Notes 2/12/2019

What is data exploration?

* A Preliminary exploration of the data to better understand its characteristics.
* Key motivations of data exploration include
  + Helping to select the right tool for preprocessing or analysis
  + Making use of humans’ abilities to recognize patterns
    - People can recognize patterns not captures by data analysis tools
  + Related to the area of Exploratory Data Analysis (EDA)
    - Created by statistician John

Techniques Used in Data Exploration

* In EDA, as originally defined by Turkey
  + The focus was on visualization.
  + Clustering and anomaly detection were viewed as exploratory techniques
  + In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory
* In our discussion of data exploration, we focus on
  + Summary statistics
  + Visualization
  + Online Analytical Processing (OLAP) [overview]

Iris Sample Data Set

* Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  + Can be obtained from the UCI Machine Learning Repository <http://www.ics.uci.edu/!mlearn/ML>

Summary Statistics

* Summary statistics are numbers that summarize properties of the data.
  + Summarized properties include frequency, location and spread
    - Examples: location – mean

Standard deviation

Frequency and Mode

Percentiles

* For continuous data, the notion of a percentile is more useful.
* Given an ordinal

Measures of Location: Mean and Median

* The mean is the most common measure of the location of a set of points.
* However, the mean is very

Measures of Spread: Range and Variance

* Range is the difference

Visualization

* Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.
* Visualization of data is one of.

Example: Sea Surface Temperature

* The following shows the Sea Surface Temperature (SST) for July 1982

Arrangement

* Is the placement of visual elements within a display
* Can make a large difference in how easy it is to understand the data.
* Example:

Selection

* Is the elimination or the?

Visualization Techniques: Histograms

Two-Dimensional Histograms

* Show the joint distribution of the values of two attributes.
* Example: petal width and petal length
  + What does this tell us?

Visualization Techniques: Box Plots

* Box Plots
  + Invented by J. Tukey
  + Another

Example of Box Plots

* Box plots can be used to compare attributes

Visualization Techniques: Matrix Plots

* Matrix plots
  + Can

Visualization of the Iris Data Matrix

Visualization Techniques: Parallel Coordinates

* Parallel Coordinates

Parallel Coordinates Plots for Iris Data

Other Visualization Techniques

* Star Plots
  + Similar approach to parallel coordinates, but axes radiate from a central point
  + The line connecting the values of an object is a polygon
* Chernoff Faces
  + Approach created by Herman Chernoff
  + This approach associates each attribute with a characteristic of a face

Creating a Multidimensional Array

* Converting tabular data into a multidimensional array:
  + Identify which attributes are to be the dimensions and which attribute is to be the target attribute

Example: Iris data (continued)

* Slices of the multidimensional array are shown by the following cross-tabulations
* What do these tables tell us?

OLAP Operations: Data Cube

* The key operation of a OLAP

Data Cube Example (continued)

* The following figure table shows one of the two multidimensional array by specifying a specific value for one or more dimensions.
* Dicing involves selecting a subset of cells by specifying a range of attribute values.
  + This is equivalent to

OLAP Operations: Slicing and Dicing

* Slicing is selecting a group of cells from the entire multidimensional array by specifying a specific value for one ormore dimensions.
* Diving involves selecting a subset of cells by specifying a range of attribute values.
  + This is equivalent to defining a subarray from the complete array.
* In practice, both operations can also be accompanied.

OLAP Operations: Roll-up and Drill-down

* Attribute values often have a hierarchical structure.
  + Each date is associated with a year, month, and week.
  + A location is associated with a continent, country, state(province, etc), and city.

ITCS

Classification: Definition

* Given a collection of records (training set)
  + Each record is by characterized by a tuple (x,y), where x is the attribute set and y is the class label
    - X: attribute, predictor, independent variable, input

Examples of Classification Task

|  |  |  |
| --- | --- | --- |
| Task | Attribute set, x | Class label, y |
| Categorizing email messages | Features extracted from email message header |  |

General Approach for Building Classification Model

* Learning algorithm 🡪

Classification Techniques

* Base Classifiers
  + Decision Tree based Methods
  + Rule-based Methods
  + Nearest-neighbor
  + Support Vector Machines
  + Naïve Bayes
  + Neural Networks (Advanced)
  + Deep Learning (Advanced)
  + Bayesian Belief Networks (Advanced)
* Ensemble Classifiers

Example of a Decision Tree

Another Example of Decision Tree

Continuous Attributes: Computing Gini Index…

* For efficient computation: for each attribute,
  + Sort the attribute on values
  + Linearly scan these values, each time updating the count matrix and computing gini index
  + Choose the split position that has the least gini index