Data Structures

Terminology and Concepts (1)

- Data
  - Dataset is a set of Data. A set implies a commonality. The commonality is expressed as a type or a relation.
  - A data type provides structure and meaning to the data. Just like there is no such thing as un-structured data, there is no such thing as un-typed data. Data can be insufficiently typed and structured.
- Rectangular Data
  - Datasets are often 2D matrices, which are organized into rows and columns. The column and row order is not important .
  - Columns are named with a header; A columns may be also referred to as an attribute or field. The number of columns is often called the dimensionality of the data.
  - Rows are not named. A row is often referred to as a case or observation. Number of rows in a category is called support.
- Data dimensionality
  - A data frame or a table can be considered a sparse multi-dimensional matrix
  - The dimensionality for un-supervised learning is #columns
  - The dimensionality for supervised learning is #columns - 1 because one column represents the value and not the dimension. This structure is very similar to a star schema

Terminology and Concepts (2)

- Predictive Analytics (Machine Learning , Artificial Intelligence)
  - Algorithms (often called Methods)
    - Supervised Learning
      - Classification
      - Estimation
    - Unsupervised Learning
      - Clustering
      - Association (Market-basket analysis)
      - Anomaly detection
    - Forecasting (Time Series)

Terminology and Concepts (3)

- Supervised Learning Algorithms
  - Classification Algorithms predict classes or categories
    - Logistic Regression (Deterministic)
    - Decision Trees (Deterministic)
    - Naïve Bayes (Deterministic)
    - Neural Net (Non-Deterministic)
  - Estimation Algorithms predict continuous (numeric) values
    - Generalized Linear Modeling abbreviated: GLM (Deterministic)
      - Linear Regression
      - Logistic Regression
    - Regression Trees (Deterministic)
    - Neural Net (Non-Deterministic)

Terminology and Concepts (4)

- Un-Supervised Learning Algorithms
  - Segmentation Algorithms, also called Clustering, create clusters or segments. These clusters can be thought of as categories.
    - Mixture of Gaussians aka Probabilistic (Deterministic)
    - Hierarchical (Deterministic)
    - K-Means (Non-Deterministic)
  - Association Algorithms associate or link items by a common attribute called the transaction ID.
    - Market Basket Analysis (Deterministic)
    - Affinity Analysis (Deterministic)
  - Anomaly Detection is used to find unusual or anomalous data like outliers

Terminology and Concepts (5)

- Forecasting (Time Series) is used to estimate future values based on past behaviors.
  - ARIMA / Auto ARIMA
  - Survival Analysis

Major types of Data Sets

- **Univariate**
- **Rectangular**
- **Time Series**
- **Nested**
- **Graphs (later in the course)**

Univariate (1)

- A collection of data. The data do not have a particular order. Example: Students' age. This type of data is often (mistakenly) called unstructured data, especially when the values are strings of indeterminate length. (Ragged Array)
- Example usage: anomaly detection.

Univariate (2)

Parent Income
40,000
53,000
60,000

Rectangular Data (1)

- The data set has columns and rows. Each cell has a value or is null.
- A Rectangular dataset is often called a matrix, data frame, or table.
- Example usage: classifications and estimations

Rectangular Data (2)

- Columns have descriptive headers like: Name, Age, Height, Weight of each student.
- Columns are also called attributes and fields.
- All values within a column have the same data type

Rectangular Data (3)

- Rows generally do not have names.   If a row has a name, then the names could be considered another column.
- Rows are also called observations or cases
- The number of rows in a category is called support.

Rectangular Data (4)

ID	IQ	Parent Income	Moral Support	Gender	College Plans
835	107	40,000	Yes	Female	Applied
016	99	53,000	Yes	Male	Applied
490	105	60,000	No	Male	Did not apply

Time Series (1)

- A rectangular data set where the independent variable is time. The observations are sorted by time.
- Example usage: forecasting.

Time Series (2)

Date	Red Wine Sales	White Wine Sales	Rose Sales
1/22/13	$103.00	$300.50	$19.00
1/23/13	$35.50	$204.00	$44.00
1/24/13	$217.50	$74.50	$80.00

Nested (1)

- A rectangular data set where the rows have a table. Such a table can have a flat representation.
- Example usage: associations (shopping basket analyses).

Nested (2)

Transaction ID	**Item**
1	Milk
	Sugar
2	Lumber
3	Milk
	Sugar
	Flour

Nested (3)

Transaction ID	**Item**
1	Milk
1	Sugar
2	Lumber
3	Milk
3	Sugar
3	Flour

Nested (4)

Transaction ID	**Item**
1	Milk
1	Sugar
2	Lumber
3	Milk
3	Sugar
3	Flour

Data Structures