Notes Jan 29, 2019

What is Data Mining?

* Many Definitions
  + Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  + Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.
  + Input Data 🡪

What is (not) Data Mining?

* What is not Data Mining?
  + Look up phone number in phone directory
  + Query a web search engine.

Origins of Data Mining

* Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems.
* Traditional techniques may be unsuitable due to data that is
  + Large-scale
  + High dimensional
  + Heterogeneous
  + Complex
  + Distributed

Data Mining Tasks

* Prediction Methods
  + Use some variables to predict unknown or future values of other variables
* Description Methods
  + Find human-interpretable patterns that describe the data.

Predictive Modeling: Classification

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

Classification Example

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Categorical | Categorical | Quantitative | Class |
| Tid | Employed | Level of Education | # years at present address | Credit Worthy |
| 1 | Yes | Graduate | 5 | Yes |
| 2 | Yes | High School |  |  |
| 3 | Yes | Undergrad |  |  |
| 4 | No | High School |  |  |
| … | … | … |  |  |

Examples of Classification Task

* Classifying credit card transactions as legitimate or fraudulent.
* Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data.
* Categorizing news stories as finance, weather, entertainment, sports, etc.

Classification: Application 1

* Fraud Detection
  + Goal: Predict fraudulent cases in credit card transactions.
  + Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      * When does a customer buy, what does he buy, how often he pays on time, etc.?
      * Label past transactions as fraud or fair transactions. This forms the class attribute.
      * Learn a model for the class of the transactions.
      * Use this model to detect fraud.

Classification: Application 2

Classifying Galaxies

* Early, Intermediate, Late
* Class

Regression

* Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
* Extensively studies in statistics, neural network fields.

Clustering

* Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

Applications of Cluster Analysis

* Understanding
  + Custom profiting for targeted marketing
  + Group related documents for browsing

Clustering: Application 1

* Market Segmentation:
  + Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  + Approach:
    - Collect di

Clustering: Application 2

* Document Clustering:
  + Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  + Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Association Rule Discovery: Definition

* 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.

|  |  |
| --- | --- |
| TID | Items |
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

Association Analysis: Applications

* Market-basket analysis
  + Rules are used

Deviation/Anomaly/Change Detection

* Detect significant deviations from normal behavior.
* Applications:
  + Credit Card Fraud Detection
  + Network Intrusion Detection
  + Identify anomalous behavior from sensor networks for monitoring and surveilling.

Motivating Challenges

* Scalability
* High Dimensionality
* Heterogeneous and Complex Data
* Data Ownership and Distribution
* Non-traditional Analysis

Outline

* Attributes and Objects
* Types of Data
* Data Quality
* Similarity and Distance
* Data Preprocessing (will begin next class with this after Python tutorial)

Collection of data objects and their attributes

An attribute is a property or characteristic of an object

* Examples: eye color of a person, temperature, etc.
* Attribute is also known as variable, field, characteristic, dimension, or feature.

A collection of attributes describes an object

* Object is also known as record, point, case, sample, entity, or instance.

A More Complete View of Data

* Data may have parts
* The different parts of the data may have relationships
* More generally, data may have structure
* Data can be incomplete
* We will discuss this in more detail later

Attribute Values

* Attribute values are numbers or symbols assigned to an attribute for a particular object
* Distinction between attributes and attribute values
  + Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  + Different attributes can be mapped to the same set of values
    - Example:

The way you measure an attribute may not match the attributes properties.

Types of Attributes

* There are different types of attributes
  + Nominal
    - Examples: ID numbers, eye color, zip codes
  + Ordinal
    - Examples: ranking (e.g., taste of potato chips on a scale of 1-10), grades, height {tall, medium, short}

Properties of Attribute Values

* The type of an attribute depends on which of the following properties/operations it possesses:
  + Distinctness: = ≠
  + Order: < >

Difference Between Ratio and Interval

* Is it physically meaningful to say that a temperature of 10° is twice that of 5° on:
  + The Celsius scale?
  + The Fahrenheit scale?
  + The Kelvin scale?
* Consider measuring the height above average
  + If Bill’s height is three inches above average and Bob’s Height is six inches above average.

Discrete and Continuous

Asymmetric Attributes

* Only presence (a non-zero attribute value) is regarded as important
  + Words present in documents
  + Items present in customer transactions
* If we met a friend in the grocery store would we ever say the following?
  + “I see our purchases are very similar since we didn’t buy most of the same things.”
* We need two asymmetric binary attributes to represent one ordinary binary attribute
  + Association analysis

Key Messages for Attribute Types

* The types of operations you choose should be meaningful for the type of data you have
  + Distinctness, order, meaningful intervals, and meaningful ratios are only four properties of data.
  + The data type you see – often numbers or strings.

Types of data sets

* Record
  + Data Matrix
  + Document Data
  + Transaction Data
* Graph
  + World Wide Web
  + Molecular Structures
* Ordered
  + Spatial Data
  + Temporal Data
  + Sequential Data
  + Genetic Sequence Data

Important Characteristics of Data

* Dimensionality (number of attributes)
  + High dimensional data brings a number of challenges
* Sparsity
  + Only presence counts
* Resolution
  + Patterns depend on the scale
* Size
  + Type of analysis may depend on size of data

Data Matrix

* If data objects have the same fixed set of numeric

Document Data

* Each document

Ordered Data

* Spatio-Temporal Data
* - Average Monthly Temperature of land and ocean

Data Quality

* Poor data quality negatively affects many data processing efforts
  + “The most important point is that poor data quality is an unfolding disaster”
    - Poor data quality costs the typical company at least.

Data Quality …

* What kinds of data quality problems?
* How can we detect problems with the data?
* What

Noise

* For objects, noise is an extraneous object
* For attributes, noise refers to modification of original values
  + Examples: distortion of a person’s voice when talking on a poor phone and “snow” on television screen.

Outliers

* Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set.
  + Case 1: Outliers are noise that interferes with data analysis
  + Case 2: Outliers are the

Missing Values

* Reasons for missing values
  + Information is not collected (e.g., people decline to give their age and weight)
  + Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
* Handling missing values
  + Eliminate data objects or variables
  + Estimate missing values
    - Example: time series of temperature
    - Example: census

Missing Values…

* Missing completely at random (MCAR)
  + Missingness of a value is independent of attributes
  + Fill in values based on the attribute
  + Analysis may be unbiased overall
* Missing at Random (MAR
  + Missingness is related to other variables
  + Fill in values based other

Duplicate Data

* Data set may include data objects that are duplicates, or almost duplicates of one another
  + Major issue when merging data from heterogeneous

Similarity and Dissimilarity Measures

* Similarity measure
  + Numerical measure of how alike two data objects are.
  + Is higher when objects are more alike.
  + Often falls in the range [0,1]
* Dissimilarity measure
  + Numerical measures of how different two data objects are

Similarity/Dissimilarity for Simple Attributes

* The following table shows the similarity and dissimilarity between two objects, x and y, with respect to a single, simple attribute

Euclidean Distance

* Euclidean Distance
  + d(x, y) = sqrt

Euclidean Distance

Minkowski Distance

* Minkowski Distance is a generalization of Euclidean Distance

Minkowski Distance: Examples

* r = 1. City block (Manhattan, taxicab, L, norm) distance.
  + A common example of this is the Hamming distance, which is just the number

Minkowski Distance

Common Properties of a Distance

* Distances, such as the Euclidean distance, have some well-known properties.
  + D(x,y)

Similarity Between Binary Vectors

* Common situation is that objects, p and q, have only binary attributes
* Compute similarities using the following quantities
  + f­­01 = 2 (the number of attributes where p was 0 and q was 1)
  + f­10 = 1 (the number of attributes where p was 1 and q was 0)
  + f00 = 7 (the number

Cosine Similarity

* If d1 and d2 are two document vectors, then cos (d1, d2) = <d1,d2>/||d1|| ||d2||, where <d1,

Extended Jaccard Coefficient (Tanimoto)

* Variation of Jaccard

Correlation measures the linear relationship between objects

* Corr (x, y) = covariance (x, y) / (standard\_

Visually Evaluating Correlation

Drawbacks of Correlation

* X = (-3,-2,-1,0,1,2,3)
* Y = (9,4,1,0,1,4,9)

Comparison of Proximity Measures

Information Based Measures

* Information theory is a well-developed and fundamental disciple with broad applications
* Some similarity measures are based on information theory
  + Mutual information in various versions
  + Maximal

Information and Probability

* Information relates to possible outcomes of an event
  + Transmission of a message, flip of a coin, or measurement of a piece of data

Entropy

* For
  + A variable (event), X,
  + With n possible values (outcomes), x1, x2, xn

Entropy Examples

* For a coin with a probability p of heads and probability q = 1 – p of tails
  + H = -p log­2p – q log2 q
  + For p = 0.5, q = 0.5 (fair coin) H = 1
  + For p = 1 or q = 1, H = 0
* What is the entropy of a fair four-sided-die?

Python

Python Packages for Data Mining

* Numpy(for multi-dimensional array operations)
* Pandas (for handling datagrams like in SQL and Excel)
* Matplotlib & seaborn (for data visualizations)
* Scikit-learn (for machine learning algorithms)
* Scipy (for scientific Python programming using Fourier transforms, linear algebra, eigenvalue problems, etc.)
* Nltk (for natural language processing)

Anaconda Distribution

* Open source software for fastest easiest Python and R data science and machine learning
* It supports Linux, Windows, and MAC OS X

Jupyter Notebook Demo

* Check the tutorial at: https://www.datacamp.com/community/tutorials/tutorial-jupyter-notebook