CS 579: Online Social Network Analysis

Data Mining Essentials

Spring 2022

Kai Shu

Reading: Chapter 5

Introduction

- Data production rate has increased dramatically and we are able store much more data than before:
  - "Big Data"
  - E.g., purchase data, social media data, mobile phone data
- Businesses and customers **need useful or actionable knowledge** and gain insight from raw data for various purposes
  - Not just searching data or databases

The process of extracting useful patterns from raw data is known as Knowledge Discovery in Databases (KDD).

Data Mining

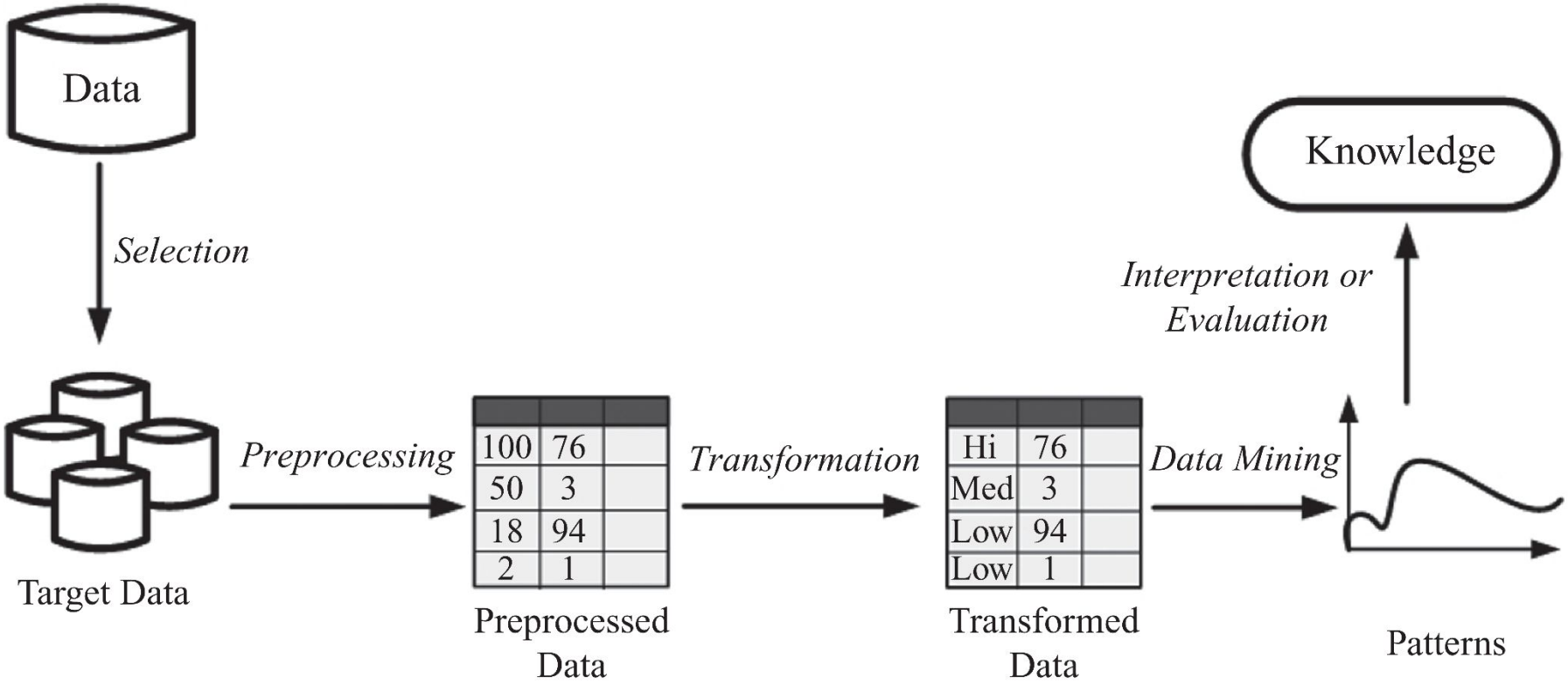
The process of discovering hidden patterns in large data sets

It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems

- *Extracting or "mining" knowledge from large amounts of data, or big data*
- Data-driven discovery and modeling of hidden patterns in big data
- Extracting implicit, previously unknown, unexpected, and potentially useful information/knowledge from data

Data

KDD Process

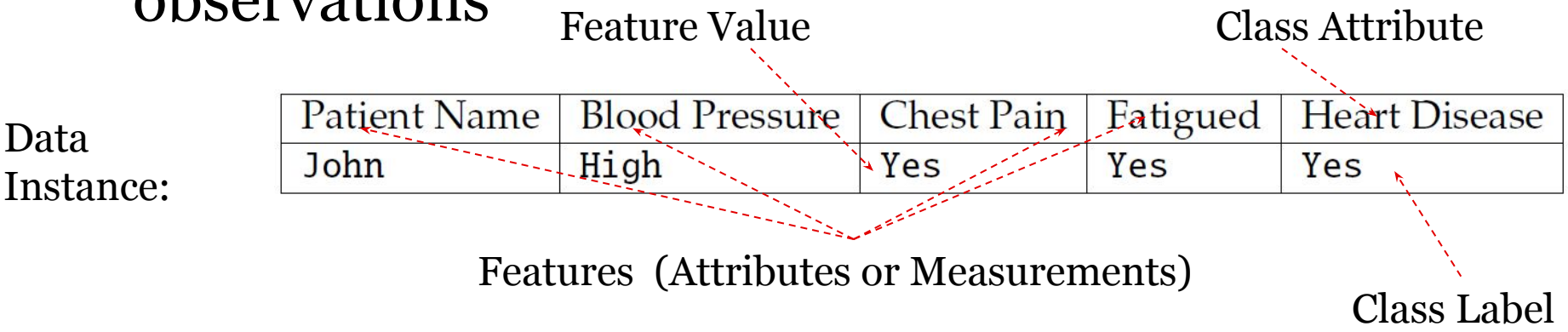

Data
Selection
Target Data
Preprocessing
100 76
50 3
18 94
2 1
Preprocessed Data
Transformation
Hi 76
Med 3
Low 94
Low 1
Transformed Data
Data Mining
Patterns
Interpretation or Evaluation
Knowledge

Examples of Data Mining Applications

- Identifying fraudulent transactions of a credit card or spam emails
  - You are given a user's purchase history and a new transaction, identify whether the transaction is fraud or not;
  - Determine whether a given email is spam or not
- Extracting purchase patterns from existing records
  - beer ⇒ dippers (80%)
- Forecasting future sales and needs according to some given samples
- Extracting groups of like-minded people in a given network

Data Instances

- In the KDD process, data is represented in a tabular format
- A collection of features related to an object or person
  - A patient's medical record
  - A user's profile
  - A gene's information
- Instances are also called points, data points, or observations



Data Instances

- Predicting whether an individual who visits an online book seller is going to buy a specific book

Attributes				Class
Name	Money Spent	Bought Similar	Visits	Will Buy
John	High	Yes	Frequently	?
Mary	High	Yes	Rarely	Yes

Unlabeled Example

Labeled Example

- Continuous feature: values are numeric values
  - Money spent: $25
- Discrete feature: Can take a number of values
  - Money spent: {high, normal, low}

Data Types + Permissible Operations (statistics)

- **Nominal** (categorical)
  - **Operations**: Mode (most common feature value), Frequency, Equality Comparison
  - E.g., {dog, cat, snake, bird}
- **Ordinal**
  - Feature values have an intrinsic order to them, but the difference is not defined
  - **Operations**: same as nominal, feature value rank
  - E.g., {small, medium, large, x-large}
- **Interval**
  - **Operations:** Addition and subtractions are allowed whereas divisions and multiplications are not
  - E.g., year, temperature (F/C)
- **Ratio**
  - **Meaningful zero point**
  - **Operations:** divisions and multiplications are allowed
  - E.g., Height, weight, money quantities

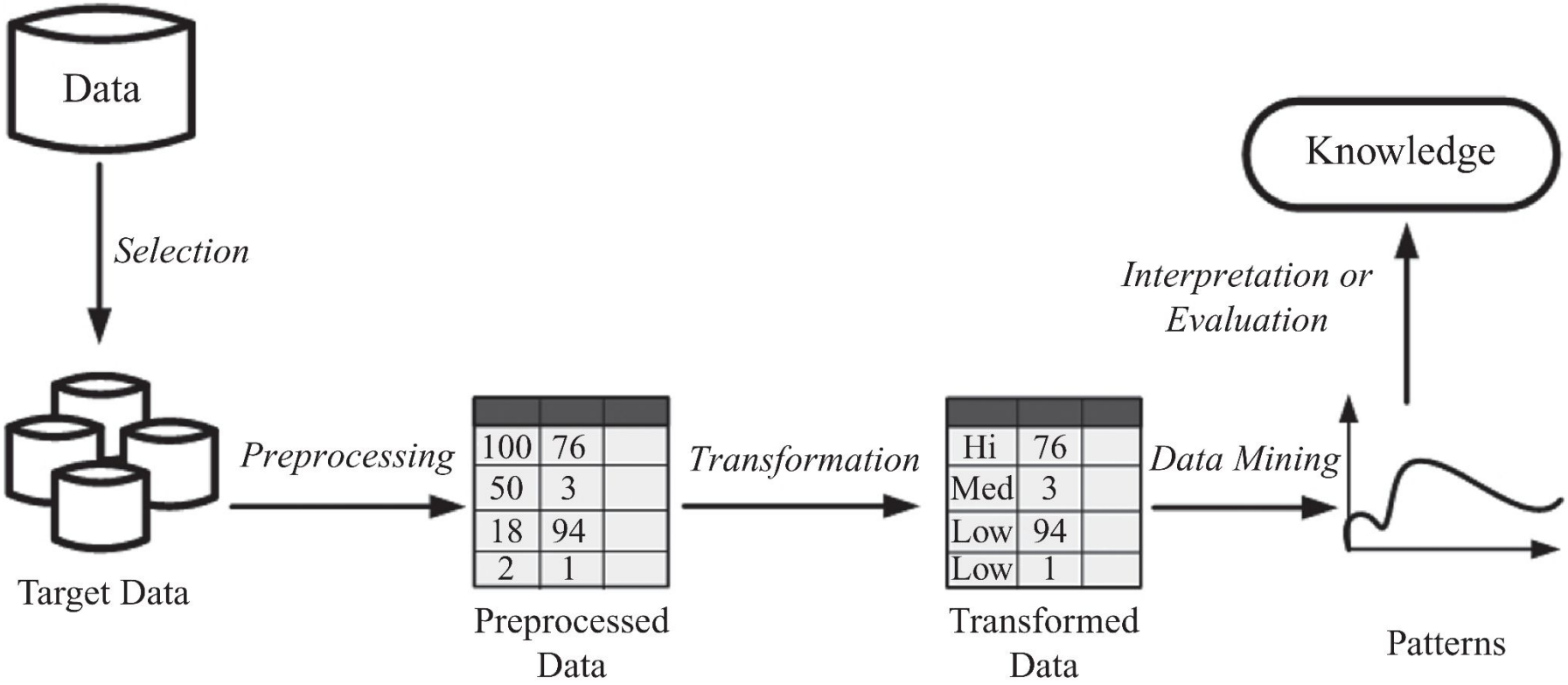
Sample Dataset

outlook	temperature	humidity	windy	play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rainy	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rainy	71	91	TRUE	no

Interval **Ratio** **Ordinal** **Nominal**

No.	Outlook (O)	Temperature (T)	Humidity (H)	Play Golf (PG)
1	sunny	hot	high	N
2	sunny	mild	high	N
3	overcast	hot	high	Y
4	rain	mild	high	Y
5	sunny	cool	normal	Y
6	rain	cool	normal	N
7	overcast	cool	normal	Y
8	sunny	mild	high	?

KDD Process


Data
Selection
Target Data
Preprocessing
100 76
50 3
18 94
2 1
Preprocessed Data
Transformation
Hi 76
Med 3
Low 94
Low 1
Transformed Data
Data Mining
Patterns
Interpretation or Evaluation
Knowledge

Text Representation

- The most common way to model documents is to transform them into sparse numeric vectors
- This representation is called “Bag of Words”
- Methods:
  - Vector space model
  - TF-IDF

Vector Space Model

- In the vector space model, we start with a set of documents, D
- Each document is a set of words
- The goal is to convert these textual documents to vectors

$$d_i = (w_{1,i}, w_{2,i}, \ldots, w_{N,i})$$

  - d_i : document i, $w_{j,i}$: the weight for word j in document i
  - **Design Choices:**
    - We can set it to 1 when the word j exists in document i and 0 when it does not.
    - We can also set this weight to the number of times the word j is observed in document i

Vector Space Model: An Example

- Documents:
  - d1: data mining and social media mining
  - d2: social network analysis
  - d3: data mining
- Vocabulary:
  - (analysis, data, media, mining, network, social)
- Vector representation:

	analysis	data	media	mining	network	social
d1	0	1	1	1	0	1
d2	1	0	0	0	1	1
d3	0	1	0	1	0	0

Term Frequency-Inverse Document Frequency (TF-IDF)

tf-idf of term *t*, document *d*, and document corpus *D* is calculated as follows:

$$w_{j,i} = tf_{j,i} \times idf_j,$$

$tf_{j,i}$ is the frequency of word j in document i

$$idf_j = \log_2 \frac{|D|}{|\{\text{document} \in D \mid j \in \text{document}\}|}$$

The total number of documents in the corpus

The number of documents where the term j appears

TF-IDF: An Example

Consider the words “apple” and “orange”

- “apple” appears 10 times in document 1 (d1)
- “orange” appears 20 in document 1 (d1)
- The corpus contains 20 documents.
- “apple” only appears in d1.
- “orange” appears in all 20 documents.

$$tf - idf(\text{“}apple\text{”}, d_1) = 10 \times \log_2 \frac{20}{1} = 43.22,$$

$$tf - idf(\text{“}orange\text{”}, d_1) = 20 \times \log_2 \frac{20}{20} = 0.$$

TF-IDF : An Example

- Documents:
  - d1: social media mining
  - d2: social media data
  - d3: financial market data
- TF values:

	social	*media*	*mining*	*data*	*financial*	*market*
d_1	*1*	*1*	*1*	*0*	*0*	*0*
d_2	*1*	*1*	*0*	*1*	*0*	*0*
d_3	*0*	*0*	*0*	*1*	*1*	*1*

$\log_2(3/2)=0.584$
$\log_2(3/1) = 1.584$

- TF-IDF

$$idf_j = \log_2 \frac{|D|}{|\{\text{document} \in D \mid j \in \text{document}\}|}$$

TF-IDF : An Example

- Documents:
  - d1: social media mining
  - d2: social media data
  - d3: financial market data
- TF values:

	social	*media*	*mining*	*data*	*financial*	*market*
d_1	*1*	*1*	*1*	*0*	*0*	*0*
d_2	*1*	*1*	*0*	*1*	*0*	*0*
d_3	*0*	*0*	*0*	*1*	*1*	*1*

$$
\begin{aligned}
idf_{social} &= \log_2(3/2) = 0.584 \\
idf_{media} &= \log_2(3/2) = 0.584 \\
idf_{mining} &= \log_2(3/1) = 1.584 \\
idf_{data} &= \log_2(3/2) = 0.584 \\
idf_{financial} &= \log_2(3/1) = 1.584 \\
idf_{market} &= \log_2(3/1) = 1.584.
\end{aligned}
$$

- TF-IDF

	social	*media*	*mining*	*data*	*financial*	*market*
d_1	*0.584*	*0.584*	*1.584*	*0*	*0*	*0*
d_2	*0.584*	*0.584*	*0*	*0.584*	*0*	*0*
d_3	*0*	*0*	*0*	*0.584*	*1.584*	*1.584*

Data Quality

When making data ready for data mining algorithms, data quality needs to be assured

- **Noise**
  - Noise is distortion of the data
- **Outliers**
  - Outliers are data points that are considerably different from other data points in the dataset
- **Missing Values**
  - Missing feature values in data instances
  - **To solve this problem:** *1) remove instances that have missing values 2) impute missing values, and 3) ignore missing values when running data mining algorithm*
- **Duplicate data**

Data Preprocessing

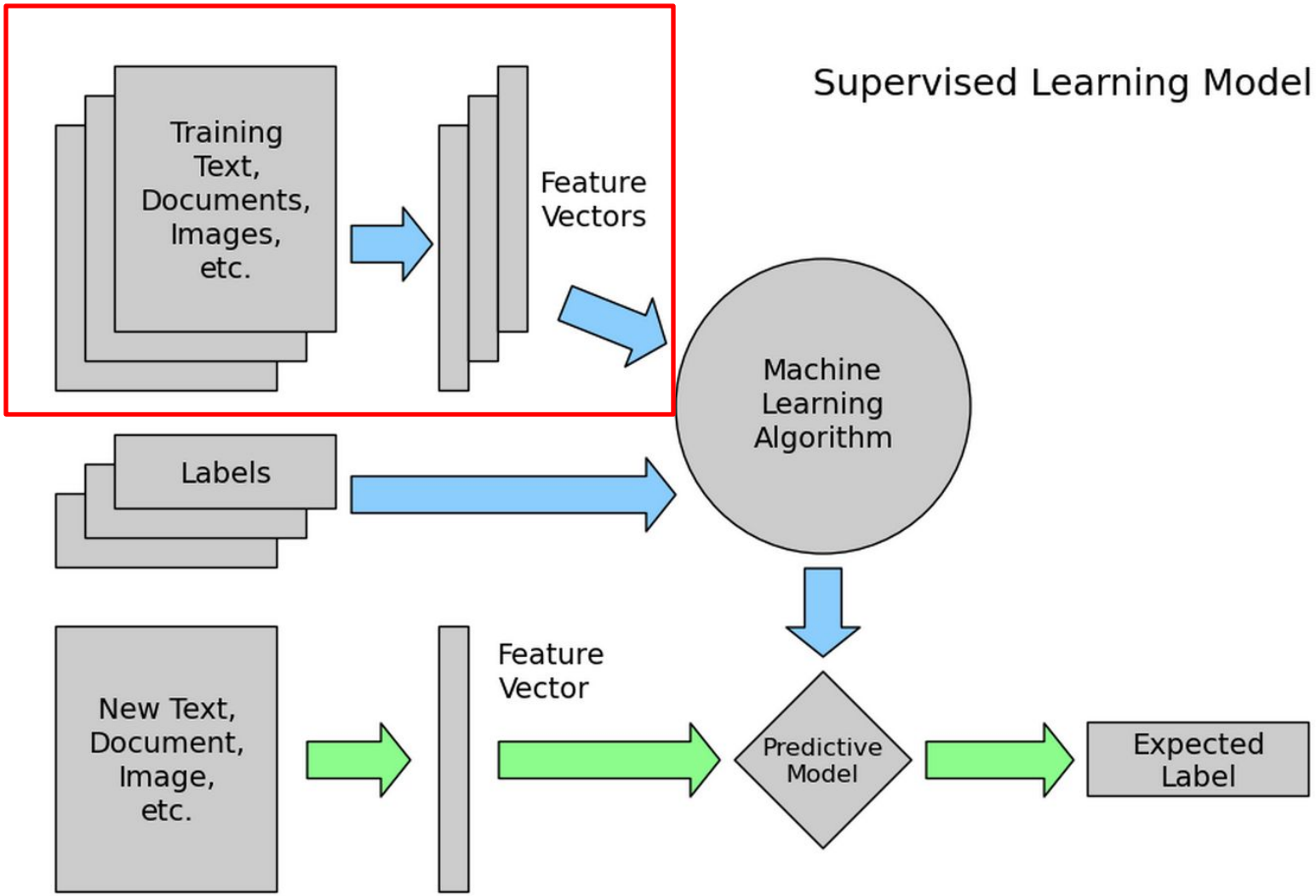
- **Aggregation**
  - It is performed when multiple features need to be combined into a single one or when the scale of the features change
  - Example: image width , image height -> image area (width x height)
- **Discretization**
  - From continuous values to discrete values
  - Example: money spent ($) -> {low, normal, high}
- **Feature Selection**
  - Choose relevant features
  - http://featureselection.asu.edu/
- **Feature Extraction**
  - Creating new features from original features
  - Often, more complicated than aggregation
- **Sampling**
  - Random Sampling
  - Sampling with or without replacement
  - Stratified Sampling: sample from each subgroup independently. useful for class imbalance
  - Social Network Sampling

Feature selection keeps a subset of the **original** features.
Feature extraction creates brand **new** ones.

Design Choices

- Finding the right features for your problem can be difficult.
- “Can I do…. ?”
- Yes!
- “Will it work?”
- Experiment
  - Early
  - Often

KDD Process


Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Data Mining Algorithms

- **Supervised Learning Algorithm**

Class attribute is available

  - **Classification (class attribute is discrete)**
    - Assign data into predefined classes
      - Spam Detection, fraudulent credit card detection
  - **Regression (class attribute takes real values)**
    - Predict a real value for a given data instance
      - Predict the price for a given house

- **Unsupervised Learning Algorithm**

Class attribute is **not** available

  - **Clustering**
    - Group similar items together into some clusters
      - Detect communities in a given social network