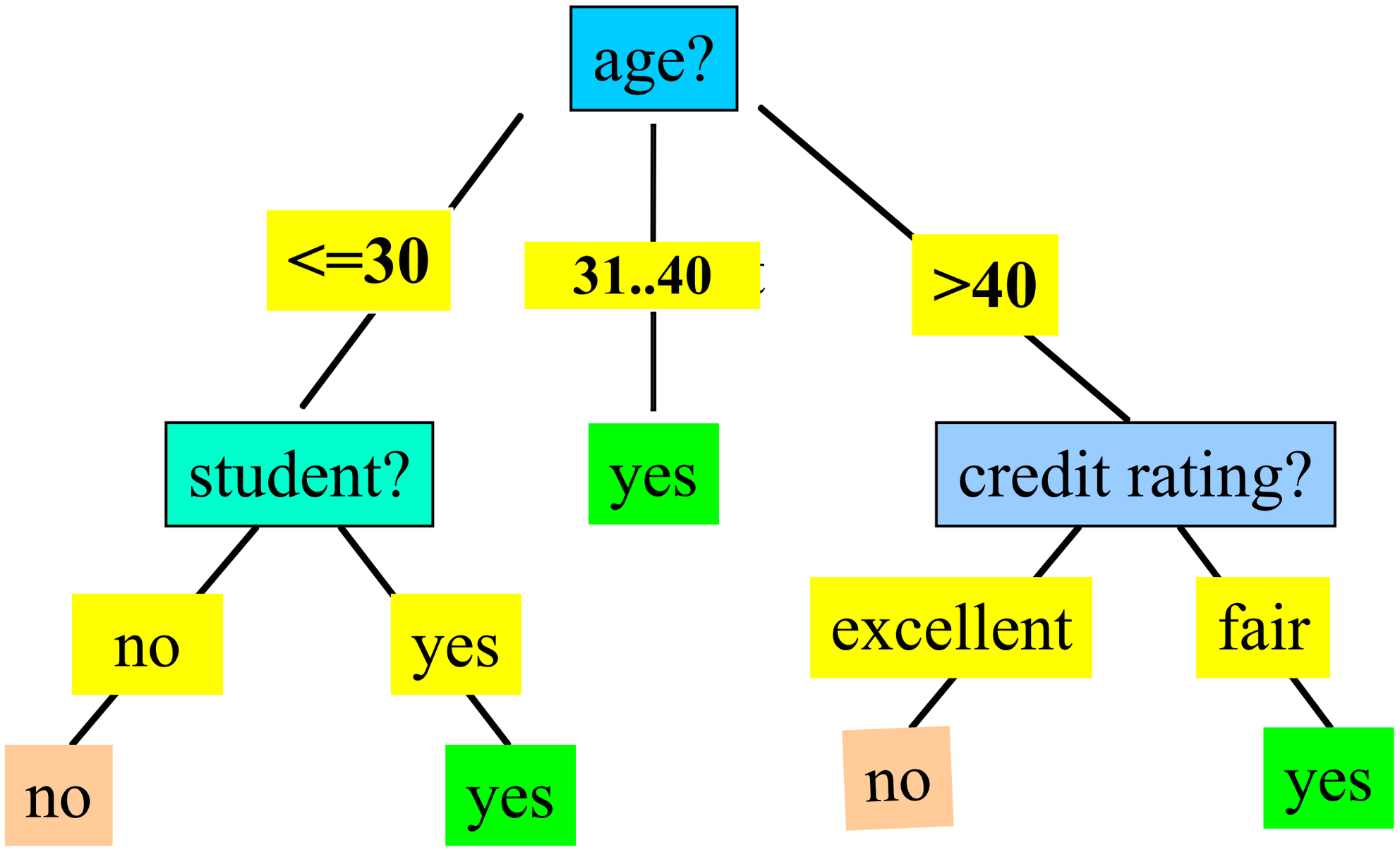
Chap 8: Classification

- Basic concepts
- Decision tree induction
- Bayesian classification
- Rule-based classification
- Model evaluation and selection
- Improve classification accuracy
- Summary

Example: Training Set

CID	age	income	student	credit_rating	buys_computer
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	fair	yes
6	>40	low	yes	excellent	no
7	31-40	low	yes	excellent	yes
8	<= 30	medium	no	fair	no
9	<= 30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<= 30	medium	yes	excellent	yes
12	31-40	medium	no	excellent	yes
13	31-40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Example: Decision Tree

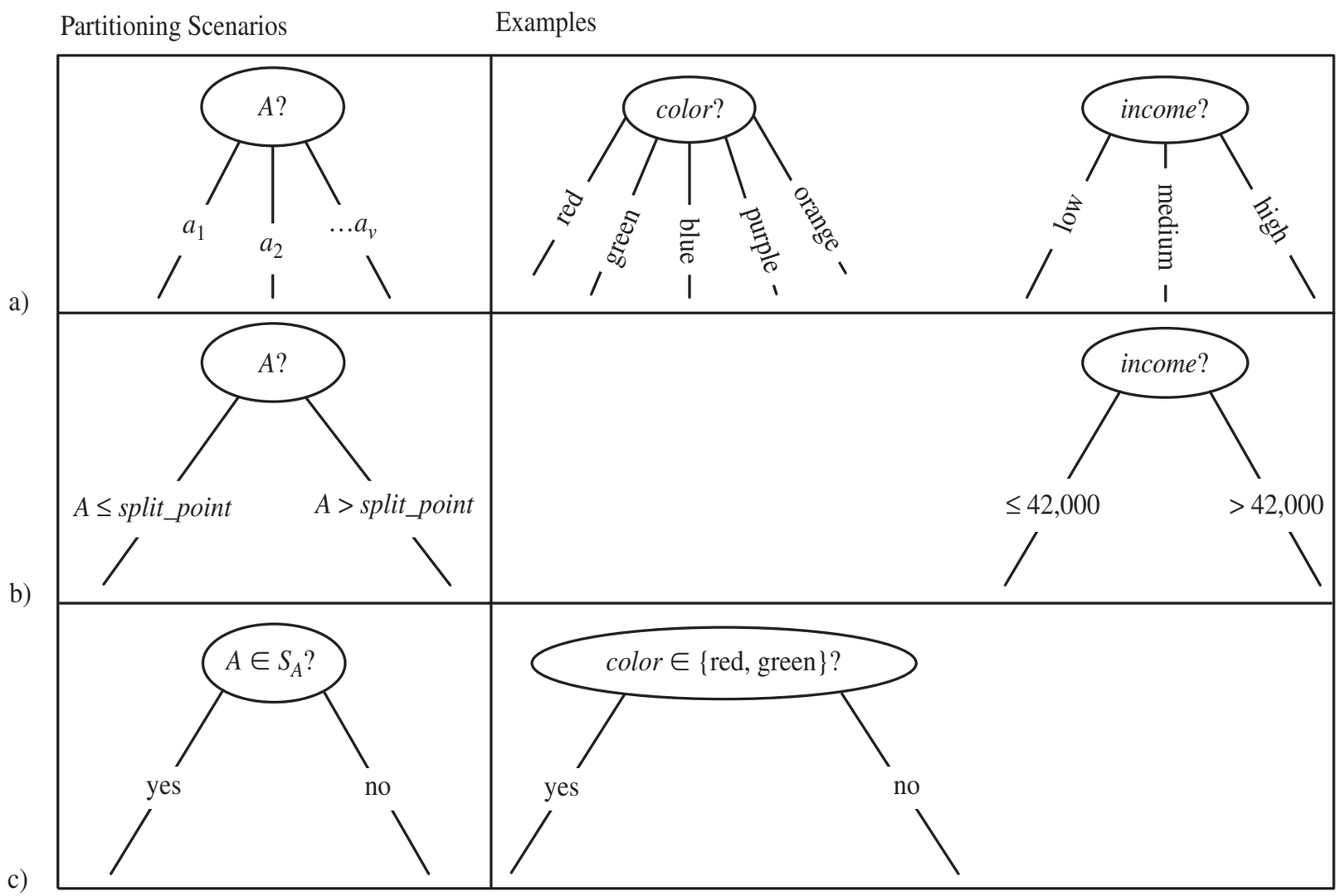


Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - top-down, recursive, divide-and-conquer
  - attribute selection
  - attribute split

Splitting Attributes

- Discrete-valued
- Continuous-valued: split_point
- Discrete-valued: binary tree, splitting_subset



Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - top-down, recursive, divide-and-conquer
  - attribute selection
  - attribute split
- Stopping conditions
  - all samples belong to the same class
  - no remaining attributes: majority voting
  - no samples left

Reminder

- Due at 9:30am, Thursday, Oct 3
  - course project announcement
    - post to moodle project discussion forum
  - course project proposal summary slides
    - submit at moodle

- Due at 9:30am, Thursday, Oct 10
  - course project proposal report