CSE 4221: Data Mining

Riyad Morshed Shoeb
2022

Data Mining and Applications

Why Data Mining is important?

The explosively growing, widely available, and gigantic body of data makes our time truly the data age. Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining. Data mining turns a large collection of data into knowledge.

The abundance of data, coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation. The fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools. As a result, data collected in large data repositories become **"data tombs"**—data archives that are seldom visited. Consequently, important decisions are often made based not on the information-rich data stored in data repositories but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. Efforts have been made to develop expert system and knowledge-based technologies, which typically rely on users or domain experts to manually input knowledge into knowledge bases. Unfortunately, however, the manual knowledge input procedure is prone to biases and errors and is extremely costly and time consuming. The widening gap between data and information calls for the systematic development of data mining tools that can turn data tombs into **"golden nuggets"** of knowledge.

What is Data Mining?

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Advantages, Disadvantages, Applications of Data Mining

Many other terms have a similar meaning to data mining, for example, knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, **Knowledge Discovery from Data (KDD)**, while others view data mining as merely an essential step in the process of knowledge discovery. The knowledge discovery process is an iterative sequence of the following steps:

1. **Data cleaning** to remove noise and inconsistent data.
2. **Data integration** where multiple data sources may be combined.
3. **Data selection** where data relevant to the analysis task are retrieved from the database.
4. **Data transformation** where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. **Data mining** an essential process where intelligent methods are applied to extract data patterns.
6. **Pattern evaluation** to identify the truly interesting patterns representing knowledge based on interestingness measures.
7. **Knowledge presentation** where visualization and knowledge representation techniques are used to present mined knowledge to users.

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

Data mining involves six common classes of tasks.

- Anomaly detection (Outlier/change/deviation detection)
- Association rule learning (Dependency modeling)
- Clustering
- Classification
- Regression
- Summarization

Is everything data mining?

???

What kinds of Data Can Be Mined?

As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data, data warehouse data, and transactional data.

Relational Database A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data. A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes and usually stores a large set of tuples. Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values. A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships.

Data Warehouse is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making. Data warehouse technology includes data cleaning, data integration, and online analytical processing (OLAP). Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing. There are three types of data warehouse: Enterprise data warehouse, Data Mart and Virtual Warehouse.

To facilitate decision making, the data in a data warehouse are organized around major subjects. The data are stored to provide information from a historical perspective, and are typically summarized. A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure. A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data.

Transactional Databases is a collection of data organized by time stamps, date, etc to represent transaction in databases. Each record in a transactional database captures a transaction. This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed. Highly flexible system where users can modify information without changing any sensitive information.

Advanced data and Information systems

data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW

There are a number of data mining functionalities. These include characterization and discrimination, the mining of frequent patterns, associations, and correlations, classification and regression, clustering analysis, and outlier analysis. Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: descriptive and predictive. **Descriptive mining** tasks characterize properties of the data in a target data set. **Predictive mining** tasks perform induction on the current data in order to make predictions. Interesting patterns represent knowledge.

Characterization and Discrimination

Data entries can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived using data characterization, data discrimination, or both.

Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query. The output of data characterization can be presented in various forms. Examples include pie charts, bar charts, curves,multidimensional data cubes, and multidimensional tables, including crosstabs. The resulting descriptions can also be presented as generalized relations or in rule form (called **characteristic rules**).

Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries. The methods used for data discrimination are similar to those used for data characterization. Discrimination descriptions expressed in the form of rules are referred to as **discriminant rules**.

Mining Frequent Patterns

Frequent patterns are patterns that occur frequently in data. There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures. A frequent itemset typically refers to a set of items that often appear together in a transactional data set. A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern. A substructure can refer to different structural forms (*e.g.*, graphs, trees, or lattices) that may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern. Mining frequent patterns leads to the discovery of interesting associations and correlations within data. Frequent itemset mining is a fundamental form of frequent pattern mining.

Associations and Correlations

???

Classification and Prediction

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The model is used to predict the class label of objects for which the the class label is unknown. The derived model may be represented in various forms, such as classification rules (*i.e.*, IF-THEN rules), decision trees, mathematical formulae, or neural networks. Regression is used to predict missing or unavailable numerical data values rather than (discrete) class labels.

The term prediction refers to both numeric prediction and class label prediction. Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution trends based on the available data.

Classification and regression may need to be preceded by relevance analysis, which attempts to identify attributes that are significantly relevant to the classification and regression process. Such attributes will be selected for the classification and regression process. Other attributes, which are irrelevant, can then be excluded from consideration.

Cluster Analysis

Clustering analyzes data objects without consulting class labels. In many cases, class labeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. Each cluster so formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

Outlier Analysis

A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are **outliers**. Many data mining methods discard outliers as noise or exceptions. However, in some applications (*e.g.*, fraud detection) the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as **outlier analysis** or **anomaly mining**.

Evolution Analysis

???

Data Preprocessing

Data have quality if they satisfy the requirements of the intended use. There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability. **Inaccurate** (or noisy; containing errors, or values that deviate from the expected, having incorrect attribute values), **incomplete** (lacking attribute values or certain attributes of interest, or containing only aggregate data), and **inconsistent** data (*e.g.*, Age="42" Birthday="03/07/1997") are commonplace properties of large real-world databases and data warehouses.

Possible reasons for inaccurate data:

- The data collection instruments used may be faulty.
- There may have been human or computer errors occurring at data entry.
- Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information (*e.g.*, by choosing the default value "January 1" displayed for birthday). This is known as **disguised missing data**.
- Errors in data transmission can also occur. There may be technology limitations such as limited buffer size for coordinating synchronized data transfer and consumption.
- Incorrect data may also result from inconsistencies in naming conventions or data codes, or inconsistent formats for input fields (*e.g.*, date).

Incomplete data may come from

- Attributes of interest may not always be available.
- Different considerations between the time when the data was collected and when it is analyzed.
- Human/hardware/software problems.
- Data that were inconsistent with other recorded data may have been deleted.

Inconsistent data may come from

- Different data sources
- Functional dependency violation (*e.g.*, modify some linked data)

Duplicate tuples also require data cleaning.

Why preprocess the Data?

Real world data tend to be dirty, incomplete, and inconsistent. No quality data means no quality mining results. Quality decisions must be based on quality data. Duplicate or missing data may cause incorrect or even misleading statistics. If users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied. Dirty data can cause confusion for the mining procedure, resulting in unreliable output. Data preprocessing techniques can improve data quality, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an important step in the knowledge discovery process, because quality decisions must be based on quality data. Detecting data anomalies, rectifying them early, and reducing the data to be analyzed can lead to huge payoffs for decision making.

Descriptive Data Summarization

Measures of central tendency measure the location of the middle or center of a data distribution, *i.e.*, given an attribute, where do most of its values fall.

Mean

$$\text{sample, } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{population, } \mu = \frac{1}{N}\sum x$$

$$\text{weighted arithmetic mean, } \overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Trimmed mean is the mean obtained after chopping off values at the high and low extremes.

Median For skewed (asymmetric) data, a better measure of the center of data is the median, which is the middle value in a set of ordered data values.

$$median = L_1 + \left(\frac{\frac{N}{2} - (\sum freq)_l}{freq_{median}} \right) width$$

L_1 → lower boundary of the median interval

N → number of values in the entire data set

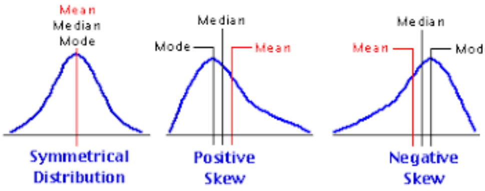
$\left(\sum freq\right)_l$ → sum of the frequencies of all of the intervals that are lower than the median interval

$freq_{median}$ → frequency of the median interval

$width$ → width of the median interval

Mode For unimodal numeric data that are moderately skewed:

$$mean - mode = 3 \times (mean - median)$$



Measuring the Dispersion of data measures how are the data spread out.

Range difference between the largest and smallest values in the set.

Quantiles are points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets.

Percentiles The 100-quantiles are more commonly referred to as percentiles; they divide the data distribution into 100 equal-sized consecutive sets.

Quartiles The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as quartiles.

The quartiles give an indication of a distribution's center, spread, and shape. The first quartile, denoted by Q_1, is the 25th percentile. It cuts off the lowest 25% of the data. The third quartile, denoted by Q_3, is the 75th percentile—it cuts off the lowest 75% (or highest 25%) of the data. The second quartile is the 50th percentile. As the median, it gives the center of the data distribution.

Interquartile range The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (IQR) and is defined as

$$IQR = Q_3 - Q_1$$

A common rule of thumb for identifying suspected outliers is to single out values falling at least $1.5 \times IQR$ above the third quartile or below the first quartile.

Five-number summary $Minimum,\ Q_1,\ Median,\ Q_3,\ Maximum$

Boxplots

- The ends of the box are at the quartiles so that the box length is the interquartile range.
- The median is marked by a line within the box.
- Two lines (called **whiskers**) outside the box extend to the smallest and largest observations.

When dealing with a moderate number of observations, it is worthwhile to plot potential outliers individually. To do this in a boxplot, the whiskers are extended to the extreme low and high observations only if these values are less than $1.5 \times IQR$ beyond the quartiles.Otherwise, the whiskers terminate at the most extreme observations occurring within $1.5 \times IQR$ of the quartiles. The remaining cases are plotted individually.

Standard deviation is denoted as σ. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

Variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \overline{x}^2 \qquad ;\ where,\ \overline{x} \text{ is the mean}$$

The normal (distribution) curve

- From $\mu - \sigma$ to $\mu + \sigma$ contains about 68% of the measurements.
- From $\mu - 2\sigma$ to $\mu + 2\sigma$ contains about 95% of it.
- From $\mu - 3\sigma$ to $\mu + 3\sigma$ contains about 99.7% of it.

Two attributes, X, and Y, are correlated if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated).

Visualization techniques

Histogram Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data.

Quantile Plot Displays all of the data, allowing the user to assess both the overall behavior and unusual occurrences.

Quantile-Quantile (Q-Q) Plot Graphs the quantiles of one univariate distribution against the corresponding quantiles of another. Allows the user to view whether there is a shift in going from one distribution to another.

Scatter plot Provides a first look at bivariate data to see clusters of points, outliers, etc. Each pair of values is treated as a pair of coordinates and plotted as points in the plane.

Loess Curve Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence. It is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression

Major Tasks in Data Preprocessing:

Data cleaning Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

Data integration Integration of multiple databases, data cubes, or files.

Data transformation Normalization and aggregation.

Data reduction Obtains reduced representation in volume but produces the same or similar analytical results.

Data discretization Part of data reduction but with particular importance, especially for numerical data.

Data Cleaning

Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, correct inconsistencies in the data, and resolve redundancy caused by data integration.

Missing Values may occur due to equipment malfunction, inconsistency with other recorded data and thus deleted, data not entered due to misunderstanding, certain data may not be considered important at the time of entry, no register history or changes of the data.

- Ignore the tuple
- Fill in the missing value manually
- Use a global constant to fill in the missing value
- Use a measure of central tendency for the attribute (*e.g.*, the mean or median) to fill in the missing value
- Use the attribute mean or median for all samples belonging to the same class as the given tuple

- Use the most probable value to fill in the missing value

Good database and data entry procedure design should help minimize the number of missing values or errors in the first place.

Noisy Data Noise is a random error or variance in a measured variable.

- **Binning** methods smooth a sorted data value by consulting its neighborhood. The sorted values are distributed into a number of buckets, or bins. Because binning methods consult the neighborhood of values, they perform local smoothing. In **smoothing by bin means**, each value in a bin is replaced by the mean value of the bin. Similarly, **smoothing by bin medians** can be employed, in which each bin value is replaced by the bin median. In **smoothing by bin boundaries**, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing.

  Equal-width (distance) partitioning divides the range into N intervals of equal size (uniform grid). **Equal-depth (frequency) partitioning** divides the range into N intervals, each containing approximately same number of samples.
- **Regression:** smooth by fitting the data into regression functions.
- **Clustering:** detect and remove outliers, also called **Outlier analysis**.
- **Combined computer and human inspection:** detect suspicious values and check by human.

Data cleaning as a process:

1. **Discrepancy detection:** Discrepancies can be caused by poorly designed data entry forms that have many optional fields, human error in data entry, deliberate errors, data decay, inconsistent data representations and inconsistent use of codes, errors in instrumentation devices that record data and system errors. Errors can also occur when the data are (inadequately) used for purposes other than originally intended. There may also be inconsistencies due to data integration. To detect discrepancy
   - use metadata (*e.g.*, domain, range, dependency, distribution).
   - check field overloading (error source that typically results when developers squeeze new attribute definitions into unused portions of already defined attributes).
   - check uniqueness rule (each value of the given attribute must be different from all other values for that attribute), consecutive rule (there can be no missing values between the lowest and highest values for the attribute, and that all values must also be unique) and null rule (specifies the use of blanks, question marks, special characters, or other strings that may indicate the null condition and how such values should be handled).
   - Use commercial tools: Data scrubbing tools use simple domain knowledge to detect errors and make corrections; Data auditing tools find discrepancies by analyzing the data to discover rules and relationships, and detecting data that violate such conditions.
2. **Data migration and integration** Data migration tools allow transformations to be specified. ETL (Extraction/Transformation/Loading) tools allow users to specify transformations through a graphical user interface.

Data Integration and Transformation

Data integration is the merging of data from multiple data stores. Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set. Schema integration is to integrate metadata from different sources.

Entity identification problem How can equivalent real-world entities from multiple data sources be matched up?

Redundancy An attribute may be redundant if it can be derived from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set. The same attribute or object may have different names in different databases. Some redundancies can be detected by correlation analysis.

Correlation Analysis Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data.

- **Correlation coefficient for Numeric data**, also called Pearson's product moment coefficient.

- χ^2 **test for Nominal data**, also known as Pearson χ^2 statistic.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

The χ^2 statistic tests the hypothesis that A and B are independent, *i.e.*, there is no correlation between them. The larger the χ^2 value, the more likely the variables are related. The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count. Correlation does not imply causality.

Find correlation between data by χ^2.

Data Transformation the data are transformed or consolidated into forms appropriate for mining.

Smoothing remove noise from data.

Aggregation summarization, data cube construction.

Generalization concept hierarchy climbing.

Normalization scaled to fall within a small, specified range

min-max normalization

$$v' = \frac{v - min_A}{max_a - min_a}(new\_max_A - new\_min_A) + new\_min_A$$

z-score normalization also called zero-mean normalization.

$$v' = \frac{v - \mu}{\sigma}$$

normalization by decimal scaling

$$v' = \frac{v}{10^j} \qquad \text{where } j \text{ is the smallest integer such that } Max(|v'|) < 1$$

Attribute/feature construction New attributes constructed from the given ones

Data discretization and concept hierarchy generation are also forms of data transformation.

Data Reduction

Complex data analysis and mining on huge amounts of data can take a long time, making such analysis is impractical or infeasible. Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. Data reduction strategies

Data cube aggregation The aggregated data for an individual entity of interest. The cube created at the lowest abstraction level is referred to as the **base cuboid**. A cube at the highest level of abstraction is the **apex cuboid**. Multiple levels of aggregation in data cubes further reduces the size of data to deal with. Each higher abstraction level further reduces the resulting data size.

Dimensionality reduction is the process of reducing the number of random variables or attributes under consideration, *e.g.*, remove unimportant attributes.

Wavelet Transforms 3.4.2 [1]

Principal Components Analysis 3.4.3 [1]

Attribute subset selection[a] irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed. The goal is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.

Data Compression transformations are applied so as to obtain a reduced or compressed representation of the original data. Lossy (reconstruct only an approximation of the original data) or Lossless (original data can be reconstructed from the compressed data without any information loss).

String compression Typically lossless, but only limited manipulation is possible without expansion.

Audio/video compression Typically lossy compression, with progressive refinement. Sometimes small fragments of signal can be reconstructed without reconstructing the whole.

Numerosity reduction techniques replace the original data volume by alternative, smaller forms of data representation. These techniques may be parametric (a model is used to estimate the data) or nonparametric (histograms, clustering, sampling, and data cube aggregation).

- 3.4.5 [1]
- 3.4.6 [1]
- 3.4.7 [1]
- 3.4.8 [1]

[a] In machine learning, feature subset selection

Data Discretization and Concept Hierarchy Generation

Discretization divides the range of a continuous attribute into intervals. Reduces the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Can be performed recursively on an attribute. Data size is reduced by discretization and prepared for further analysis. Can be Supervised or unsupervised, and split (top-down) or merge (bottom-up).

Concept Hierarchy recursively reduces the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as young, middle-aged, or senior).

Discretization by Binning top-down splitting technique based on a specified number of bins. Does not use class information and is therefore an unsupervised discretization technique.

Discretization by Histogram Analysis unsupervised top-down split.

Discretization by Cluster Either top-down split or bottom-up merge, unsupervised

Entropy-Based Discretization supervised, top-down split. Entropy is calculated based on class distribution of the samples in the set. The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization. The process is recursively applied to partitions obtained until some stopping criterion is met. Such a boundary may reduce data size and improve classification accuracy.

Interval merging by χ^2 Analysis unsupervised, bottom-up merge

Segmentation by natural partitioning top-down split, unsupervised

Classification, Clustering and Prediction

Classification by Decision Tree Induction

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily be converted to classification rules.

The strengths of decision tree methods are:

- able to generate understandable rules.
- perform classification without requiring much computation.
- able to handle both continuous and categorical variables.
- provide a clear indication of which fields are most important for prediction or classification.

Weaknesses of decision tree methods:

- less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- prone to errors in classification problems with many class and relatively small number of training examples.
- can be computationally expensive to train.

The major challenge is to identify the attribute for the root node in each level. This process is known as **attribute selection**. Two popular attribute selection measures:

- **Information Gain** is a measure of change in entropy. Entropy is the measure of uncertainty of a random variable.

It characterizes the impurity of an arbitrary collection of examples. The lower the entropy more the information content.

$$Entropy,\ H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

ID3 (Iterative Dichotomizer 3) uses Entropy function and Information gain as metrics.

- **Gini Index** or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. If all the elements belong to a single class, then it can be called pure The degree of Gini index varies between 0 and 1 where 0 denotes that all elements belong to a certain class or if there exists only one class, and 1 denotes that the elements are randomly distributed across various classes. A Gini Index of 0.5 denotes equally distributed elements into some classes. CART(Classification And Regression Tree) uses Gini Index (Classification) as metric.

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

Gini Index isn't computationally intensive as it doesn't involve the logarithm function used to calculate entropy in information gain, which is why Gini Index is preferred over Information gain.

Tree construction Principle

- Splitting Attribute: An attribute selection measure is a heuristic for selecting the splitting criterion that best separates a given data partition, D, of class-labeled training tuples into individual classes. The attribute having the best score for the measure is chosen as the splitting attribute for the given tuples.
- Splitting Criterion

3 main phases

- Construction Phase
- Pruning Phase: When decision trees are built, many of the branches may reflect noise or outliers in the training data. Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data.
- Processing the pruned tree to improve the understandability.

Difference between Logistic regression and Decision tree classification

Bayesian Classification

A statistical classifier, performs probabilistic prediction, based on Bayes' Theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad posteriori = \frac{likelihood \times prior}{evidence}$$

Classification Is to Derive the Maximum Posteriori. A simple Bayesian classifier, Naïve Bayesian classifier, has comparable performance with decision tree and selected neural network classifiers. Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured.

Naïve Bayesian Classifier

- Class Conditional Independence
- Effect of an attribute value on a given class is independent of the values of other attributes
- Simplifies Computations
- If we have n-dimensional attribute vector and m-classes, then the probability of X belonging to class C_i is

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

$P(X)$ is same for all classes, $P(X|C_i)P(C_i)$ needs to be maximized.

- Naïve Bayesian prediction requires each conditional probability be non zero. Otherwise, the predicted probability will be zero. Use **Laplacian correction** (or Laplacian estimator), the probability estimates are close to their uncorrected counterparts.

Bayesian Belief Networks

- Graphical models
- Represent dependencies among subsets of attributes

Support Vector Machines

SVM is related to statistical learning theory. It is a classification method for both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyperplane. With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. SVM finds this hyperplane using support vectors (input points closest to the decision boundary) and margins (the width of separation between classes).

Other Regression-Based Methods of Prediction

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.

Mining Frequent Pattern

FP-Growth Algorithm

If you want to upgrade this shit, copy the project from Overleaf.

References

[1] Jiawei Han and Micheline Kamber. *Data mining: concepts and techniques*. 3rd ed. Burlington, MA: Elsevier, 2012. ISBN: 9780123814791.