Introduction to Machine Learning and Data Mining

Version 2.0, August 2024

1 Data

1.1 Sources of Data

- Data explosion – society produces and stores huge amounts of data
- Business, science, medicine, economics, ...
- Current trend: Gather whatever data you can, whenever and wherever possible

1.2 Sources of Data

- Category data: Nominal scale
- Continuous data: numeric (including integer data, real number data)

1.3 Data Pre-processing

- Data cleaning (Not always, but simplest) scale
  - Noise: Irrelevant or erroneous information that may affect data analysis and model performance. (e.g. Random Noise, Systematic Noise,...)
  - Missing value
- Data processing
  - Data aggregation: The process of combining or aggregating multiple data points or data sets according to certain rules (e.g. Statistics, Rolling aggregation, ...)
  - Feature extraction: The process of extracting the most representative information (features) from raw data for use by machine learning models (e.g. Fourier Transform, Statistics, ...)
  - Feature subset selection: Select the most useful subset of features from the original dataset to improve the performance and efficiency of the machine learning model (e.g. Correlation Coefficient, Regularization Methods,...)
  - Converting features from one type to another: During data preprocessing, features are converted from one data type to another (e.g. Discretization, Grayscale Conversion, ...)
  - Normalization of feature values: Refers to scaling data features to a uniform scale, usually to improve the performance and stability of machine learning models
    - Min-Max Normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

* Zero - Score Standardization:

$$X' = \frac{X - \mu_X}{\sigma_X}$$

– Similarity measures: To figure out two variables $X = (X_1, X_2, X_3, \ldots, X_n)$ and $Y = (Y_1, Y_2, Y_3, \ldots, Y_n)$ are similar or not through Quantitative calculations.

* Simple Distance:
  · Euclidean:

$$d(X, Y) = \sqrt{(\sum_{i=1}^{n} |x_i - y_i|^2)}$$

  · Manhattan:

$$d(X, Y) = (\sum_{i=1}^{n} |x_i - y_i|)$$

  · Minkowski:

$$d(X, Y) = (\sum_{i=1}^{n} |x_i - y_i|^q)^{\frac{1}{q}}$$

  Note 1. In Minkowski Equation, the variable X and Y should be the $n \times q$ matrices, which means for each x_i, y_i, they should be like:

$$x_i = (x_{1i}, x_{2i}, x_{3i}, \ldots, x_{qi})^T$$

* Distance between categorical variables:
  · Hamming:

$$H(X, Y) = \sum_{i=1}^{n} |x_i - y_i|$$

  · SMC:

$$SMC(X, Y) = \frac{n_{11} + n_{00}}{n_{11} + n_{00} + n_{10} + n_{01}}$$

  · Jaccard coefficient:

$$J(X, Y) = \frac{|A \cap B|}{|A \cup B|}$$

* Cosine similarity:

$$cos(X, Y) = \frac{X^T Y}{|X||Y|}$$

* Correlation:

$$Corr(X, Y) = \frac{cov(X, Y)}{sd(X)sd(Y)}$$

2 Machine Learning and Data Mining

2.1 Machine Learning and Data Mining are concerned with finding patterns in data

- Row data is useless
- Patterns should be meaningful, useful and actionable
- The process is automatic or semi - automatic

2.2 Some differences between Machine Learning and Data Mining

- Machine Learning is a core part of Artificial Intelligence
- Most of the algorithms used for Data Mining have developed in Machine Learning
- Data Mining only deals with large and multidimensional data
- Data Mining can be seen as applied Machine Learning (use Machine Learning to do Data Mining)

3 Main tasks in Machine Learning

3.1 Supervised learning – classification and regression

- Classification: the variable to be predicted is categorical
- Regression: the variable to be predicted is numeric

3.2 Unsupervised learning – clustering

Clustering is a fundamental technique in unsupervised learning, where the goal is to group a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups.

Note 2. Distinguishing between classification and clustering:

- Classification is to classify new data into existing groups based on experience when the existing data is classified.
- Clustering is to provide a possible classification standard when the classification of existing data is unknown.

3.3 Association rule mining

A technique for discovering interesting associations or patterns in large-scale data sets. It is widely used in market basket analysis, recommendation systems, web page mining, and other fields. The classic task of association rule mining is to find sets of items that often appear together from a set of transaction data and generate rules to represent the relationship between these sets of items.

3.4 Reinforcement learning

By interacting with the environment, it learns how to take actions to maximize the accumulated rewards. It has a wide range of applications in robot control, game AI, autonomous driving, ... The core idea of reinforcement learning is that the agent learns a strategy through trial and error to take the best action in different states.

3.5 Outlier detection

The process of identifying data points in a dataset that significantly deviate from the majority of the data. These outliers can represent errors, anomalies, or rare events that are of interest in various applications such as fraud detection, network security, and quality control.