
School of Computing
UNIVERSITY OF GEORGIA

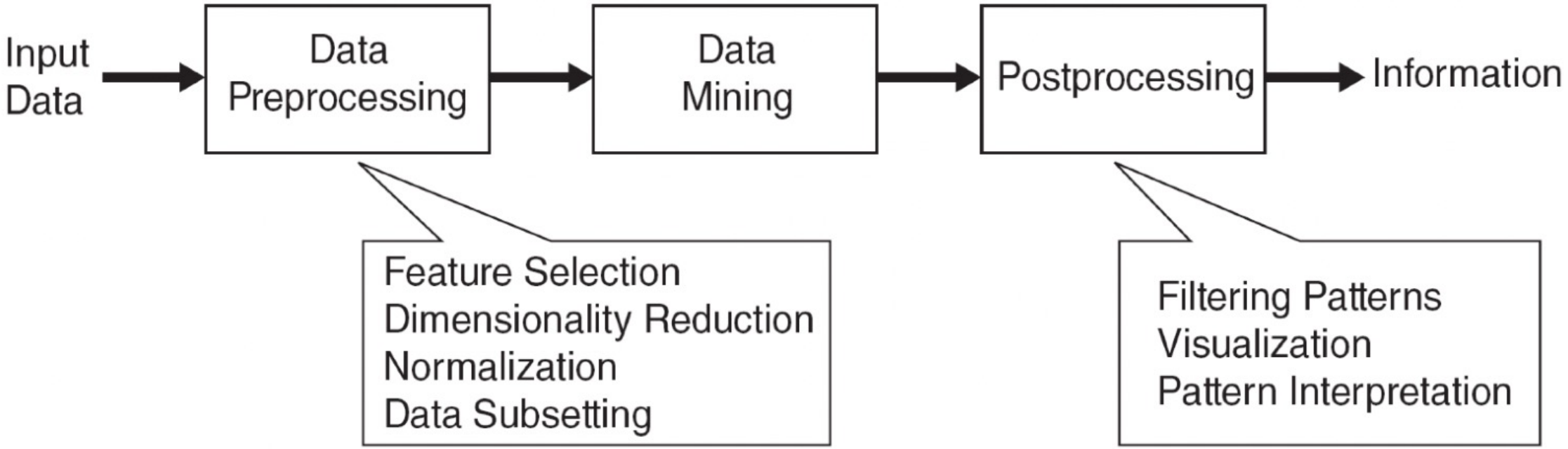
# CSCI 4380/6380 DATA MINING

Fei Dou

Assistant Professor
School of Computing
University of Georgia

September 27, 2023

# Recap: Data Mining Process

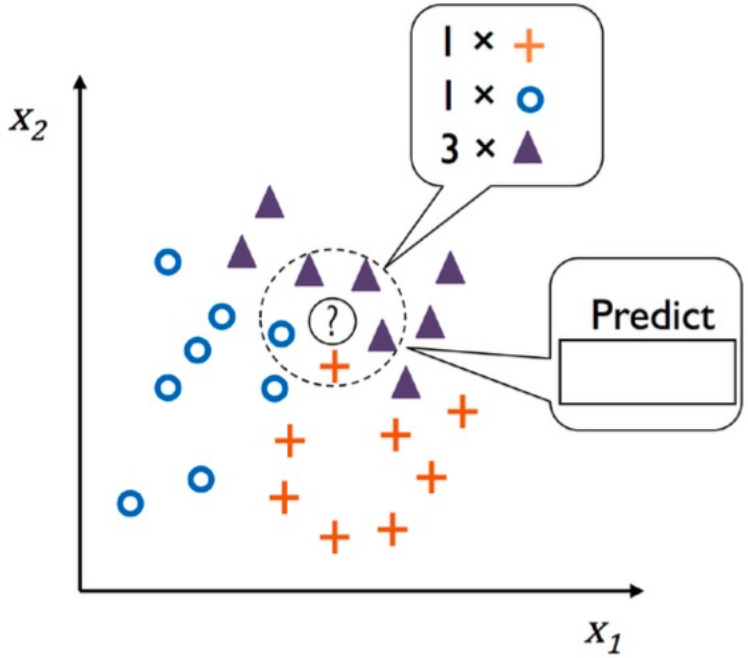


# Classification - Nearest-Neighbor Classifiers

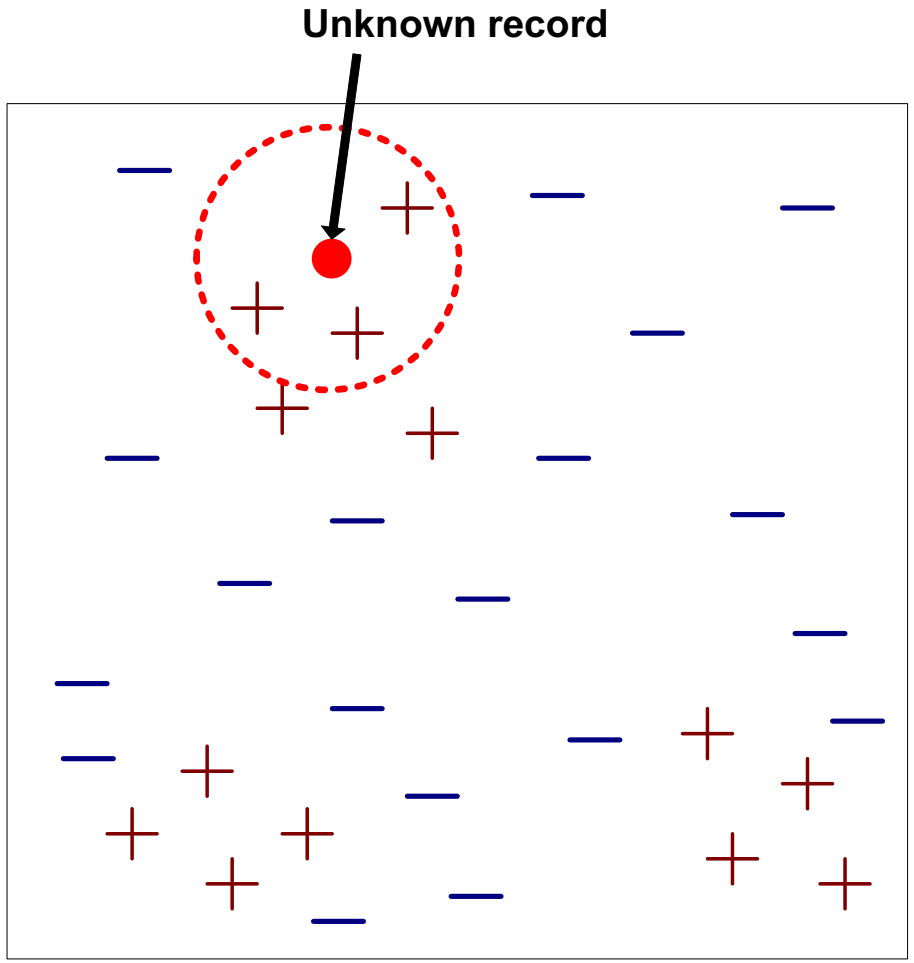
# Classification - Nearest-Neighbor Classifiers

## k-Nearest Neighbor Classifier

- Training set: $D=\{(\mathbf{x}_i, y_i), i=1,\cdots,n\}, \mathbf{x}_i \in R^d$
- $d(\mathbf{x}_q, \mathbf{x}_i)$ for $i = \{1,2,...,n\}$, $d(\cdot)$ can be Euclidean distance or other distance measures.
- Assign label by majority (purality) vote with k nearest neighbors



# Classification - Nearest-Neighbor Classifiers



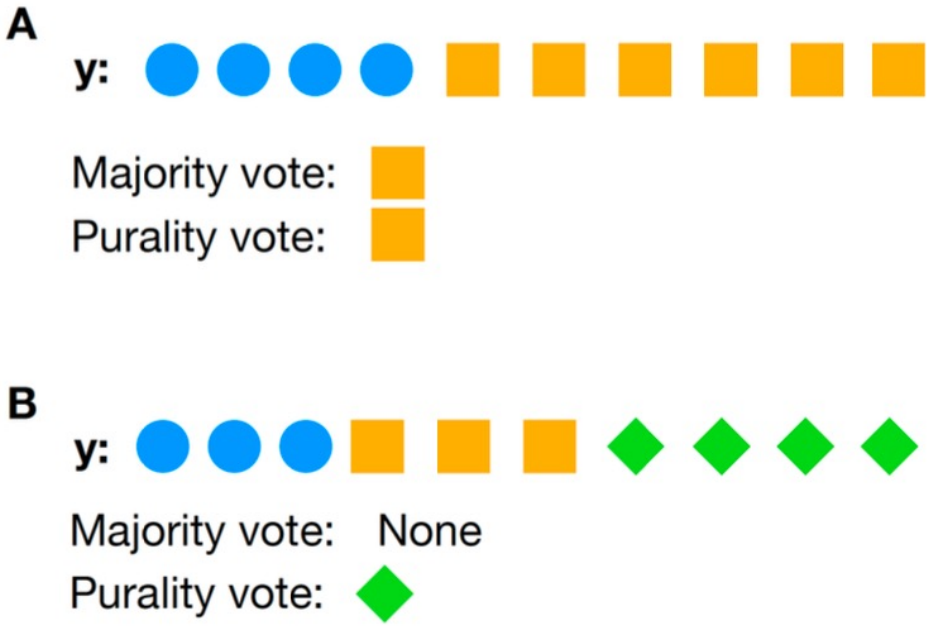

**Requires the following:**

- A set of labeled records
- Proximity metric to compute distance/similarity between a pair of records
  - e.g., Euclidean distance
- The value of *k*, the number of nearest neighbors to retrieve
- A method for using class labels of K nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

# How to Determine the class label of a Test Sample?

## k-Nearest Neighbor Classifier

- Training set: $D=\{(\mathbf{x}_i,y_i),i=1,\cdots,n\},\mathbf{x}_i \in R^d$
- $d(\mathbf{x}_q,\mathbf{x}_i)$ for $i = \{1,2,...,n\}$, $d(\cdot)$ can be Euclidean distance or other distance measures.
- Assign label by majority (purality) vote with k nearest neighbors



# How to Determine the class label of a Test Sample?

**Others**

- Weighted Majority Vote:
  - e.g. weight factor, $w = \frac{1}{d}$ , $w = 1/d^2$

# Choice of Proximity Measure Matters

- For documents, cosine is better than correlation or Euclidean

| | | | |
|---|---|---|---|
| 1 1 1 1 1 1 1 1 1 1 1 0 | | 0 0 0 0 0 0 0 0 0 0 0 1 |
| | vs | |
| 0 1 1 1 1 1 1 1 1 1 1 1 | | 1 0 0 0 0 0 0 0 0 0 0 0 |

Euclidean distance = 1.4142 for both pairs, but the cosine similarity measure has different values for these pairs.

# Nearest Neighbor Classification

- **Data preprocessing is often required**
  - Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
    - Example:
      - height of a person may vary from 1.5m to 1.8m
      - weight of a person may vary from 90lb to 300lb
      - income of a person may vary from $10K to $1M

**Practice:**

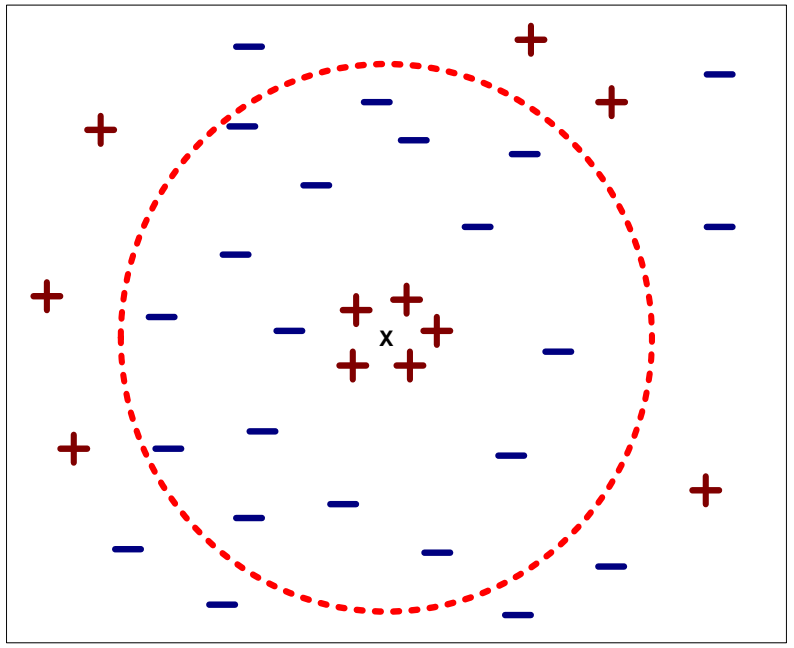
Height: 1.65m
Weight: 150lb
Income: $50,000

Using min-max scaling

  - Time series are often standardized to have 0 means a standard deviation of 1

# Nearest Neighbor Classification

- **Choosing the value of k:**
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes



# Classification - kNN

**Advantages**:

- Easy to implement
- No optimization or training is required
- Classification accuracy can be very good; can outperform more complex models

# Classification - kNN

- **Scaling issues**
  - Multiple attributes/features will need to be scaled to prevent distance measures from being dominated by one of the time series.

- **kNN classifier is a lazy learner**
  - It does not build models explicitly
  - Different from eager learners such as decision tree induction
  - Classifying unknown sample/example is relatively expensive

- **Curse of Dimensionality**
  - For high-dimensional space, the problem is hat the nearest neighbor may not be very close at all.

# Improving kNN Efficiency

- Avoid having to compute distance to all objects in the training set
  - Multi-dimensional access methods (k-d trees)
  - Fast approximate similarity search
  - Locality Sensitive Hashing (LSH)
- Condensing
  - Determine a smaller set of objects that give the same performance
- Editing
  - Remove objects to improve efficiency