0.1 Classification analysis

- Classification Analysis is a systematic process for obtaining important and relevant information about data, and metadata data about data. The classification analysis helps identifying to which of a set of categories different types of data belong. Classification analysis is closely linked to cluster analysis as the classification can be used to cluster data.
- Your email provider performs a well-known example of classification analysis: they use algorithms that are capable of classifying your email as legitimate or mark it as spam. This is done based on data that is linked with the email or the information that is in the email, for example certain words or attachments that indicate spam.

Classification

- Classification is one step in the process of data mining. It is used to group items based on certain key characteristics. There are several techniques used for data mining classification, including nearest neighbor classification, decision tree learning, and support vector machines.
- Generally a representative sample is chosen from the pool of data and then manipulated and analyzed to find patterns.
- In addition to data mining classification, researchers may also use clustering, regression, and rule learning to analyze the data.
- There are several algorithms that can be used in data mining classification. Nearest neighbor classification is one of the simplest of the data mining classification algorithms. It relies on a training set. A training set is a set of data used to train the computer into paying attention to certain variables. In nearest neighbor classification, the computer simply classifies all data as part of the group that contains data closest in value to the input.

0.2 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

0.2.1 Example: Credit Risk

A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

0.2.2 Binary Classification

The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating.

In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model.

Scoring a classification model results in class assignments and probabilities for each case. For example, a model that classifies customers as low, medium, or high value would also predict the probability of each classification for each customer.

Classification has many applications in customer segmentation, business modeling, marketing, credit analysis, and biomedical and drug response modeling.

Testing a Classification Model

- A classification model is tested by applying it to test data with known target values and comparing the predicted values with the known values.
- The test data must be compatible with the data used to build the model and must be prepared in the same way that the build data was prepared. Typically the build data and test data come from the same historical data set.
- A percentage of the records is used to build the model; the remaining records are used to test the model.
- Test metrics are used to assess how accurately the model predicts the known values. If the model performs well and meets the business requirements, it can then be applied to new data to predict the future.

0.2.3 Binary Classification

Binary or binomial classification is the task of classifying the elements of a given set into two groups on the basis of a Classification rule. Some typical binary classification tasks are

- medical testing to determine if a patient has certain disease or not (the classification property is the presence of the disease)
- quality control in factories; i.e. deciding if a new product is good enough to be sold, or if it should be discarded (the classification property is being good enough)
- deciding whether a page or an article should be in the result set of a search or not (the classification property is the relevance of the article, or the usefulness to the user)

Statistical classification in general is one of the problems studied in computer science, in order to automatically learn classification systems; some methods suitable for learning binary classifiers include the decision trees, Bayesian networks, support vector machines, neural networks, probit regression, and logit regression.

Sometimes, classification tasks are trivial. Given 100 balls, some of them red and some blue, a human with normal color vision can easily separate them into red ones and blue ones. However, some tasks, like those in practical medicine, and those interesting from the computer science point of view, are far from trivial, and may produce faulty results if executed imprecisely.

0.2.4 Lift (Marketing)

- ***Lift*** measures the degree to which the predictions of a classification model are better than randomly-generated predictions. Lift applies to binary classification only, and it requires the designation of a positive class. If the model itself does not have a binary target, you can compute lift by designating one class as positive and combining all the other classes together as one negative class.
- Numerous statistics can be calculated to support the notion of lift. Basically, lift can be understood as a ratio of two percentages: the percentage of correct positive classifications made by the model to the percentage of actual positive classifications in the test data.
- For example, if 40% of the customers in a marketing survey have responded favorably (the positive classification) to a promotional campaign in the past and the model accurately predicts 75% of them, the lift would be obtained by dividing 0.75 by 0.40. The resulting lift would be 1.875.
- (Lift is computed against quantiles that each contain the same number of cases. The data is divided into quantiles after it is scored. It is ranked by probability of the positive class from highest to lowest, so that the highest concentration of positive predictions is in the top quantiles. A typical number of quantiles is 10.(
- Lift is commonly used to measure the performance of response models in marketing applications. The purpose of a response model is to identify segments of the population with potentially high concentrations of positive responders to a marketing campaign. Lift reveals how much of the population must be solicited to obtain the highest percentage of potential responders.

(Source: Wikipedia)

- Lift is a measure of the performance of a targeting model (association rule) at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model. A targeting model is doing a good job if the response within the target is much better than the average for the population as a whole. Lift is simply the ratio of these values: target response divided by average response.
- For example, suppose a population has an average response rate of 5%, but a certain model (or rule) has identified a segment with a response rate of 20%. Then that segment would have a lift of 4.0 (20%/5%). Typically, the modeller seeks to divide the population into quantiles, and rank the quantiles by lift. Organizations can then consider each quantile, and by weighing the predicted response rate (and associated financial benefit) against the cost, they can decide whether to market to that quantile or not.
- Lift is analogous to information retrieval's average precision metric, if one treats the precision (fraction of the positives that are true positives) as the target response probability.
- The Lift curve can also be considered a variation on the Receiver operating characteristic (ROC) curve, and is also known in econometrics as the Lorenz or power curve.

0.2.5 Some Approaches

Decision tree learning

Decision tree learning uses a branching model to classify the data. The computer basically asks a series of questions about the data. If the answer to the first question is true, it asks question 2a. If the answer is false, it asks question 2b. When drawn out, this method forms a tree of branching paths.

Naive Bayes classification

Naive Bayes classification relies on probability. It asks a series of questions about each piece of data and then uses the answers to determine the probability that the data belong in a particular classification. This is different from decision tree learning because the answer to the first question does not influence which question will be asked next.

Advanced methods

More complicated methods of data mining classification include neural networks and support vector machines. These methods are computer-based models that would be difficult to do by hand. Neural networks is often used in artificial intelligence programming because it mimics the human brain. It filters information through a series of nodes that find patterns and then classify the information.

Support vector machines

Support vector machines use training samples to build a model that will classify information, usually visualized as a scatter plot with a wide space between categories. When new information is fed into the machine, it is plotted on the graph. The data are then classified based on which category the information falls closest to on the graph. This method works only when there are two options to choose from.