Knowledge Discovery & Data Mining
– Data Mining Tasks–

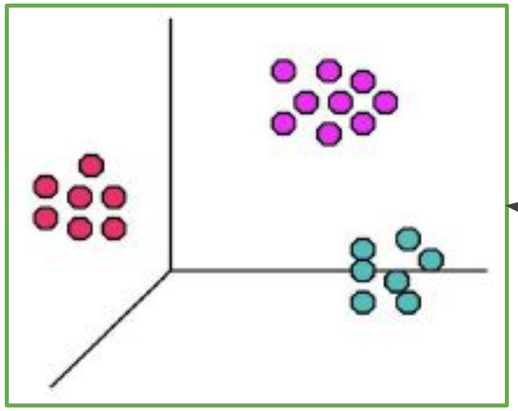
Instructor: Yong Zhuang

yong.zhuang@gvsu.edu

Outline

- Common data mining tasks
  - Predictive modeling
    - Classification
    - Regression
    - Ranking
  - Clustering
  - Association rule mining
  - Anomaly detection

Four common data mining tasks



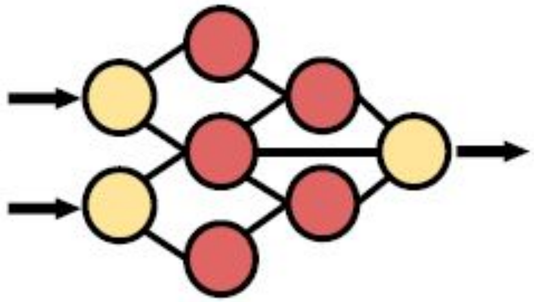
Clustering

Association Rules

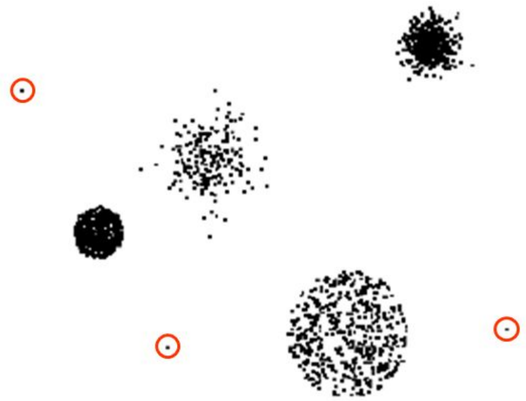
Data

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	Species
0	4.9	3.6	1.4	0.1	0
1	6.7	3.0	5.0	1.7	1
2	5.8	2.7	3.9	1.2	1
3	4.9	3.0	1.4	0.2	0
4	6.8	3.2	5.9	2.3	2
5	4.9	2.4	3.3	1.0	1
6	4.9	3.1	1.5	0.1	0
7	5.5	2.5	4.0	1.3	1
8	5.1	3.8	1.5	0.3	0
9	6.1	2.8	4.0	1.3	1
10	6.7	3.0	5.2	2.3	2

Predictive Modeling



Anomaly Detection



Predictive Modeling

Goal: Predict the value of an attribute based on the values of other attributes.

- The attribute to be predicted is often called the target attribute (also known as the dependent variable or response variable).
- The attributes used to make the prediction are often called explanatory attributes (also known as independent variables or predictors).

Examples

- Predicting future price of a stock
- Predicting the annual rainfall at a location for the next 20 years
- Predicting whether a customer will buy something at a website
- Predicting who should be a friend of whom
- Predicting which web page to display when a user entered a search query

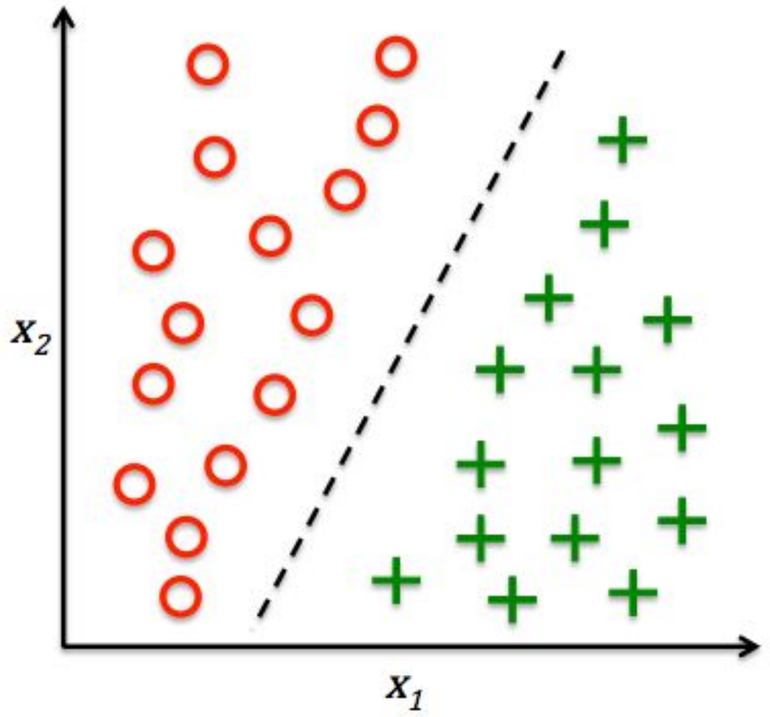
Predictive Modeling: Classification

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.

- The model is derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known).
- The model is used to predict the class labels of objects for which the class labels are unknown.

Examples:

- Text categorization
- Image classification
- Medical diagnosis
- Spam detection



Predictive Modeling: Classification

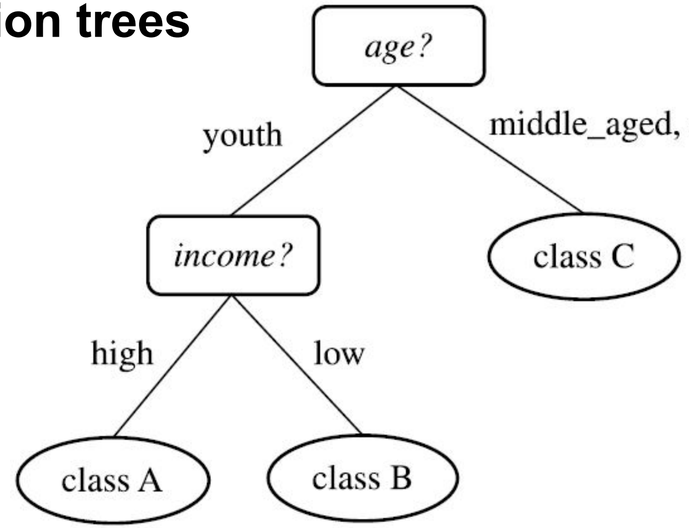
Typical methods:

Rule-based classification

age(X, "youth") AND income(X, "high") ⟶ *class(X, "A")*
age(X, "youth") AND income(X, "low") ⟶ *class(X, "B")*
age(X, "middle_aged") ⟶ *class(X, "C")*
age(X, "senior") ⟶ *class(X, "C")*

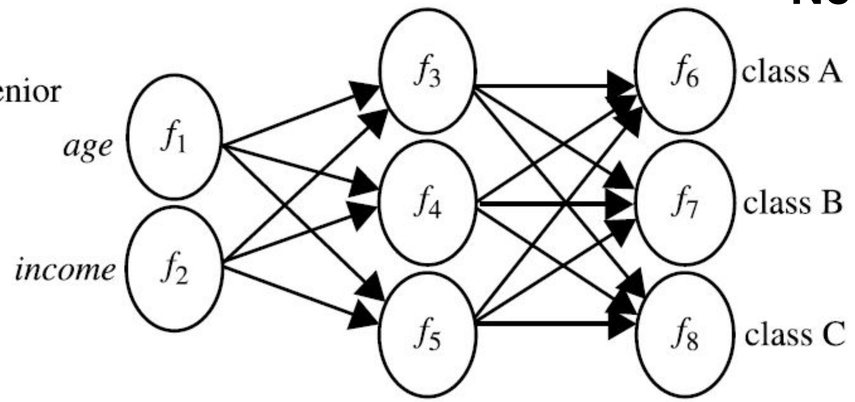
(a)

Decision trees



(b)

Neural networks



(c)

Predictive Modeling: Regression

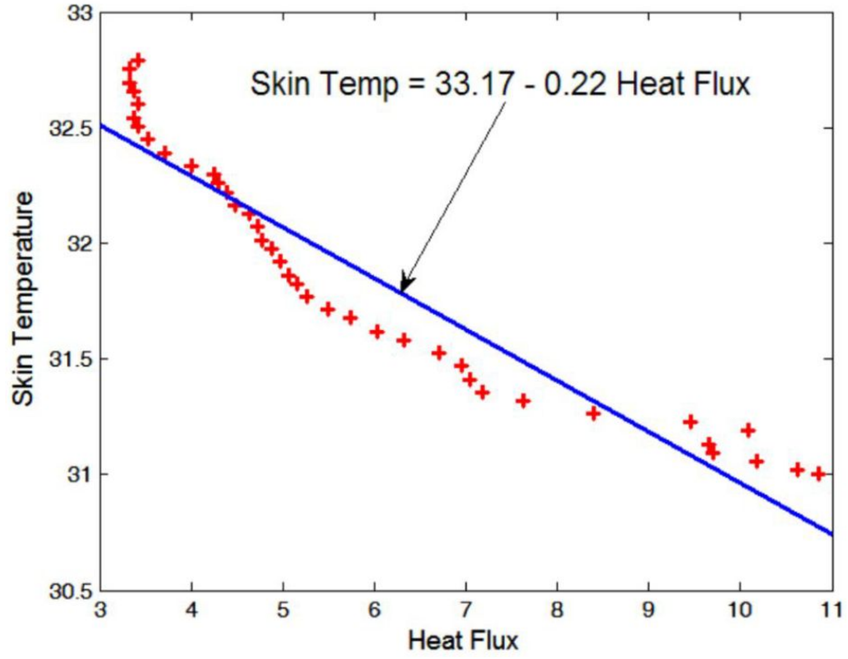
Classification predicts **symbolic** (categorical, discrete, unordered, nominal) labels.

Regression predict missing or unavailable **numerical** data values.

Heat Flux	Skin Temperature
10.858	31.002
10.617	31.021
10.183	31.058
9.7003	31.095
9.652	31.133
10.086	31.188
9.459	31.226
8.3972	31.263
7.6251	31.319
7.1907	31.356
7.046	31.412
6.9494	31.468
6.7081	31.524

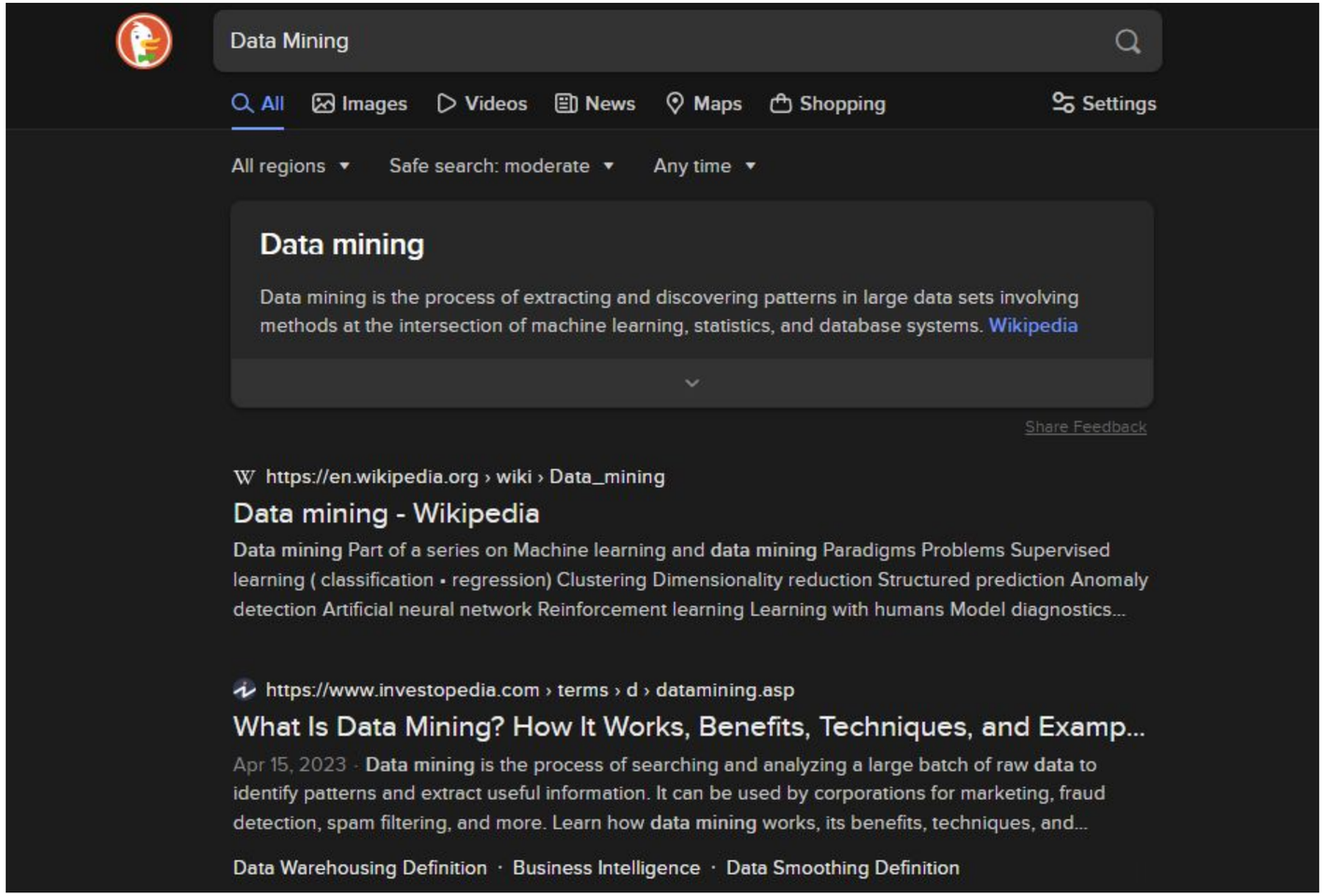
Heat Flux	Skin Temperature
6.3221	31.581
6.0325	31.618
5.7429	31.674
5.5016	31.712
5.2603	31.768
5.1638	31.825
5.0673	31.862
4.9708	31.919
4.8743	31.975
4.7777	32.013
4.7295	32.07
4.633	32.126
4.4882	32.164

Example: Physiological data from wearable device



Predictive Modeling: Ranking

The target attribute to be predicted is **ordinal**-valued.

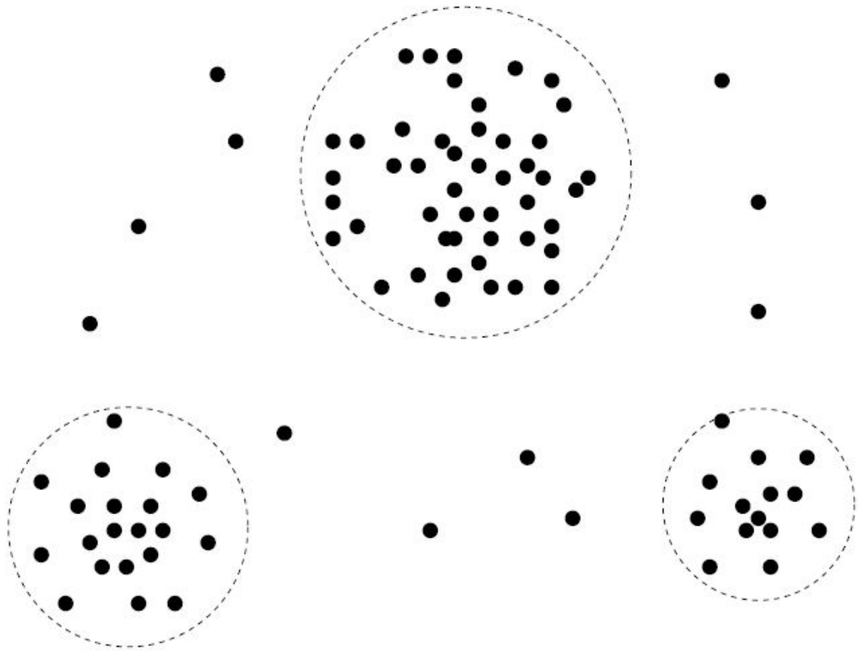


Clustering

Goal: Find groups of objects such that the objects in the same group are more similar to each other than objects from other groups.

Examples:

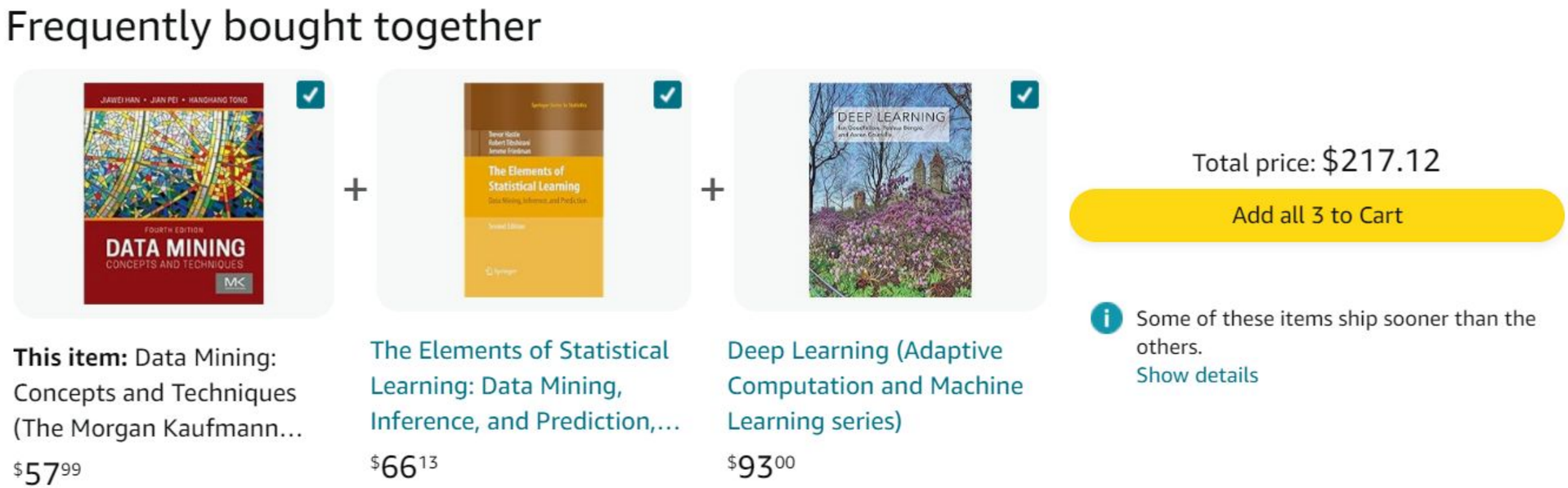
- Document clustering
- Time series clustering



Association Rule Mining

Goal: Find associations (e.g., relationships, dependencies) in sets of data items.

What items are frequently purchased together in your Amazon transactions?

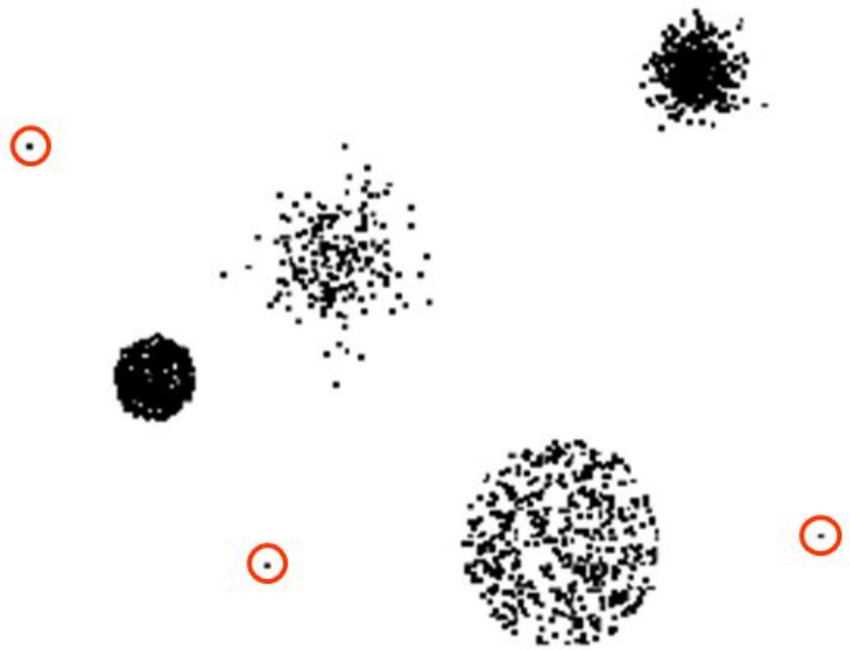


Anomaly(Outlier) Detection

Goal: Identify objects that are different from most other objects.

Examples:

- Credit card fraud detection
- Network intrusion detection
- Event detection in sensor networks
- Defect detection



Outlier analysis

- Outlier: A data object that does not comply with the general behavior of the data
- Tasks: Outlier analysis or anomaly mining.
- Detection:
  - Statistical tests: assume a distribution or probability model for the data, or using distance measures where objects that are remote from any other cluster are considered outliers.
  - Density-based methods: identify outliers in a local region, although they look normal from a global statistical distribution view.

Many data mining methods discard outliers as noise or exceptions. However, **in some applications, the rare events can be more interesting than the more regularly occurring ones.**

Summary

- Common data mining tasks
  - Predictive modeling
    - Classification
    - Regression
    - Ranking
  - Clustering
  - Association rule mining
  - Anomaly detection