Midterm Review

March 5, 2019

0.0.1 Chapter 1: Introduction

Why data mining?

- The proliferation of devices collecting data, decreased cost to store data, and increased speed and ability to process data has increased the opportunity to turn large sets of data into knowledge with commercial and societal benefits.
- Example: Turning Google's search data on flu symptoms into flu trends faster than traditional reporting systems.

What is data mining?

- Data mining the process of discovering interesting patterns from massive amounts of data. It is extracting non-trivial, previously unknown knowledge from large quantities of data by automated or semi-automated means.
- The process is:
  - *preprocessing*: feature selection, dimensionality reduction, normalization, data subsetting
  - *mining*: methods applied to extract patterns
  - *postprocessing*: filtering patterns, visualization, interpretation

What kind of data can be mined?

- While nearly any type of data can be mined, relational and transactional data in databases are common sources.

What kind of patterns can be mined?

- Data mining tasks can be classified as:
  - *descriptive*: find human interpretable patterns that describe the data.
  - *predictive*: Use variables to predict unknown future values of other variables.
- Descriptive tasks (descriptive, supervised):
  - *data characterization*: summarizing the data of the class in general terms. For example, summarize the general characteristics of customers who spend more than $5,000 / yr "big spenders". The general profile may include income, credit, age, etc.

  - *data discrimination*: similar to data characterization, but comparing two classes. For example, how does the general profile of a "frequent shopper" differ from an "infrequent shopper".

- Frequent patterns, associations and correlations (predictive, supervised):
  - *frequent patterns*: patterns that occur frequently in the data
  - *association analysis*: a predictive attribute such as 'buy' that repeats in the data.

- Classification and regression (predictive, supervised):
  - *classification*: the process of finding a model that distinguishes data classes or concepts. The process relies on labeled data. Decision trees, neural networks, naive Bayes, k-nearest neighbor are common classification methods. Example: model for predicting creditworthiness (decision tree), classifying a breed of dog (neural network).
  - *regression*: similar to classification, but models continuous valued functions, used to predict missing numerical value. Example: predict sales based on ad spend, stock price, wind speed, etc.

- Cluster Analysis (descriptive, unsupervised):
  - *cluster analysis*: finding groups of objects that minimize intra-cluster distance and maximize inter-cluster distance. Does not consult labels. Examples include market segmentation and text similarity analysis.

- Outlier analysis:
  - Data that does not comply with the model. In some situations, such as fraud detection, network invasion, deforestation: anomaly mining is more interesting than the regular occurrences.

- Not all patterns are interesting. A pattern is interesting if it is: easily understood, valid on new data, useful and new. Association analysis examples that may be interesting: market basket analysis, alarm diagnosis, medical diagnosis.

Which technologies are used?

- Data mining has incorporated many techniques from other domains such as stats, ML, database systems. Data mining is closely related to all of these disciplines.
- It is not uniquely a transformation or application of one of these disciplines but rather an evolution of all of them in response to the need for "effective, scalable, and flexible data analysis in our society".
- As data sets become massive and diverse, data mining techniques must be computationally efficient, handle various data types and produce effective results. This is distinct from the more constrained and controlled worlds of traditional disciplines such as statistics.

Major issues in data mining

- The main challenge presented when mining a huge amount of data is the efficiency and scalability of the data mining algorithm used to store and process the data effectively. Finding an algorithm that can process the data effectively, efficiently and increasingly in real-time.

- Another challenge is dealing with the increasing diversity of data types. Data is no longer of a uniform data type and stored neatly in a relational database. It is diverse, dynamic and distributed. One way to deal with these challenges is to utilize parallel and distributed data-intensive mining algorithms.

0.0.2 Chapter 2: Getting to Know Your Data

Data objects and attribute types

- Data is a collection of objects and their attributes.
- A data attribute is a property or characteristic of an object (eye color, income).
- A collection of attributes describe an object.
- Attribute values can be nominal, binary, ordinal or numeric.
  - *nominal*: categorical, the values have no meaningful order.
  - *binary*: two possible outcomes.
    - symmetric: equally likely/valuable such as gender
    - asymmetric: one value, such as a positive medical test, is more valuable
  - *ordinal*: categorical, the values have a rank order.
  - *numeric*: quantitative, measurable value.
    - interval-scaled: relative, no true zero exists (temp, year)
    - ratio-scaled: absolute, true zero exists (temp in K, length)
- The type of an attribute depends on the properties/operations it possesses:
  - Nominal attribute: distinctness (=, !=)
  - Ordinal attribute: distinctness & order (<, >)
  - Interval attribute: distinctness, order & meaningful differences (+, -)
  - Ratio attribute: distinctness, order & meaningful differences and ratios (*, /)
- Discrete vs. continuous variables:
  - Discrete values are finite or countably infinite (age 0-110)
  - Continuous values have real numbers as attributes, typically represented by floating point.

Types of data sets

- Important characteristics of data include:
  - *Dimensionality*: number of attributes
  - *Sparsity*: Only presence counts
  - *Resolution*: Patterns depend on the scale
  - *Size*: Type of analysis may depend on size
- Types of data sets:
  - *Record*: A collection of records, each with a fixed set of attributes. This can be represented by a matrix or a vector. Transaction data is a special type of record data.
  - *Graph*: Represented by edges and vertices, and may represent things like molecular structures or websites.
  - *Ordered*: May be a sequence of transactions, temporal data, genomic sequence, etc.

Data quality issues

- Noise and outliers: For objects, extraneous objects; for attributes: modification of original values. Outliers are objects with considerably different characteristics than the rest of the data set.
- Missing values: Information is either not collected or not relevant in all cases.
- Duplicate data: same person with multiple entries

Basic statistical descriptions of data

- Measures of central tendency
  - *mean*: average
  - *median*: middle
  - *mode*: most common
  - *midrange*: average of max/min
  - *negative skew*: long left tail
  - *positive skew*: long right tail

- Measures of dispersion of data
  - *quartiles*: Q1(25th), Q3(75th)
  - *interquartile range*: Q3 - Q1
  - *five number summary*: min, Q1, median, Q3, max
  - *boxplot*: outliers marked (>1.5x IQR), whiskers are min/max, quartiles are marked
  - *variance*: sum(x^2)/n - mean^2
  - *standard deviation*: sqrt(variance), measures the spread around the mean
    - mean +/- 1 sigma: 68% under normal curve
    - mean +/- 2 sigma: 95% under normal curve
    - mean +/- 3 sigma: 99.7% under normal curve

Graphic displays of basic statistical descriptions of data

- Univariate:
  - Quantile plot: Sort in ascending order, calculate the fraction of each data point as: f = (i-0.5)/N. Plot x = f-value, y = data
  - Quantile-quantile plot: Calculate the quantile of each, plot the data at each f-value against each other.
  - Histogram: plot the data distribution in bins

- Bivariate:
  - Scatter plot: treat each pair as (x, y)
  - Correlation: positive, negative or no correlation

Similarity and dissimilarity measures

- Dissimilarity measures:
  - *nominal*: [0,1] are 0 for similar objects, 1 for dissimilar objects.