# 1 Lecture 1

Course Overview

1. Introduction to Data Mining
2. Important data mining primitives:
   (a) Classification
   (b) Clustering
   (c) Association Rules
   (d) Sequential Rules
   (e) Anomaly Detection
3. Commercial and Scientific Applications

DataMining is a Non trivial extraction of nuggets from large amounts of data
Data Mining Tasks:

1. Prediction methods (Use some variables to predict unknown or future values of the same or other variables)
2. Description methods (Find human interpretable patterns that describe data)

Data Mining Tasks

1. Classification (predictive)
2. Clustering (descriptive)
3. Association Rule Discovery (descriptive)
4. Sequential Pattern Discovery (descriptive)
5. Regression (predictive)

6. Deviation Detection (predictive)

Why is Data Mining prevalent?

1. Lots of data is collected and stored in data warehouses
    (a) Business. Wal-Mart logs nearly 20 million transactions per day
    (b) Astronomy. Telescope collecting large amounts of data (SDSS)
    (c) Space. NASA is collecting peta bytes of data from satellites
    (d) Physics. High energy physics experiments are expected to generate 100 to 1000 tera bytes in the next decade
2. Quality and richness of data collected in improving
    (a) Retailers. Scanner data is much more accurate than other means
    (b) E-Commerce. Rich data on consumer browsing
    (c) Science. Accuracy of sensors is improving
3. The gap between data and analysts is increasing
    (a) Hidden information is not always evident
    (b) High cost of human labor
    (c) Much of data is never analyzed at all

Origins of Data Mining

1. Drawn ideas from
    (a) Machine Learning,
    (b) Pattern Recognition,
    (c) Statistics, and
    (d) Database systems for applications that have

i. Enormity of data

ii. High dimensionality of data

iii. Heterogeneous data

iv. Unstructured data

Regression

1. Predict the value of a given continuous valued variable based on the values of other variables, assuming a linear or non-linear model of dependency

Association Rule Discovery: Definition

1. Given a set of records, each of which contain some number of items from a given collection: Produce dependency rules which will predict occurrence of an item based on occurrences of other items

Classification: Definition

1. Given a set of records (called the training set). Each record contains a set of attributes. One of the attributes is the class

2. Find a model for the class attribute as a function of the values of other attributes

3. Goal: Previously unseen records should be assigned to a class as accurately as possible. Usually, the given data set is divided into training and test set, with training set used to build the model and test set used to validate it. The accuracy of the model is determined on the test set.

Clustering. Determine object groupings such that objects within the same cluster are similar to each other, while objects in different groups are not

1. Typically objects are represented by data points in a multidimensional space with each dimension corresponding to one or more attributes. Clustering problem in this case reduces to the following: Given a set of data points, each having a set of attributes, and a similarity measure, find clusters such that

(a) Data points in one cluster are more similar to one another

(b) Data points in separate clusters are less similar to one another

(c) Similarity measures: Euclidean distance if attributes are continuous; Other problem-specific measures

Deviation / Anomaly Detection

1. Some data objects do not comply with the general behavior or model of the data. Data objects that are different from or inconsistent with the remaining set are called outliers

2. Outliers can be caused by measurement or execution error. Or they represent some kind of fraudulent activity.

3. Goal of Deviation / Anomaly Detection is to detect significant deviations from normal behavior

Deviation / Anomaly Detection: Definition

1. Given a set of n data points or objects, and k, the expected number of outliers, find the top k objects that considerably dissimilar, exceptional or inconsistent with the remaining data

2. This can be viewed as two sub problems

   (a) Define what data can be considered as inconsistent in a given data set

   (b) Find an efficient method to mine the outliers so defined