An Introduction To Data Mining

COSC757: Data Mining

Devere Anthony Weaver

---

1 What is Data Mining?

Definition 1. ***Data mining*** *is the process of discovering useful patterns and trends in large data sets.*

---

2 The Cross-Industry Standard Practice for Data Mining (CRISP-DM)

CRISP-DM provides a nonproprietary and freely available standard process for fitting data mining into the general problem solving strategy of a business or research unit.

2.1 CRISP-DM: The Six Phases

The six phases of CRISP-DM are:

1. Business/Research Understanding Phase
    - clearly enunciate the project objectives and requirements
    - translate these requirements into the formulation of a data mining problem
    - prepare a preliminary strategy for achieving these objectives
2. Data Understanding Phase
    - collect data
    - use EDA to familiarize yourself with the data
    - evaluate the quality of the data
    - select any interesting subsets if desired
3. Data Preparation Phase
    - select the cases and variables you want to analyze and that are appropriate for the analysis at hand
    - perform any transformations as needed
    - clean the raw data so that it is ready for the modeling tools
4. Modeling Phase
    - select and apply appropriate modeling techniques
    - calibrate model settings to optimize results
5. Evaluation Phase
6. Deployment Phase

---

3 What Tasks Can Data Mining Accomplish?

3.1 Description

Sometimes researchers and analysts are simply trying to find ways to describe patterns and trends lying within the data.

Data mining models should be as transparent as possible. That is, the results of the data mining model should describe clear patterns that are amenable to intuitive interpretation and explanation.

3.2 Estimation

In estimation, we approximate the value of a numeric target variable using a set of numeric and/or categorical predictor variables.

3.3 Classification

Classification is similar to estimation, except that the target variable is categorical rather than numeric.

3.4 Clustering

Clustering refers to the grouping of records, observations, or cases into classes of similar objects.

A cluster is a collection of records that are similar to one another, and dissimilar to records in other clusters. Clustering differs from classification in that there is no target variable for clustering.

Clustering algorithms seek to segment the whole data set into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized, and the similarity to records outside of this cluster is minimized.

Clustering is often performed as a preliminary step in a data mining process, with the resulting clusters being used as further inputs into a different technique.

3.5 Association

The association task for data mining is the job of finding which attributes "go together".