MODULE 5

Machine LearningAlgorithmsfor Big Data Analytics

LEARNING OBJECTIVES

After studyingthis chapter,you will be able to:

LO 6.1 Understand metric, feature and category variables, evaluate and estimate the relationships, outliers, variances, probability distribution, and the correlations between the variables in variables, items or entities

LO 6.2 Apply regression analysis using linear, non-linear and multiple regression models,
K-Nearest Neighbours (KNN) distance measures, and predict the expected results

LO 6.3 Discover similar items and the similarities using distance measures

LO 6.4 Apply methods for Frequent-Itemsets Mining (FIM), market-basket model, association-rules mining, Apriori algorithm, and method of evaluation of Candidate Rules. Get knowledge of FIM and association rule applications and find the associations and similarities

LO 6.5 Apply methods of unsupervised machine-learning for clustering a collection, K-means, determine the number of clusters, and perform

cluster diagnostics

LO 6.6 Apply methods of supervised machine-learning for classification; K-Nearest Neighbour (KNN) classifier, decision trees and Random-Forest, AdaBoost and other ensemble, Naive Bayes classifiers, Artificial Neural Networks and SVM-based classifiers

LO 6.7 Design a recommender system using approaches of Collaborative Filtering (CF), Contents-based Filtering (CBF), Knowledge-based Filtering (KBF) and hybrid approaches for making recommendations

LO 6.8 Get knowledge of Apache Mahout Architecture, the ML algorithms for clustering, classification, collaborative filtering, and design recommender in Big Data environment

RECALL FROM EARLIER CHAPTERS

The most popular open-source analytics-tools are Apache Spark, Python, R, Apache Pig, and Hive, according to a study (Section 5.1). Spark with multiple languages, Python and Scala shells provide *great ease in programming for complex analytics, machine learning* and other solutions (Section 5.2).

Big Data analysis requires scalable distributed computations. Spark scalable MLib (machine-learning library) consists of the widely used ML algorithms and utility functions for large datasets. MLib includes the algorithms for optimization-primitives, regression, collaborative filtering, dimensionality reduction, cluster analysis, classification and recommender.

Uses of Machine Learning (ML) and Artificial Neural Networks (ANN) are in analytics, predictive modeling and decisions. The analytics methods include data mining, pattern mining, clusters analysis and detection of anomalies.

Apache Mahout consists of ML algorithms for Big Data analysis (Section 2.2.3 Figure 2.2).

This Chapter focusses on the ML methods of regression analysis based predictions, finding similarities, FIM, clustering, classifiers, recommenders, and introduces Mahout Architecture and features for the ML applications in Big Data analytics.

6.1 ! INTRODUCTION

Analytics uses the mathematical equations, formulae and models. Analytics also uses the statistics, AI, ML and DL, and predict the behaviour of entities, objects and events. Statistics refers to studying organization, analysis of a collection of data, making interpretations and presentation of analyzed results.

Artificial Intelligence (AI) refers to the science and engineering of making computers perform tasks, which normally require human intelligence. For example, tasks such as predicting future results, visual perception, speech recognition, decision making and natural language processing.

Two concepts in AI, *'machine learning'* and 'deep learning' provide powerful tools for advanced analytics and predictions.

Google-owned company *Deep Mind* developed an Artificial Intelligence (AI) program called AlphaZero, which played 100 chess games in 24 hours, and defeated Stockfish, the highest-rated chess program by 28 games to 0 with 72 games drawn. This was a historical moment. It became a milestone in the history of AI, ML and DL.

The former world champion, Garry Kasparov, noted that achievement of AlphaZero has history-shaping potential. "The ability of a machine to replicate and surpass centuries of human knowledge, is a world-changing tool". (Garry Kasparov, *"Deep Thinking - Where Artificial Intelligence Ends and Human Creativity Begins"*, published by the author himself, 2017)

Machine Learning - Definition and Usage Examples

Machine Leaming (ML) is a field of computer science based on AI which deals with learning from data in three phases, i.e. *collect, analyze* and *predict.* It does not rely on explicitly programmed instructions.

An ML program learns the behavior of a process. The program uses data generated from various sources for training. Learning from the outcomes from common inputs improves future performance from previous outcomes. Learning applies in many fields of research and industry. Learning from study of data enables efficient and logical decisions for future actions.

Advanced ML techniques use unsupervised, semi-supervised or supervised

learning. Supervised learning uses a known dataset (called *training dataset).* Learning enables creation of a *model program* for evaluating outcomes. The program makes future predictions and leads to knowledge discovery. Supervised learning uses output datasets, which are used to train a machine (program) such that the program leads to the desired outputs. Unsupervised learning does not use output datasets to train a machine.

Deep Leaming (DL) refers to structured learning (DSL) or hierarchical learning. DL methods are advanced methods, such as artificial neural networks (ANN) such as artificial neural networks (ANN) or neural nets, deep neural networks, deep belief networks and recurrent neural networks. Learning can be unsupervised, semi-supervised or supervised. Applications of DL and ANN include computer vision, speech recognition, Natural Language Processing (NLP), audio recognition, social network filtering, machine translation, bioinformatics and drug design. DL methods give results comparable to and in some cases superior to human experts.

The present chapter describes the ML methods and introduces Mahout Architecture, features and its ML applications. Section 6.2 describes methods of estimating relationships, outliers, variances, probability distribution, errors and correlations in variables, items and entities. Section 6.3 describes regression analysis using linear, non-linear and multiple-regressor models and KNN distance measures for making predictions. [Regressor means an independent (explanatory) variable in regression equation.]

Section 6.4 describes methods of finding similar items, similarities and filtering of similar items. Section 6.5 describes frequent-itemset mining by collaborative filtering of similar itemsets. Section 6.5 also describes associations and association rules mining. Section 6.6 describes methods of finding the clusters. Section 6. 7 describes the classifiers for classifying data in datasets. Section 6.8 introduces recommendation system and collaborative, content, knowledge and hybrid recommendation approaches. Section 6.9 describes Apache Mahout and ML algorithms for Big Datasets.

The following sections use a convention for fonts when denoting an absolute value, mean value, function value, vector element, set member, entity or variable using a character or set of characters, entities
or elements.

1. [u] represents absolute value of u, means value without sign. For example, consider 1- 31 and |+ 31, the value of both is 3.
2. x represents mean, average or expected value of x.
3. F (y, x) represents a function with an expression, which finds value of F from the given values of y and x. F (y, x) values depend on one or more dependent variables as a function of one or more independent variables. For example, F depends y as well as xis F (y, x) = 1/sqrt {(y + x)2 + k2}. The F also depends on constant k. Another example is y = F (x), for example, y = cos (x). F (x) represents a function F, which gives value of y, is a dependent variable. The xis an independent variable.
4. V denotes a vector V (vi, V2 ...). V is in bold font. Vl and V2 are in text font and are elements 1 and 2 of V. The V consists of number of elements Vl, V2 ...
5. [o] represents length of vector U.
6. S denotes a set S (A, B, C ...). Font S is in French script MT or distinct font for English S. The A, Band Care in text font (no bold), and are the members of S. The members can be vectors or subsets. They, when denoted in bold, represent vector elements.

6.21 ESTIMATING THE RELATIONSHIPS, OUTLIERS, VARIANCES, PROBABILITY DISTRIBUTIONS AND CORRELATIONS

LO 6.1

~et1iic.. feawre and categoiry wrialbl es, Re'latiollilslmi ps, oll.!ltiliers- vari aaoas. probabl lty distributiion, and the comrelatioliils betivooiiti tihe van altiles- "rtems, or e:ntitiies

Methods of studying relationships use variables. Types of variables used are as follows:

Independent variables represent directly measurable characteristics. For example, year of sales figure or semester of study. Dependent variables represent the characteristics. For example, profit during successive years or grades awarded in successive semesters. Values of a dependent variable depend on the value of the

independent variable.

Predictor variable is an independent variable, which computes a dependent variable using some equation, function or graph, and does a prediction. For example, predicts sales growth of a car model after five years from given input datasets for the sales, or predicts sentiments about higher sales of particular category of toys next year.

Outcome variable represents the effect of manipulation(s) using a function, equation or experiment. For example, CGPA (Cumulative Grade Points Average) of the student or share of profit to each shareholder in a year using profit as the dependent variable. CGPA of a student computes from the grades awarded in the semesters for which student completes his/her studies. A company declares the share of profit to each shareholder in a year after subtracting requirements of money for future growth from the profit.

Explanatory variable is an independent variable, which explains the behavior of the dependent variable, such as linearity coefficient, non-linear parameters or probabilistic distribution of profit-growth as a function of additional investment in successive years.

Response variable is a dependent variable on which a study, experiment or computation focuses. For example, improvement in profits over the years from the investments made in successive years or improvement in class performance is measured from the extra teaching efforts on individual students of a class.

Feature variable is a variable representing a characteristic. For example, apple feature red, pink, maroon, yellowish, yellowish green and green. Feature variables are generally represented by text characters. Numbers can also represent features. For example, red with 1, orange with 2, yellow with 3, yellowish green 4 and green 5.

Categorical variable is a variable representing a category. For example, car, tractor and truck belong to the same category, i.e., a four-wheeler automobile. Categorical variables are generally represented by text characters.

Independent and dependent variables may exhibit a relation or correlation. The relationships may be linear, nonlinear, positive, negative, direct, inverse, scattered or spread. A data point for dependent variable can be an outlier with no relationship.

Data analysis requires studying relationships graphically, mathematically and statistically, studying the outliers, anomalies, variances, correlations, features, categories and probability distributions using a set of variables, and other characteristics. The relationship involves some quantifiable independent variables and the resulting dependent variable or entity. The following subsections explain methods of estimating the relationships, outliers, variances, correlations and probability distributions between a set of variables.

6.2.1 Relationships-Using Graphs, Scatter Plots and Charts

A relationship between two or more quantitative dependent variables with respect to an independent variable can be well-depicted using graph, scatter plot or chart with data points, shown in distinct shapes. Conventionally, independent variables are on the x-axis, whereas the dependent variables on the y-axis in a graph. A line graph uses a line on an x-y axis to plot a continuous function.

A scatter plot is a plot in which dots or distinct shapes represent values of the dependent variable at the multiple values of the independent variable [Section 10.5]. Whether two variables are related to each other or not, can be derived from statistical analysis using scatter plots.

A data point is (xi, Yi) when dependent variable value = Yi at the independent variable value= xi. The
i = 1, 2 ... n for number of data points = n. The **i** varies with the position of projection of the point on X-axis. Scatter plot represents data points by dots. The dot can also be a bubble, triangle, circle, cross or vertical bar. Size or colour of dot distinguishes the dependent variables on the same plot.

Another method is quantifying two or more dependent variables by columns of different widths with filled colours, shades or patterns. The width quantifies the dependent variable. The column-position quantifies the independent variable.

Examples of dependent variables are sales of five car models in a year, grades in five courses taken in a semester.

6.2.1.1 Linear and Non-linear Relationships

A linear relationship exists between two variables, say x and y, when a straight line ($y = a_0 + a_1.x$) can fit on a graph, with at least some reasonable degree of accuracy. The a_1 is the linearity coefficient. For example, a scatter chart can suggest a linear relationship, which means a straight line. Figure 6.1 shows a scatter plot, which fits a linear relationship between the number of students opting for computer courses in years between 2000 and 2017.

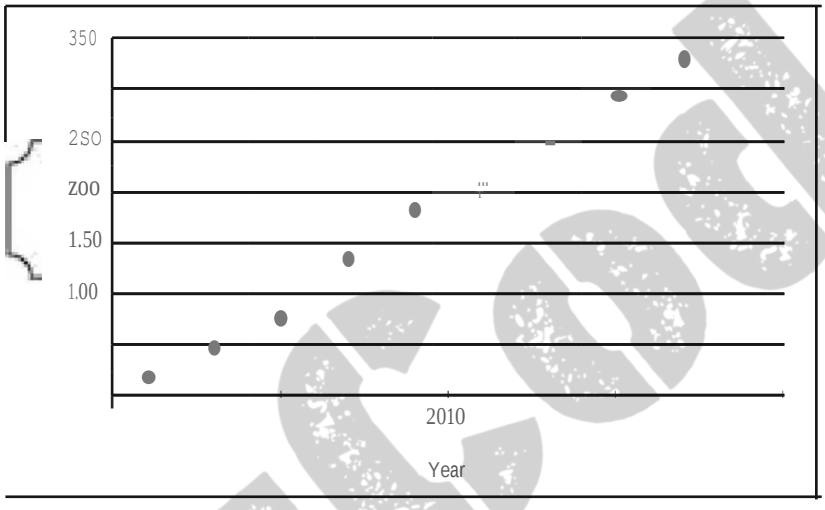


Figure 6.1 Scatter plot for linear relationship between students opting for computer courses in years between 2000 and 2017

A linear relationship can be positive or negative. A positive relationship implies if one variable increases in value, the other also increases in value. A negative relationship, on the other hand, implies when one increases in value, the other decreases in value. Perfect, strong or weak linearship categories depend upon the bonding between the two variables.

A non-linear relationship is said to exist between two quantitative variables when a curve ($y = a_0 + a_1.x + a_2.x_2 + \cdots$) can be used to fit the data points. The fit should be with at least some reasonable degree of accuracy for the fitted parameters, a_0, a_1, a_2 ··· Expression for y then generally predicts the values of one quantitative variable from the values of the other quantitative variable with considerably more accuracy than a straight line.

Consider an example of non-linear relationship: The side of a square and its area are not linear. In fact, they have quadratic relationship. If the side of a square doubles, then its area increases four times. The relationship predicts the area from the side.

Figure 6.2 shows a scatter plot in case of a non-linear relationship between side of square and its area.

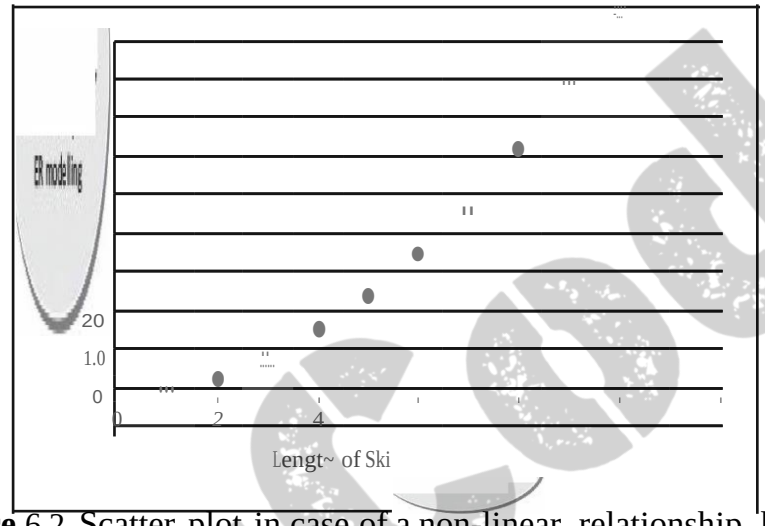


Figure 6.2 Scatter plot in case of a non-linear relationship between side of square and its area

6.2.2 Estimatingthe Relationships

Estimating the relationships means finding a mathematical expression, which gives the value of the variable according to its relationship with other variables. For example, assume Ym = sales of a car model *m* in xth year of the start of manufacturing that model. Assume that computations show that the y_m relates by a mathematical expression ($ym = a_0 + a_1.xm + a_2.xm_2$) up to an acceptable degree of accuracy, when $a_0 = 490, a_1 = 10$ and $a_2 = 5$.

Estimated first year sales, $Ym(l) = (490 + 10) = 500$, second year $Ym(2) = (490 + 10 \times 2 + 5 \times 2_2) = 530$, third year $Ym(3) = (490 + 10 \times 3 + 5 \times 3_2) = 565$, if fit with the desired accuracy, then the results are showing that the expression of y_m estimates the relationship between model *m* sales in next and other years. The y_m can also predict the sales in 6th or later years. Predictions are up to a certain

degree of certainty.

6.2.3 Outliers

Outliers are data, which appear as they do not belong to the dataset (Section 5.3.3.1). Outliers are data points that are numerically far distant from the rest of the points in a dataset, are termed as outliers. Outliers show significant variations from the rest of the points (Section 1.5.2.2). Identification of outliers is important to improve data quality or to detect an anomaly. The estimating parameters mathematically, statistically, describing an outcome, predicting a dependent variable value, or taking the decisions based on the datasets given for the analysis are sensitive to the outliers.

There are several reasons for the presence of outliers in relationships. Some of these are:

- Anomalous situation
- Presence of a previously unknown fact
- Human error (errors due to data entry or data collection)
- Participants intentionally reporting incorrect data (This is common in self-reported measures and measures that involve sensitive data which participant doesn't want to disclose)
- Sampling error (when an unfitted sample is collected from population).

Population means any group of data, which includes all the data of interest. For example, when analysing 1000 students who gave an examination in a computer course, then the population is 1000. 100 games of chess will represent the population in analysis of 100 games of chess of a grandmaster.

Sample means a subset of the population. Sample represents the population for uses, such as analysis and consists of randomly selected data.

6.2.4 Variance

A random variable is a variable whose possible values are outcomes of a random phenomenon. A random variable is a function that maps the outcomes of unpredictable processes to numerical quantities. A random variable is also called stochastic variable or random quantity. Randomness can be around some

expected mean value or outcome, and with some normal deviation.

Variance measures by the sum of squares of the difference in values of a variable with respect to the expected value. Variance can alternatively be a sum of squares of the difference with respect to value at an origin. Variance indicates how widely data points in a dataset vary. If data points vary greatly from the mean value in a dataset, the variance is large; otherwise, the variance is less. The variance is also a measure of dispersion with respect to the expected value.

A high variance indicates that the data in the dataset is very much spread out over a large area (random dataset), whereas a low variance indicates that the data is very similar in nature.

No variance is sometimes hard to understand in real datasets. The following example illustrates no variance:

EXAMPLE 6.1

Consider an examination where everyone gets the same grades. What does it signify?

SOLUTION

Some measurement problem may have taken place in a situation where either the semester examination questions were so easy that everyone got full marks, or it was so hard that everyone got a zero. Now consider the two types of examinations. After each examination, everyone gets the same score on the test, i.e., everyone gets 'A' grade in one test and everyone gets 'B' in the second test. This is again not telling much

about the study or intelligent quotient of the students. Now, these no variance results signify the extreme case and hard to understand or explain. But in general, differences in scores are always found.

6.2.4.1 Standard Deviation and Standard Error Estimates

The variance is not a standalone statistical parameter. Estimations of other statistical parameters, such as standard deviation and standard error are also used.

Standard Deviation With the help of variance, one can find out the standard

deviation. Standard deviation, denoted bys, is the square root of the variance. The s says, "On an average how far do the data points fall from the mean or expected outcome?" Though the interpretation is the same as variance but s is squared rooted, therefore, less susceptible to the presence of outliers. The formulae for the population and the sample standard deviations are as follows:

Tbe Population Standard Deviation: $a = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ (6.la)

The Sample Standard Deviation: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ {6.1 bi

where N is number of data points in population, S is number in the sample, m is expected in the population or average value of x, and X is expected x in the sample.

StandardErrorThe standard error estimate is a measure of the accuracy of predictions from a relationship. Assume the linear relationship in a scatter plot of y (Figure 6.1). The scatter plot line, which fits, is defined as the line that minimizes the sum of squared deviations of prediction (also called the **sum of squares error).** The **standard error of the estimate** is closely related to this quantity and is defined below:

$$s_{est} = \sqrt{\frac{\sum (y - y')^2}{N}} \quad \text{... (6.2)}$$

where sest is the standard error in the estimate, y is an observed value, y¢ is a predicted value, and N is the number of values observed. The standard error estimate is a measure of the dispersion (or variability) in the predicted values from the expression for relationship. Following are three interpretations from the sest:

1. When sest is small, most of the observed values (y) dots are fairly close to the fitting line in the scatter plot, and better is the estimate based on the equation of the line.
2. When the sest is large, many of the observed values are far away from the line.
3. When the standard error is zero, then no variation exists corresponding to the computed line for predictions. The correlation between the observed

and estimation is perfect.

6.2.5 Probabilistic Distribution of Variables, Items or Entities

Probability is the chance of observing a dependent variable value with respect to some independent variable. Suppose a Grandmaster in chess has won 22 out of 100 games, drawn 78 times, and lost none. Then, probability P of winning P_W is 0.22, P of drawn game P_0 is 0.78 and P of losing, $P_L = 0$. The sum of the probabilities is normalized to 1, as only one of the three possibilities exist.

Probability distribution is the distribution of P values as a function of all possible independent values, variables, situations, distances or variables. For example, if P is given by a function P(x), then P varies as *x* changes. Variations in P(x) with *x* can be discrete or continuous. The values of P are normalized such that sum of all P values is 1. Assuming distribution is around the expected value r, the standard normal distribution formula is:

$$P(x) = \frac{1}{a,Jii} e^{\frac{-(i:-:1)2-}{1a}} \quad (45.3)$$

Normal distribution relates to Gaussian function. Figure 6.3 shows a PDF with normal distribution around x = X standard deviation = s and variance = s^2.

Figure 6.3 Probability distribution function as a function of *x* assuming normal distribution around *x* = X, and standard deviation = s

The figure also shows the percentages of areas in five regions with respect to the total area under the curve for P(x). The variance for probability distribution represents how individual data points relate to each other within a dataset. The

variance is the average of the squared differences between each data value and the mean.

Moments (0, 1, 2 ...) refer to the expected values to the power of (0, 1, 2,) of random variable variance (Section 6.2.5.3). The variance is the second central moment of a distribution, which equals to the square of the standard deviation, and the covariance of the random variable with itself, and it is often represented by s^2 or var (x). The variance is computed as follows:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N} \tag{6.4}$$

Assume that probability distribution (PDF) is normal, called Gaussian distribution, which is like a bell-shaped curve (Figure 6.3). The PDF of the normal distribution is such that 68% of area under the PDF is within $(\bar{x} + s)$ and $(\bar{x} - s)$, 95% of area under the PDF is within $(\bar{x} + 2s)$ and $(\bar{x} - 2s)$ and 99.7% is within $(\bar{x} + 3s)$ and $(\bar{x} - 3s)$.

Standard deviation and empirical rule help in computing the population distribution over 68%, 95% and 99.7% of data under normally distributed population. This further helps in forecasting. The following example explains the meaning of population, expected values, normalized probabilities, PDF and interpretation using mean value.

EXAMPLE 6.2

Assume that N students gave the examination. Let N_1 is number of students obtained grade pointer average = 1, N_2 got 2, ..., N_{10} got 10. Highest-grade pointer is 10.0. Grade pointer obtained is not a random variable. Grade pointer variation is a random variable with an expected value and standard deviation.

Expected value among the distributed x_i values, where i varies discretely from 0.0 to 10.0 will depend on the expected performance of the student. If teaching in the class is very good and students prepare for the examination very well, then expected value of GPA is 8.0 for very good performing students and standard deviation found is 1.0.

(i) What do you mean by population? What do you mean by sample?

(ii) What will be the normalized probabilities?

(iii) How will you define Probability Distribution Function (PDF)?

(iv) How will you interpret the results in terms of normal distribution?

(v) When will you interpret the results as poor and poorer in terms of normal distribution?

SOLUTION

(i) Population is GPA of all the students of the university who gave the examination. Population size is N. Sample means datasets used in the analysis. It can be N or less than N students and GPA of each one.

(ii) Probability that students obtained grade pointer 1 is $(\frac{\cdot\cdot}{\cdot,,})$, 2 is (~), ... on normalization of probability. ($N = N_1 + N_2 + \cdots$)

(iii) PDF represents a curve for independent variable *x* between GPA = 0 and GPA = 10, such that the sum of all *P* values is 1, where *P;* is the ratio of number of students getting GPA = *i* with respect to the total population *N* or the sample.

$$P(x) = \frac{1}{a\{2;} e^{-(\text{\"-xt2}} \quad -?-,\bullet \quad \text{"i.tr} \qquad (6.5)$$

between *x* = 0 and 10.0, wheres= 1.0 and x- = 8.0.

(iv) GPA value is 8.0 and standard deviation is 1.0, which means 68% of the students will get GPAs between 7.0 and 9.0, 95% between 6.0 and 10.0, and 99.7% between 5.0 and 10.0.

(v) The expected value of 3.0 (less than 3.0) and standard deviation of 1.0 means poor performance of students because 68% students get between 2.0 and 4.0. The expected value of 3.0- {less than 3.0, say 2.5) and standard deviation of 1.5 means poorer performance of students because 68% students get between 1.0 and 4.0.

6.2.5.1 *Kernel Functions*

A probability or weight can be represented by a kernel function[1] like a Gaussian or tri-cube function. (Kernel in English means some thing central and key (important) part. For example, the kernel inside a walnut's shell is important because it is the edible part. Kernel in an operating system is key or central component.)

Kernel function is a function which is a central or key part of another function. For example, Gaussian kernel function is the key part of the probability distribution function [Equation (6.5)]. Figure 6.3 shows the probability normal distribution, which is a Gaussian function based on the Gaussian kernel function.

A kernel function [1], K* defines as

$$K^*(u) = A.K(u). \quad (6.6u)$$

where $A > 0$. Gaussian kernel function is

$$K^*(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \quad (6.6b)$$

and when $u = \frac{T}{a}$ the distribution function is proportional to

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

$A = \left(\sigma\sqrt{2\pi}\right)$ in Equation (6.3).

Tricube kernel function is:

$$K^*(u) = (70/81)(1 - |u|^3)^3 1_{|u|}. \quad (6.6c)$$

where $|u| \le 1$.

6.2.5.2 *Moments*

Moments (0, 1, 2, ...) refer to expected values to the powers of (0, 1, 2 ...) of random variable variance. 0th moment is 1, 1st moment = $E(x) = \mu$, (expected value), 2nd moment is squared $V[(x_i - x)^2]$ = sum of product of $(x_i - x)^2$, and $P(x = x_j)$

Here, *P* is the probability at $x = xi$ when *i* is varying from 1 to n, for *n* values of random variable x. The ,th moment is ,th power of variance v((xi - xY]. Moments are evaluated from the results obtained for the randomly distributed probabilistic values of the variable, such as sales. 1st moment assigns equal weight to variances of outliers and inliers, i.e., equal weight for variance of each. znd moment assigns higher weight to outliers compared to inliers. 3rd moment assigns greater weight to outliers compared to inliers. Moment can be defined with respected to the origin, and in that case, x- is considered 0.

Let *P* is along *y* axis and variable *x* on *x* axis. Central moment means that moments compute taking X, equals to variable *x* at *x* axis point where the probability curve partitions equally by a vertical axis, parallel to they axis.

6.2.5.3 Unequal Variance Welch's t-test

A test in statistics is unequal-variance *t* test, also called Welch *t* test.

(i) The test assumes that two groups of data are sampled data which consist of Gaussian distributed populations (Equation (6.3)).

(ii) The test does not assume those two populations have the same standard deviation.

Unequal variances t-test is a two-sample location test. It tests the hypothesis that two populations have equal means. *(Hypothesis* means making assumption statements about certain characteristics of the population. For example, an assumption that most students of a specific professor will excel as a programmer. Hypothesis when tested for a decade may pass or fail depending up on whether the statistically significant results show that the students of that professor really excelled as programmers.)

Welch's t-test is an adaptation of student's t-test in statistics. The t-test is more reliable when the two samples have unequal variances and unequal sample sizes.

6.2.5.4 Analysis of Variance (ANOVA)

An ANOVA test is a method which finds whether the fitted results are significant or not. This means that the test finds out (infer) whether to reject or accept the null hypothesis. Null hypothesis is a statistical test that means *the hypothesis that "no significant difference* exists between the specified populations". Any observed

difference is just due to sampling or experimental error.

Consider two specified populations (datasets) consisting of yearly sales data of Tata Zest and Jaguar Land Rover models. The statistical test is for proving that yearly sales of both the models, means increments and decrements of sales are related or not. Null hypothesis starts with the assumption that no significant relation exists in the two sets of data (population).

The analysis (ANOVA) is for disproving or accepting the null hypothesis. The test also finds whether to accept another alternate hypothesis. The test finds that whether testing groups have any difference between them or not.

Analysis of variance (ANOVA) is a useful technique for comparing more than two populations, samples, observations or results of computations. It is used when multiple sample cases are involved. Variation between samples and also within sample items may exist. For example, compare the effect of three different types of teaching methodologies on students. This may be done by comparing the test scores of the three groups of 20 students each. This technique provides inferences about whether the samples have been drawn from populations having the same mean. It is done by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.

F-test F-test requires two estimates of population variance- one based on variance between the samples and the other based on variance within the samples. These two estimates are then compared for F-test:

$$F = \frac{El(V)}{E2(V\_)} \tag{6.7}$$

where *El(V)* is an estimate of population variance between the two samples and *E2(V)* is an estimate of population variance within the two samples. Several different F-tables exist. Each one has a different level of significance. Thus, look up the numerator degrees of freedom and the denominator degrees of freedom to find the critical value.

The value of F calculated using the above-mentioned formula is to be compared to the critical value of F for the given degrees of freedom. If the F value calculated is equal or exceeds the critical value, then significant differences between the means of samples exist. This reveals that the samples are not drawn from the same population and thus null hypothesis is rejected.

6.2.5.5 No Relationship Case

Statistical relationship is a dependence or association between two random variables or bivariate data. Bivariate means 'two variables'. In other words, there are two types of data. Relationships between variables need to be studied and analyzed before drawing conclusions based on it. One cannot determine the right conclusion or association when no relationship between the variables exists.

6.2.6 Correlation

Correlation means analysis which lets us find the association or the absence of the relationship between two variables, *x* and *y.* Correlation gives the strength of the relationship between the model and the dependent variable on a convenient 0-100% scale.

R-Square Risa measure of correlation between the predicted values *y* and the observed values of *x. R-squared* (R^2) is a goodness-of-fit measure in linear• regression model. It is also known as the coefficient of determination. R^2 is the square of R, the coefficient of multiple correlations, and includes additional independent (explanatory) variables in regression equation.

Interpretationof R-squared The larger the R^2, the better the regression model fits the observations, i.e., the correlation is better. Theoretically, if a model shows 100% variance, then the fitted values are always equal to the observed values, and therefore, all the data points would fall on the fitted regression line.

Correlation differs from a regression analysis. Regression analysis predicts the value of the dependent predictor or response variable based on the known value of the independent variable, assuming a more or less mathematical relationship between two or more variables within the specified variances.

6.2.6.1 Correlation Indicators of Linear Relationships

Correlation is a statistical technique that measures and describes the 'strength' and 'direction' of the relationship between two variables. Let us explore the relations between only two variables. The significant questions are:

Does y increase or decrease with x? For example, expenditure increases with income or does the number of patients decrease with proper medication. (Direction)

(i) Suppose y does increase with x; then, how fast?

(ii) Is this relationship strong?

(iii) Can reliable predictions be made? That is, if one tells the income, can the expenditure be predicted?

Relationships and correlations enable training model on sample data using statistical or ML algorithms. Statistical correlation is measured by the coefficient of correlation. The most common correlation coefficient, called the *Pearson product-moment correlation coefficient.* It measures the strength of the linear association between variables. The correlation *r* between the two variables *x* and *y* is:

$$r = \left[\sim \sum_{1)}^{t.n} X L\{[(.t'i-x)Jx[(}';-Y>J\}\sim \right. \quad (.6.8u)$$

where n is the number of observations in the sample, xi is the *x* value for observation *i,* x- is the sample mean of *x, Yi* is they value for observation *i, y-* is the sample mean of *y,* sx is the sample standard deviation of *x,* and *sy* is the sample standard deviation of *y.*

Summation is over all *n* values of *i,* i = 1, 2, ..., *n.*

[r2 is square of sample correlation coefficient between the observed outcomes and the observed predictor values, and includes intercept on y-axis in case of linear regression.]

Use of Statistical Correlation Assume one sample dataset is {u1, ---, u) containing *n* values of a parameter *r.* The ru,i is i-th data point in dataset u. (i = 1, 2, ..., *n).* Another sample dataset is {v1, ---, vn}
containing *n* values of *r. rv,i* is i-th data point in dataset v. Let the correlation among two samples is being measured. Sample Pearson correlation metric *c;* measures how well two sample datasets fit on a
straight line.

$$C,(U, 1,) = \frac{I..-(1';.._{iii} - \sim)(.')\cdot j - \sim)}{\sim Li(r6'.i - \sim \cdot 1 2 Li(r\sim.t - \sim)2} \quad \ldots (6.8b)$$

where the summations are over the values of parameter in the datasets.

Three other similarities based on correlation are:

(i) Constrained Pearson correlation - It is a variation of Pearson correlation that uses midpoint instead of mean rate.

(ii) Spearman rank correlation - It is similar to Pearson correlation, except that the ratings are ranks.

(iii) Kendall's G correlation - It is similar to the Spearman rank correlation, but instead of using ranks themselves, only the relative ranks are used to calculate the correlation.

Numerical value of correlation coefficient ranges from +1.0 to -1.0. It gives an indication of both the strength and direction of the relationship between variables.

In general, a correlation coefficient $r > 0$ indicates a positive relationship; $r < 0$ indicates a negative relationship; $r = 0$ indicates no relationship (or that the variables are independent of each other and not related). Here $r = +1.0$ describes a perfect positive correlation and $r = -1.0$ describes a perfect negative correlation.

The closer the coefficients are to +1.0 and -1.0, the greater is the *strength* of the relationship between the variables.

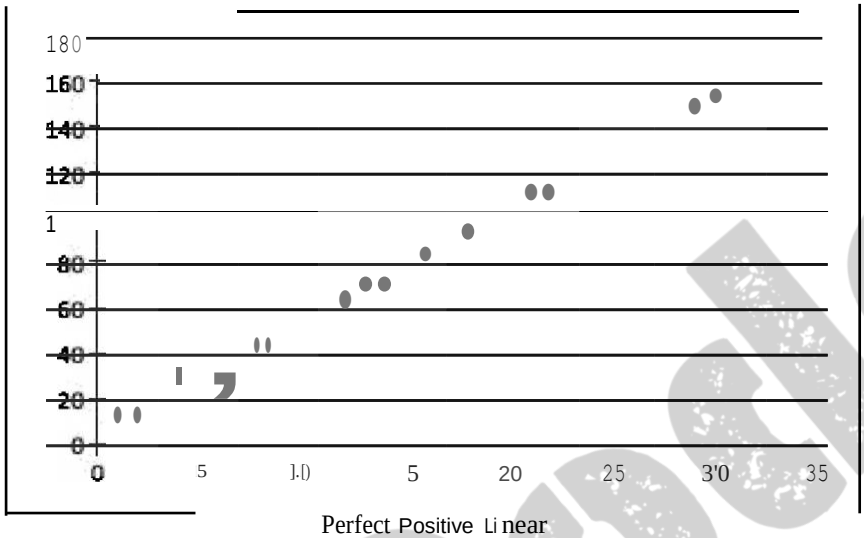
Table 6.1 gives rough guidelines on the strength of the relationship (though many experts would somewhat disagree on the choice of boundaries).

Table 6.1 The strength of the relationship as a function of *r*

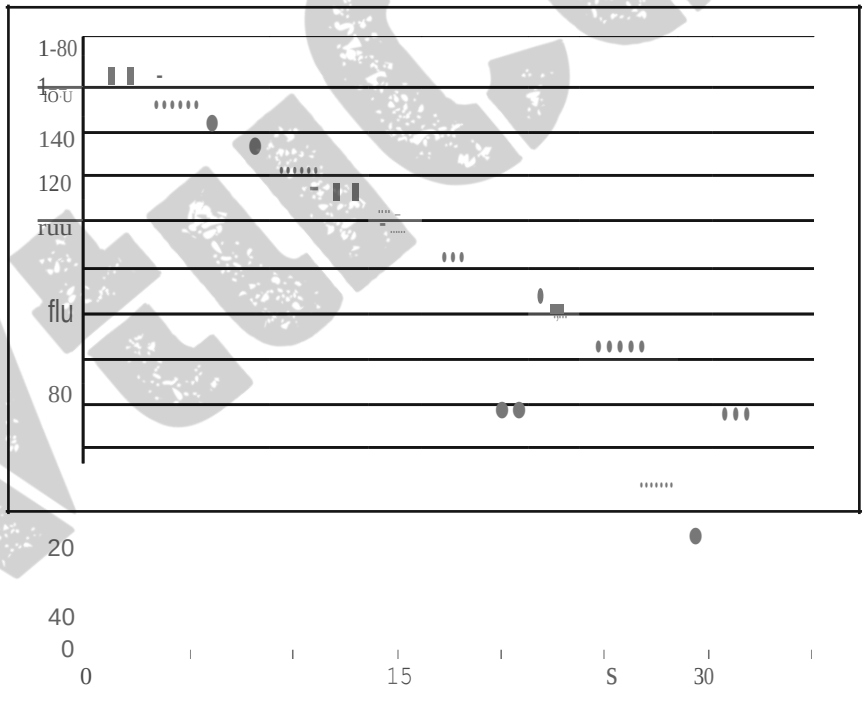
Value ofr	Strength of relationship
-1.0 to -0.5 or 1.0 to 0.5	Strong
-0.5 to -0.3 or 0.3 to 0.5	Moderate
-0.3 to -0.1 or 0.1 to 0.3	Weak
-0.1 to 0.1	None or very weak

Correlation is only appropriate for exammmg the relationship between meaningful quantifiable data (such as, temperature, marks, score) rather than categorical data, such as gender, color etc. Figure 6.4 shows perfect and

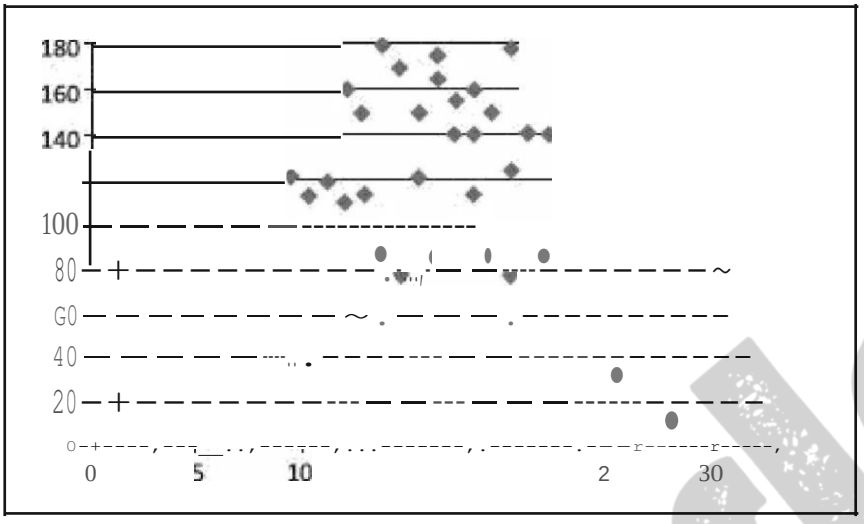
imperfect, linear positive and negative relationships, and the strength and direction of the relationship between variables



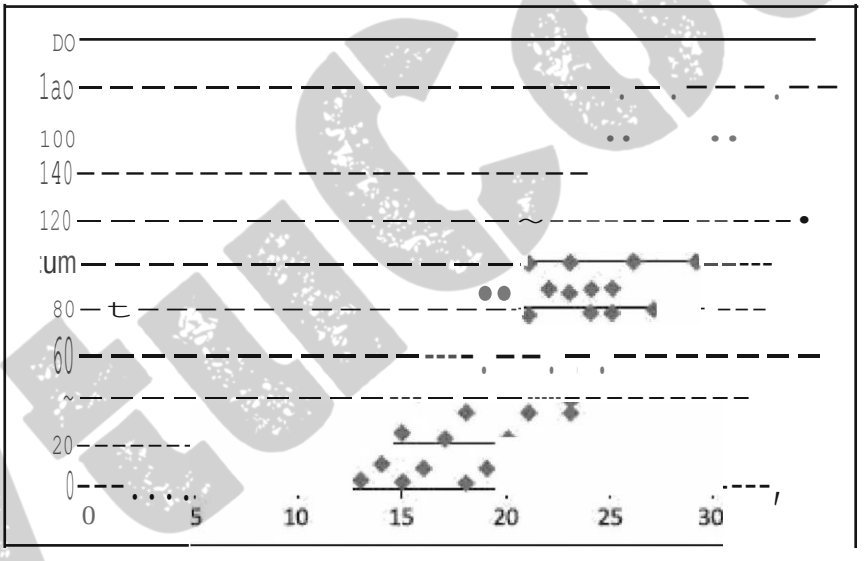
Perfect Positive Li near



Perfect Negative Linear
RelatlonstIp tr= -1~



No R@latlonsMp *(r* ... 0)



PosJtivl! Linear R.elatlonship (r = 0.9)

Figure 6.4 Perfect and imperfect, linear positive and negative relationships, and the strength and direction of the relationship between variables

Self-Assessment Exercise linked to LO 6.1

1. Define non-linear relation. Plot on the same graph, a company car sales,

y for its two models every year between 2012 to 2017, using the formula $(ym = a_0 + a_1 .xm + a_2.xm^{2)}$. How will you predict the sales in 2010? Assume for first model $a_0 = 490$, $a_1 = 10$ and $a_2 = 5$. Assume for second model $a_0 = 4900$, $a_1 = 100$ and $a_2 = 50$. Assume, $xm = 0$ for year 2011, $xm = 1$ for 2012 and $xm = 6$ for 2017.

2. How does the P(x) vary in normal distribution when expected mean is at $x = 6.0$ and standard deviation s is 1.0? Show a plot of P(x) and x and points at deviations of 1.0, 2.0 and 3.0 (means at a, $2a$ and $3a$).
3. Define mean, variance and standard deviation. How do the 0th moment, 1st moment, 2nd moment and 3rd moment compute from the values and their probabilities?
4. When will you perform t-test and F-test?
5. What does variable R-squared mean? How is the correlation parameter between predicted valued and observed value evaluated? When do you use R, r, R^2 and when r2?
6. Consider correlation r between two variables. How do you interpret $r > 0$, $r < 0$ and $r = 0$?
7. How is the inference made that two variables do not correlate?

6.3 REGRESSION ANALYSIS

LO 6.2

R~ rassien :a n:alysis u sillilg l irneJ r a ndl non-ii near reg res.sion mod E!ls. K-Neairest-Nei qhbou rs, and u sillllg di s:t.3nce measures *tor* predictions

Correlation and regression are two analyses based on multivariate distribution. A multivariate distribution means a distribution in multiple variables.

Suppose a company wishes to plan the manufacturing of Jaguar cars for coming years. The company looks at sales data regressively, i.e., data of previous years' sales. Regressive analysis means estimating relationships between variables. Regression analysis is a set of statistical steps, which estimate the relationships among variables. Regression analysis may require

many techniques for modeling and performing the analysis using multiple variables. The aim of the analysis is to find the relationships between a dependent variable and one or more independent, outcome, predictor or response variables. Regression analysis facilitates prediction of future values of dependent variables.

It helps to find how a dependent variable changes when variation is in an independent variable among a set of them, while the remaining independent variables in the set are kept fixed.

Non-linear regression equation is as follows:

$$y = a0 + a1.x + a2 \cdot x^{n} + a3..i'. \quad (,6.9)$$

where number of terms on the right-hand side are 3 or 4. Linear regression means only the first two terms are considered. The following subsections describe regression analysis in detail.

6.3.1 Simple Linear Regression

Linear regression is a simple and widely used algorithm. It is a supervised ML algorithm for predictive analysis. It models a relationship between the independent predictor or explanatory, and the dependent outcome or variable, y using a linearity equation.

$$)''' = ftao \sim a1) = "o + a1.x \sim \quad ('6.lO)$$

where a_0 is a constant and a_1 is the linearity coefficient.

Simple linear regression is performed when the requirement is prediction of values of one variable, with given values of another variable. The following example explains the meaning of linear regression.

EXAMPLE 6.3

How can a university student's GPA be predicted from his/her high school percentage (HSP) of marks?

SOLUTION

Consider a sample of ten students for whom their GPAs and high school scores, HSPs, are known. Assume linear regression. Then,

GPA= **b1.HSP** +***A*** …({iJ 1)

Figure 6.5 shows a simple linear regression plot for the relationship between the college GPA and the percentage of high school marks. Plot the values on a graph, with high school scores in percentage on the *x* axis and GPA on they axis.

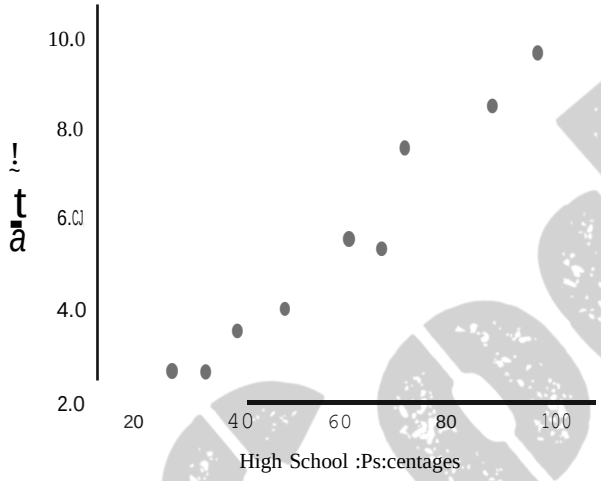


Figure 6.5 Linear regression relationship between college GPAand percentage of high school marks

Whenever a perfect linear relationship between GPA and high school score exists, all 10 points on the graph would fit on a straight line. However, this is never the case. Whenever an imperfect linear relationship exists between these two variables, a cluster of points on the graph, which slope upward, may be obtained. In other words, students who got more marks in high school should get more GPA in college as well.

One variable, denoted by x, is regarded as the predictor, explanatory or independent variable. The other variable, denoted by y, is regarded as the response, outcome or dependent variable.

The purpose of regression analysis is to come up with an equation of a line that fits through a cluster of points with minimal amount of deviation from the line. The best-fitting line is called the *regression line.* The deviation of the points from the line is called an 'error'. Once this regression equation is obtained, the GPA of a student in college examinations can be predicted provided his/her high

school percentage is given. Simple linear regression is actually the same as a correlation between independent and dependent variables.

Figure 6.6 shows a simple linear regression with two regression lines with different regression equations. Looking at the scatter plot, two lines can fit best to summarize the relation between GPA and high school percentage.

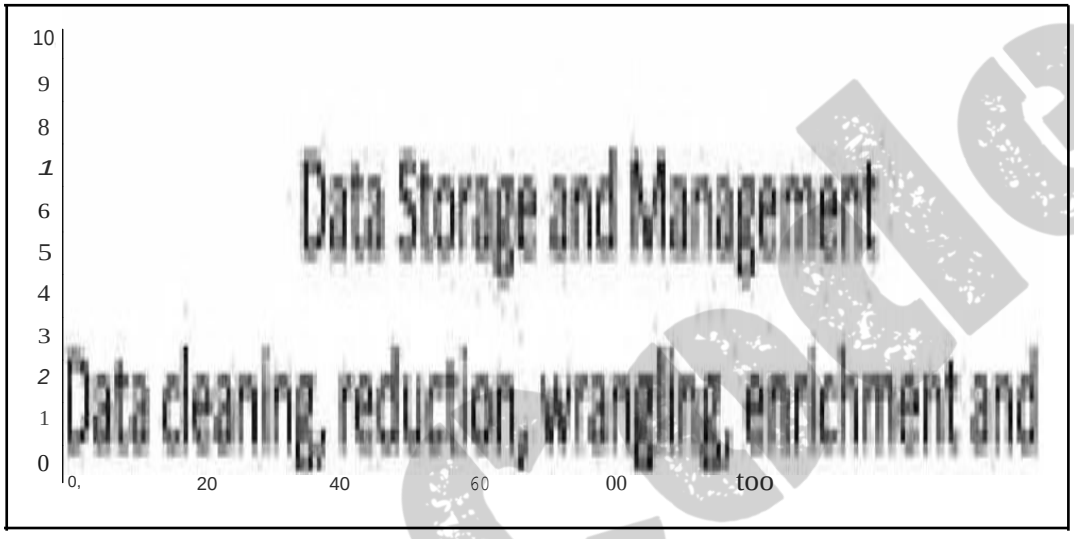


Figure 6.6 Linear regression relationship with two regression lines with different coefficient in regression equation

Following notations can be used for examining which of the two lines is a better fit:

1. *Yi* denotes the observed response for experimental unit *i*
2. *xi* denotes the predictor value for experimental unit **i**
3. *Yi* is the predicted response (or fitted value) for experimental unit **i**

Then, the equation for the best fitting line using a sum of the error estimating function is:

(<iJ 2)

where a'_0 and $a'i$ are the coefficients in Equation (6.10). Use of the above equation to predict the actual response Yi, leads to a prediction error (or residual error) of size:

6.3.2 Least SquareEstimation

Assume *n* data-points, i = 1, 2, ..•, *n.* A line out of two lines (Figure 6.6) that fits the data *best* will be one for which the sum of the squares of the *n* prediction errors (one for each observed data point) is as small as possible. This is the 'least squares criterion', which says that the best fit is one, which 'minimizes the sum of the squared prediction errors'. This implies that when the equation of the best fitting line is:

where b_0 and b_1 are the coefficients which minimize the errors. The coefficients values make the sum of the squared prediction errors as small as possible. Thus,

oop Cloud Service (IBM BigInsight, Microsoft Azu Insights, Oracle Big Data Cloud Services)

(,6. 1 5)

Q is also called chi-square function. To minimize~= .L:t~·,-(b$_0$ + *bLx/,* compute the derivative with respect to b_0 and b_{11} set to 0, respectively, and get the 'least squares estimates' for b_0 and b_1 as follows:

(6.16)

Cloud Services

and

$$b_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{--- (6.17)}$$

The derivative of a dependent variable with respect to the independent variable is also called a gradient. Sections 6.7.1.3 and 6.7.3 describe the use of 'gradient descent', i.e., a gradient's descent towards convergence when optimizing for minimum values of gradient descent.

For obtaining the best-fit line here, the sum of the squared prediction error Q is minimized. Since the objective in the regression analysis is to minimize Q, Q is called objective function.

6.3.3 Multiple Regressions

A criterion variable can be predicted from one predictor variable in simple linear regression. The criterion can be predicted by two or more variables in *multiple regressions.* The following example explains the meaning of multiple regression and coefficients.

EXAMPLE 6.4

Recall Example 6.3 where an assumption that university examination GPA depends on past examination HSP was made. Now assume that GPA depends on HSP as well as internal assessment (IA) at the university.

(i) How will you predict a student GPA on the basis of the HSP and IA during university study?

(ii) What do the coefficients tell?

SOLUTION

(i) Regression analysis requirement is to find a linear combination of HSP and IA that best predicts overall GPA. Regression relation gives GPA:

$$GPA = b_1.HSP + b_2.IA + b_0 \quad (6.18)$$

where b_0, b_1 and b_2 are regression coefficients.

(ii) With multiple independent variables, the coefficients tell how much the dependent (response) variable is expected to increase when the independent (predictor) variable increases by unit value, holding all the other independent variables constant. Remember, the units by which the variables are measured differ for different models. For example, assume $y = 1 + 2x_1 + 3x_2$. When x_2 is constant, for each change of 1 unit in x_1, y changes 2 units.

Multiple regressions are used when two or more independent factors are involved. These regressions are also widely used to make short- to mid-term predictions to assess which factors to include and which to exclude. Multiple regressions can be used to develop alternate models with different factors.

More than one variable can be used as a predictor with multiple regressions. However, it is always suggested to use a few variables as predictors necessarily, to get a reasonably accurate forecast. The prediction takes the form:

(6.19)

where a is the intercept of line on the y axis (means value of y when all independent variable values = 0). The c_1, c_2, ..., and en are coefficients, representing the contributions (weights) of the independent variables x_1, x_2, ..., xn in the calculation of y.

Multiple regression analysis, often referred to simply as regression analysis, examines the effects of multiple independent variables on the value of a dependent variable or outcome.

Statistical significance means that the observer can be confident that the findings are real, and not just a coincidence, for the given data. Regression calculates a coefficient for each independent variable and its statistical significance, to estimate the effect of each independent variable on the dependent variable. An example of a regression study is to examine the effect of education, experience, gender and social background on income.

6.3.4 Modelling Possibilities using Regression

Regressions range from simple models to highly complex equations. Two primary uses for regression are forecasting and optimization. Consider the following examples:

1. Using linear analysis on sales data with monthly sales, a company could forecast sales for future months.
2. For the funds that a company has invested in marketing a particular brand, an analysis of whether the investment has given substantial returns or not can be made.
3. Suppose two promotion campaigns are running on TV and Radio in parallel. A linear regression can confine the individual as well as the combined impact of running these advertisements together.
4. An insurance company exploits a linear regression model to obtain a tentative premium table using predicted claims to Insured Declared Value

ratio.

5. A financial company may be interested in minimizing its risk portfolio and hence want to understand the top five factors or reasons for default by a customer.
6. To predict the characteristics of child based on the characteristics of their parents.
7. A company faces an employment discrimination matter in which a claim that women are being discriminated against in terms of salary is raised.
8. Predicting the prices of houses, considering the locality and builder characteristics in a locality of a particular city.
9. Finding relationships between the structure and the biological activity of compounds through their physical, chemical and physicochemical traits is most commonly performed with regression techniques.
10. To predict compounds with higher bioactivity within groups.

6.3.5 Predictions using Regression Analysis

Regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of another variable. Regression analysis is generally a statistical method to deal with the formulation of a mathematical model depicting the relationship amongst dependent and independent variables. The dependent variable is used for the purpose of prediction of the values. One or more variables whose values are hypothesized are called independent variables. The prediction for the dependent variable can be made by accurate selection of independent variables to estimate a dependent variable.

Two steps for predicting the dependent variable:

1. *Estimation* step: A function is hypothesized and the parameters of the function are estimated from the data collected on the dependent variable.
2. *Prediction* step: The independent variable values are then input to the parameterized function to generate predictions for the dependent variable.

Consider an example of data that contain two variables, viz., crop yield and rainfall. Assume that the yield depends on rainfall (in certain critical growth phases). Using past yield data as a function of rainfall, the crop yield can be predicted. The application of linear regression upon these two variables will generate a linear equation, $y = a + b.x$, where y and x variables denotes crop yield and rainfall, respectively. Constants, a and bare the model's parameters known as the intercept and slope of the equation.

6.3.6 K-Nearest-Neighbour Regression Analysis

Consider the saying, 'a person is known by the company he/she keeps.' Can a prediction be made using neighbouring data points? K-Nearest Neighbours (KNN) analysis is an ML based technique using the concept, which uses a subset of $K = 1, 2$ or 3 neighbours in place of a complete dataset. The subset is a training dataset.

Assume that population (all data points of interest) consist of k-data points. A data point independent variable is xi, where $i = 1$ to k. K-Nearest Neighbours (KNN) is an algorithm, which is usually used for classifiers. However, it is useful for regression also. Predictions can use all k examples (global examples) or just K examples (K-neighbours with $K = 1, 2$ or 3). It predicts the unknown value Yp using predictor variable xP using the available values at the neighbours. The training dataset consists of available values of y_{ni} at x_{ni} with $n, = 1$ to K, where n, is the K-the neighbour, means just the local examples.

A subset of training dataset restricts k to K-neighbours, where $K = 1, 2$ or 3. This means using local values near the predictor variable. $K = 1$ means the nearest neighbour data points. $K = 2$ means the next nearest neighbour data points (xi,YJ $K = 3$ means the next to next nearest neighbour data points (xi, Yi).

First find all available neighbouring target (xi, yi) cases and then predict the numerical value to be predicted based on a similarity measure. Prediction methods are as follows:

(i) Simple interpolation, when predictor variable is outside the training subset

(ii) Extrapolation, when predictor variable is outside the training subset

(iii) Averaging, local linear regression or local-weighted regression.

KNN analysis assumes that weight is inversely proportional to the square of distance (w a n-2), inverse of the distance (a n') or inverse of qth power of the distance (a *v-q)* called Euclidean DEu' Manhattan DMa and Minkowski DMi distances, respectively. When predicting, a weight assignment may require computations using a kernel function1 like a Gaussian or tri-cube function (Section 6.2.5.1) in cases where the dependent variable varies according to the kernel function.

Assume continuously varying values as a function of independent variables. Assume v denotes the number of variables, independent as well as dependent. The following equations give the KNN distances in *v-dimensional space* for the purpose of using weights.

Euclidean Distance The following equation computes the Euclidean distance DEu:

Sum of the squared Euclidean distance, [r»; L--"i:iiJ]2 = ["" ..£...i=J (x - x~i J. and

Euclidean distance DEu = [L;~1 *(xi* - .t; >2 J^lf2^ (6.20~.")

Sum is over v dimensions. If one independent and one dependent variable, then v = 2. For example, ifv = 2 and two data points are *(xyy)* and (x_;+i,Yj+1), then Euclidean distance between the points is as follows:

Euclidean distance DEu = [(xj - xj+1)2 + (yj - Yj+1)2]1!2 (6.20b)

Euclidean distance for three variables v = 3 (two independent variables and one dependent variable case) consists of three terms on the right-hand side in Equation (6.20b).

Manhattan Distance The following equation computes the Manhattan distance DMa:

Manhattan distance DMa

DMa = L;~J 0-\- x;j] c_6.20c)

Manhattan distance for three variables v = 3 (two independent variables and one dependent variable case) consists of three terms on the right-hand side in Equation (6.20c).

Comparison between Euclidean and Manhattan Distances Basically,

Euclidean distance is the direct path distance between two data points in v• dimensional metric spaces. Manhattan distance is the staircase path distance between them. Staircase distance means to move to the next point, first move along one metric dimension (say, *x* axis) from the first point, and then move to the next along another dimension (say, *y* axis).

When v = 2, Euclidean distance is the diagonal distance between the points on an *x-y* graph. Manhattan distances are faster to calculate as compared to Euclidean distances. Manhattan distances are proportional to Euclidean distances in case of linear regression.

Minkowski Distance The following equation computes the Minkowski distance DMi:

$$\text{Minkowski distance } DM{:}i = \left\{ \sum_{\iota} [c\,xi - xf/l] \right\}^{liq} \qquad (6.20d)$$

Hamming Distance When predictions are on the basis of categorical variables, then use the Hamming distance. It is a measure of the number of instances in which corresponding values are found.

$$\text{Hamming Distance} \sim\sim = L,;Jx.- - \cdot 11^{\sim} \qquad (.6.10e)$$

when xi= xe, then DH= 0 and when xi not equal to xc, then DH= 1. For example, Hamming distance DH= 1 between 10100111100 and 11100111100 because just one substitution is needed, change second bit from Oto 1 at 10th place from the right to left positioned bits. Hamming distance DH= 4 between 111001 00000 and 011001 11100 because we need four substitutions, change 3rd, 4th, 5th and from 0 to 1 and 11th bit from 1 to 0

An application is in text analytics. Hamming distance DH = 3 between 'Bank notes' and 'Java notes'. The distance= 3 because the required number of changes is 3 at B, n and k among two strings. Another application of Hamming distance is in counting the number of data points off from the regression curve (Refer Section 6.4.4.6). Another application is in counting the wrong or distinct characters when comparing two document sentences.

Normalization Concept Normalization factor in p-norm form in a v• dimensional space is

$$x, = Jr \cdot x, \text{ where } N = \left(L;\sim| lxir\ rip \right. \qquad c\,6.:!\ l)$$

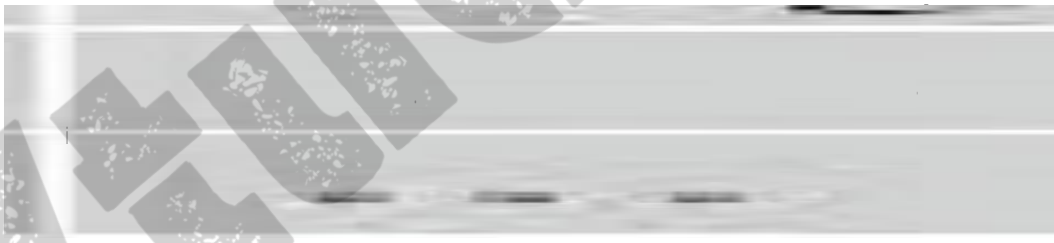
Here, *xi* is *fh* component of the vector X. The total number of components are v. Two-dimensional space v = 2, three-dimensional v = 3.

The following example explains the meaning of distances, use of Euclidean and Manhattan distances, use distances for predictions, and the KNN regression analysis.

EXAMPLE 6.5

Assume dataset S with two subsets of sets Jspi and Zspi for sales and sales percent increase (SPI) for Jaguar Land Rover and Zest models of Tata Motors Company. Assume S is training dataset and consists of data points as per the following table.

Table 6.2 An example of two car models,Jaguar, and Zest (JLRS and ZS), sales and sales percent increase (SPI) in years between 2012 and 2018



M = 0 means Jaguar Land Rover, and M = 1 means Zest.

(i) Draw two plots, one with scatter set of points with Y and ZS, which means columns 1 and 5 data, second plot between Yb and Jspi with columns 2 and 4.

(ii) Find the Euclidean 2-NN distance between third and first row data points (2014, 11232) and (2012, 10000).

(iii) What is the Manhattan 2-NN distance between the third and first row data points (2014, 11232) and (2012, 10000)?

(iv) What are seven Hamming distance terms of Equation (6.20e) between fourth and six column vectors Jspi and Zspi? Interpret the summed DHa·

(v) Assume data point for JLRS as missing for 2015. How do you predict car sales in 2015 assuming missing row for 2015? Use Euclidean distances using 1-NN. How do the results differ when using 2-NN and 3-NN?

(vi) How do you predict car sales for 2011? Use Euclidean distances.

(vii) How will you use 1-NN, 2-NN and 3-NN for estimation regression coefficient?

(viii) How will you calculate Euclidean distances DEu between values Jspi in columns 4 for Yb = 3 and 5?

(ix) How will you calculate DEu between value for column 2 Yb = 0 and column 2 Zspi value in column 6 for Yb = 1, 3 and 4?

SOLUTION

(i) Figure 6.7 shows scatter set of points one for Y and ZS, which means data points in columns 1 and 5, and second for Yb andJspi' which means data points in columns 2 and 4.

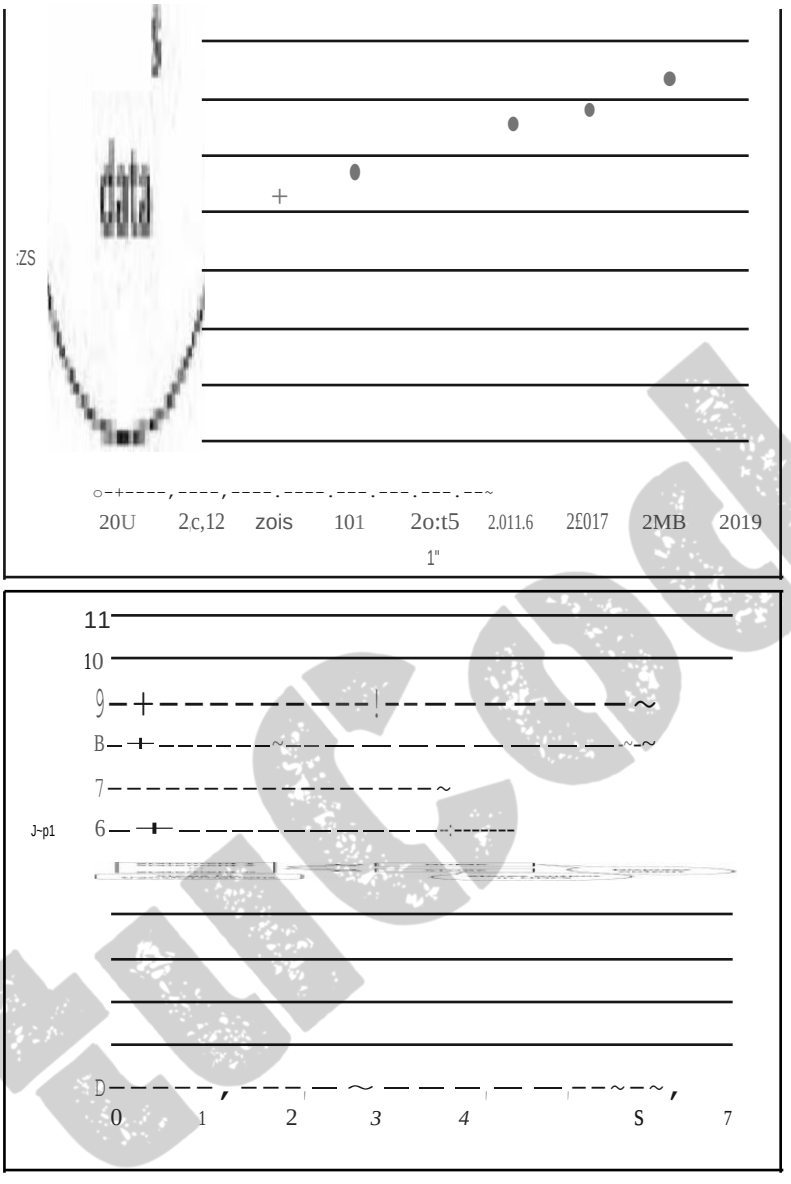


Figure 6.7 Scatter plots for two set of data points, one between Y and ZS, and second between Yb andJspi

(ii) Using the equation, $DEu = [(X_i - X_{i+1})2 + (yj - yj_{+1})2]\backslash$ find the Euclidean 2-NN distance between third and first row data points (2014, 11232) and {2012, 10000) $DEu = [\{2014 - 2012)_2 + \{11232 - 10000)2]Yz = [\{2)_2 + \{1232)2]Yz = 1232.001$.

(iii) Using the equation, $DMa = [(xj - xj_{+1)} + (yj - Yj_{+1)}]$. Manhattan 2-NN distance between third and first row data points (2014, 11232) and (2012, 10000)= [(2014- 2012)+ (11232- 10000)]= [2 + 1232]= 1234.

(iv) Hamming distances need to compute between fourth and six column vectors Jspi and Zspi are {1, 0, 0, 0, 0, 1, 0} because only in these the Jspi and Zspi differ. That also means that in two years out of seven, increase in sales percentage differs for Jaguar Land Rover and Zest models.

(v) Lets JLRS missing for 2015 (independent or predictor variable). Since 2014 and 2016 are its 1-NN. Let us choose 1-NN of year 2014, that is 2013. DEu (2014, 2013) = v{(2014- 2013)$_2$ + (1123-1040)2} = 83. Predicted JLRS (2015) = 1123 + 83 = 1246 by extrapolation, assuming DEu (2014, 2013) = DEu (2014,2015).(weight factors 1)

Years 2012 and 2016 are 2-NNsof 2014. Let us consider DE)2014, 2016) = 175. Thus, the predicted JLRS(2015) = (1298 - 175/2) = 1210 using interpolation (weight factor= 1 per year change).

Similar computations can be made for DE)2014, 2017) as 3-NN of 2014 is 2017. DEu 3-NN= 305. PredictedJLRS(2015) = (1123+ 305/3) = 1225.

(vi) Predicting the car sales for 2011 is an example of extrapolation, when predictor variable is outside the training subset. JLR(2012) is closet point. DEu (2012,2013)= 40. PredictedJLRS(2011) = 1000 - 40 = 960.

(vii) K-NN algorithm is used for estimating regression coefficient. For example, use a weighted average of the k-nearest neighbours, weighted by the inverse of their distance. Compute the Euclidean from the query example to the labeled examples.

1. Order the labeled examples by increasing distance.
2. Find a heuristically optimal number *k* of nearest neighbours.
3. Calculate an inverse distance weighted average with the k-nearest multivariate neighbours.

(viii)Euclidean distances between values Jspin in columns 4 for Yb = 3 and 5

DE11 = [(Y~ -Y~) ' 2 +c_J5P~ -JJPi5) 2]1n

= ((3 - 5.,2 + (9 - 10/]1!2 = ((-1/ + (-1/in

DEu = [5]~ = :!.~36

(ix) DEu between its value for column 2 Yb = 0 and value of Zspi in column 6 for Yb= 1, 3, and 4

$$D_{Eu} = \left[(Y_{b_0} - Y_{b_1})^2 + (Z_{spi_0} - Z_{spi_1})^2\right]^{1/2}$$
$$= [(0-1)^2 + (3-4)^2]^{1/2} = [(-1)^2 + (-1)^2]^{1/2} = 1.414$$
$$D_{Eu} = \left[(Y_{b_0} - Y_{b_3})^2 + (Z_{spi_0} - Z_{spi_3})^2\right]^{1/2}$$
$$= [(0-3)^2 + (3-9)^2]^{1/2} = [(-3)^2 + (-6)^2]^{1/2} = 6.708$$
$$D_{Eu} = \left[(Y_{b_0} - Y_{b_4})^2 + (Z_{spi_0} - Z_{spi_4})^2\right]^{1/2}$$
$$= [(0-4)^2 + (3-6)^2]^{1/2} = [(-4)^2 + (-3)^2]^{1/2} = 5.0$$

Self-Assessment Exercise linked to LO 6.2

1. How does regression analysis predict the value of the dependent variable in case of linear regression?
2. (i) Define objective function for least square fitting of coefficients in regression equation.
   (ii) How are the best-fitting regression coefficients evaluated?
3. When are multiple regressions used? How do multiple regressions predict intermediate term? How do multiple regressions assess which factors to include and which to exclude? How do multiple regressions help in developing alternate models with different factors?
4. How is KNN regression used for predicting, considering two variables and K = 3? Use training dataset given in Example 6.5.
5. How do KNN regression computations differ when using Euclidean and Manhattan distances? Consider two variables and K = 3. Use the training dataset given in Example 6.5.

6.41 FINDING SIMILAR ITEMS, SIMILARITY OF SETS AND COLLABORATIVE FILTERING

MapReduce

Fi1i1di'ng si mii~r items, applications ofNear• NEigliilbouir search. Jacc::ard similairiey of —s; sirmil:riiiicy of documents; col la borativa filtering 3s ~ sJirmilar..S@t proiiil'en;i: and Eu ~lidea. Jaccardl., Cosi11T1e. ~dlit and Ham m ifilg cfiis.ta ncas

Similar item search refers to a data mmmg method which helps in discovering items which have similarities in datasets. *(Data mining* means discovering previously unknown interesting patterns and knowledge from apparently unstructured data. The process of data mining uses the ML algorithms. Data mmmg enables analysis, categorization and summarization of data and relationships among data.)

The following subsections describe methods of finding similar items using similarities, application of near-neighbour search, Jaccard similarity of sets, similarity of documents, Collaborative Filtering (CF) as a similar-set problem, and the distance measures for finding similarities.

6.4.1 Finding Similar Items

An analysis requires many times to find similar items. For example, finding similar excellent performance of students in Python programming, similar showrooms of a specific car model which show high sales per month, recommending books on similar topic such as in Internet of Things by Raj Kamal from McGraw-HillHigher Education, etc.

6.4.1.1 Application of Near Neighbour Search

Similar items can be found using Nearest Neighbour Search (NNS). The search finds that a point in a given set is most similar (closest) to a given point. A dissimilarity function having larger value means less similar. The dissimilarity function is used to find similar items.

NNS algorithm is as follows: Consider set S having points in a space **M.** Consider a queried point q EM, which means q is member of M. k-NNS algorithm finds the k-closet (1-NN) points to q in S.

Three problems with the Pearson similarities (6.2.6.1):

1. Do not consider the number of items in which two users' preferences overlap. (e.g., 2 overlap items==> 1, more items may not be better.)
2. If two users overlap on only one item, no correlation can be computed.
3. The correlation is undefined if series of preference values are identical.

Greater distance means greater dissimilarity. Dissimilarity coefficient relates to a distance metric in metrics space in v-dimensional space. An algorithm computes Euclidean, Manhattan and Minkowski distances using Equations (6.20a) to (6.20d).

Distance metric is symmetric and follows triangular inequality. Meaning of triangular inequality can be understood by an example. Consider three vectors of lengths *x.y,* and z. Then, triangular inequality means $z < x + y$. It is similar to the theorem of inequality that the third side of a triangle is less than the sum of two other sides, and never equal. The theorem applies to v-dimensional space also. Dissimilarity can be asymmetric, i.e., triangular inequality is not true (Bergman divergence).

Consider a linear search (also referred as Naive search) algorithm, Naive, one of meaning is *simple* in English. Search requires computations of distances to every other point. The algorithm running time is large. The time function, O (v.c) which measures the efficiency of the search algorithm in terms of means v.c. The v is dimensionality of M and c is cardinality of S. Cardinality refers to the number of relationships. For example, one independent variable and two dependent variables in a relationship, then cardinality is 3. Cardinality in the context of databases means the uniqueness of values contained in a column fields.

Note: Space partitioning followed by the search algorithm is an efficient method using a k-d tree or R-tree data structure. Search is made after arranging the tree-like data structure. Space partitioning problems become complex in case of high dimensionality.

Naive search algorithm outperforms space partitioning approaches when using high dimensional spaces Mand high cardinality.2

The following example explains the NNS approach to find similar items.

EXAMPLE 6.6

Assume a set S consists of data of a large number of students. The dataset consists of grade points (GPs) in each of the five subjects of study in a semester. The total dataset is for six semesters. Each semester examination awards SGPAs (Semester Grade Point averages). CGPAi (Cumulative GPA of ith semester) calculates after end of ith semester after adding the SGPAs of previous semesters. Assume that each student GP is on a 10-point scale. A student performance in a subject is high (H) if GP is 8.0 or close within ±1.0. A student performance in a subject is excellent (E) if GP is 9.0 or close by within ±1.0.

(i) How will you choose independent and dependent variables? What does metric space mean?

(ii) How will you define a metric space for finding similar performances in a specific subject? How will you define a metric space M for finding similar performance from SGPAs of the first semester? How will you define a metric space for finding similar performances from CGPAs of a semester?

(iii) What will you consider S for finding similarities by NNS?

(iv) What does nearest neighbour search mean when search is for students with similar excellent performance?

(v) How will you find students of similar excellent performance by the GPs of a subject, say Java Programming in the second semester?

(vi) How will you find similar excellent performances by the CGPA?

(vii) How will you find similar high performances by the SGPA?

(viii) How will you compute Euclidian and Manhattan distances with respect to query point GP = 8.0 ±1.0? How will you compute dissimilarity?

(ix) What do you mean by dimensionality of M? What do you mean by cardinality of S?

SOLUTION

(i) Independent variables are student ID, year of study, semester period, name and type (theory or practical) of five subjects. Dependent variables are GP, GPA, SGPA and CGPA. Metric space means a space in which variables are quantifiable. For example, GP, GPA, SGPA and CGPA.

(ii) Metric space for finding similar performances in a specific subject, Metric space M for finding similar performance in SGPA of first semester, Metric space for finding similar performances from CGPA of a semester:

(iii) Members of set S are input vectors, each having elements {studentID, CGPA [or SGPA, GPA,
T_GPA (GPA of theory subject), P_GPA (GPA of practical subject)]} for each student for finding the similarities by NNS using three distances Dl, 02, 03 of first, second and third nearest neighbours.

(iv) Nearest neighbour search for students with similar excellent performance means search of studentIDs awarded CGPA within the distance 1.0 from 9.0.

(v) Students of similar excellent performance by the GPs of a subject, say Java Programming, in the second semester means Student IDs with GPs inJava programming within the distance ±1.0 from 9.0.

(vi) Similar excellent performance by the CGPA means similar performance of students_IDs with CGPA within the distance ±1.0 from 9.0.

(vii) Similar high performance by the SGPA means similar performance of students_IDs with SGPA within the distance ±1.0 from 8.0.

(viiikomputation of Euclidian distances with respect query point GP = 8.0 ± 1.0

Computation of Manhattan distances with respect query point GP = 8.0 ± 1.0

(ix) Dimensionality of M equals the number of independent and dependent variables in metric space for which distances are quantifiable.

Cardinality of S means number of relationships, number of independent but unrelated and dependent unrelated variables. For example, subject_name and subject_type is related to each other. Therefore, subject_name and subject_type are counted as one variable when computing cardinality.

6.4.2 Jaccard Similarity of Sets

Let A and B be two sets. Jaccard similarity coefficient of two sets measures using notations in set theory as shown below:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$A \cap B$ means the number of elements or items that are same in sets A and B. $A \cup B$ means the number of elements or items present in union of both the sets. Assume two set of students in two computer courses, Computer Applications CA, and Computer Science CS in a semester. Set CA 40 students opted for Java out of 60 students. Set CS 30 students opted for Java out of 50 students. Jaccard similarity coefficient $J_{java}(CA, CS) = 30/(60 + 50) \times 100\% = 27\%$. Two sets are sharing 27% of the members for Java course.

($\cap$ is symbol for intersection in set theory. $\cup$ is symbol for union in set theory.)

6.4.2.1 Similarity of Documents

An application of Jaccard similarity coefficient is in Natural Language Processing (NLP) and text processing. It quantifies the similarity in documents. Computational steps are as follows:

1. Find Bag of Words (Section 9.2.1.4) and remove words such as is, are, does, at, in, ...
2. Assign weighting factor is the Term frequency and Inverse Document Frequency (TF-IDF). Consider the frequency of words in the document.
3. Find k-shingles. A shingle is a word of fixed length. The k-shingles are the number of times the similar shingles extracted from a document or text. Examples of a shingle are Java, GP, 8.0, Python, 80%, Programming.
4. Find n-grams. A gram is a contiguous sequence of fixed length item (word

or set of characters, letters, words in pairs, triplets, quadruplets, ...) in a document or text. The n-grams are the number of times the similar items (1-grams, 2-grams, ..) extracted from a document or text. The 3-gram examples are lava GP 8.0, Python Programming 7.8, Big Data Analytics, 23A 240C 8LP, the numbers of which are extracted from the text.

5. Compute Jaccard similarity coefficient using Equation (6.22) between the documents.

A number of other methods exist for computing similarity of documents. One method is Latent Semantic Indexing method (LSI). The computational steps are as follows:

- Steps 1 and 2 are the same as above.
- Consider documents into word space. Reduce dimensionality of the projection space. An algebraic model is one that represents text documents as vectors or identifiers, such as how many times a word is present in a document, the index terms or deploy singular value decomposition method.
- Use Cosine Similarity measure between the documents.

Refer Section 9.2 for details on text analysis.

6.4.3 Collaborative Filtering as a Similar-Sets Finding Problem

An analysis requires finding similar sets using collaborative filtering. Collaborative filtering refers to a filtering algorithm, which filters the items sets that have similarities with different items in a dataset.

CF finds the sets with items having the same or close similarity coefficients. Following are some examples of applications of CF:

- Find those sets of students in computer application, and computer science who opt for the Java Programming subject in a semester.
- Find sets of students in Java Programming subjects to whom same teacher taught and they showed excellent performance.

An algorithm finds the similarities between the sets for the CF. Applications of

CF are in many ML methods, such as association rule mining, classifiers, and recommenders.

6.4.4 Distance Measures for Finding Similar Items or Users

Distance measures compute the dissimilarities. Complement of dissimilarity gives similarity. The following subsections describe the distance measures.

6.4.4.1 Definition of a Distance

Distance can be defined in a number of ways. Distance is the measure of length of a line between two values in a two-dimensional map or graph. Set of Equations (6.20) measures distances.

For example, distance between (2014, 6%) and (2018, 8%) on a scatter plot when year is on the x axis and profit% on they axis is Distance = $\sqrt{[(2014 - 2018)^2 + (6 - 8)^2]} = \sqrt{(16 + 4)} = 4.47$, using Equation (6.20b). Distance can also be similarly defined in v-dimensional space using Equation (6.20a).

Distances between all members in a set of points can be computed in metrics space using a mathematical equation. Metrics space means measurable or quantifiable space. For example, profit and year on a scatter plot are in metric space of two dimensions. Probability distribution function values are in metric space.

Consider student-performance measures 'very good' and 'excellent'. These parameters are in non-metric space. How are they made measurable? They become measurable when very good is specified as grade point average 8.5 which implies that a score between 8.0 to 9.0 is very good, and define 9.5 which implies that a score between 9.0 to 10.0 is excellent on a 10-point scale.

Consider a chart between number of students passing in examination with best grades vs languages C++, Java, Node.js and Python. Languages are in non-metric space. They become measurable when numbers, say 0, 1, 2 and 3 are assigned for a language for the purpose of using distance measure for similarity analysis.

Distance can be defined as the reciprocal of weight in v-dimensional space. For example, a point at unit distance can be taken as weight w = 1, and a point at distance = 2, w = ½ and so on.

Distance can also be defined as dissimilarity coefficient in v-dimensional

space. Greater distance means greater dissimilarity. Subtracting dissimilarity coefficient from 1 gives similarity coefficient. Many different algorithms exist to compute distance and thus similarity between entities, number of users or items. An algorithm computes the distances DEu, DMa, DMi' DHa [Equations (6.20a to e)] or any other distance metric, for example, Jaccard distance *D1a,* cosine distance *Dew* edit distance DEd·

Jaccard similarity, Cosine similarity, edit distance or correlation methods are used to find out similarities between users.

6.4.4.2 Euclidean Distance

Euclidean distance ~u= [r:]~.--x/Jt2 refer Equation (6.20a) in Section 6.3.6 for details.)

6.4.4.3 Jaccard Distance

Equation (6.22) gives J (A, B). Jaccard distance, DJa (A, B) measures the dissimilarity between two sets. It is equal to result of subtraction of Jaccard similarity coefficient] (A, B) from 1.

$$Dla(a', ffi) = 1 - J(8, -3) \qquad (0.23'')$$

(Refer Section 6.4.2 for details.)

6.4.4.4 Cosine Distance

Cosine similarity is a measure of similarity in the inner-product space between two vectors of finite magnitudes. Cosine distance Dcos is measure of dissimilarity between vectors. A measure of cosine distance is in terms of the angle between the vectors. Cosine similarity has low complexity. Cosine distance has applications in text mining, finding similarity of documents, and similarities in sparse vectors, column-vectors (fields) and matrices (Section 3.3.3.1).

Let U and V be two non-zero vectors, two documents in the vector space.

$$Dcai<.ll\sim V) = \frac{L\ u.r}{\widetilde{Kv}\ \widetilde{\pounds.i,}} \qquad (6.23£l)$$

where *U,* and *Vi* are components of U and V, respectively, and summation in numerator is over *i* = 1 to *N,* where N is the number of elements of the vectors.

No Triangular Inequality Property Cosine distances do not exhibit triangular

inequality property, while the Euclidean distances exhibit triangular inequality (Section 6.4.1.1).

Vector Cosine-Based Similarity Vector cosine similarity in terms of angle between two vectors U and V is given by equation:

$$\sim \ldots = \cos\text{-}1\ (U, V, = \frac{U.V}{\|U\|\|V\|} \quad (6.23h)$$

Consider Example 6.5. Let each model have Sales Percentage Increase (SPI) values in successive years. The similarity between SPIs of two models, M_1 and M_2, is measured by treating each model as a vector of SPIs and computing the cosine of the angle formed by the SPI vectors.

Formally, if P is $m \times n$ SPI matrix for a model M, then the similarity between two models, Mi and Mj is defined as the cosine in the n-dimensional vectors space corresponding to the ith and /h columns of P.

The following example illustrates computing of cosine and Euclidean similarities to find similar items.

EXAMPLE 6.7

Consider members of a dataset S in five-dimensional metric space. Assume S subsets are JLR and Z. Data subset members consist of values in two column vectorsJspi and Zspi of the elements SPIs. Data subset JLR consists of percentage increase in sales number in a year of Tata Jaguar Land Rovers cars, and Z consists of Zest cars SPIs. (Example 6.5) Assume dataset S consists of data points as per Table 6.2.

(i) Represent members of dataset S of table data points in five• dimensional metric space consisting of three independent variables Y, Yb, Mand two dependent variables Jspi and Zspi·

(ii) Represent the members of six subsets for values in columns 1, 2, 3 and 5 as elements of vectors Y, Yb, and matrices (M, Jspi) and (M, Zspi), respectively.

(iii) Represent the table data points in two-dimensional metric spaces (Yb, Jspi) and (Yb, Zspi).

(iv) Represent the table data points in three-dimensional metric space (Yb, Jspi, Zspi).

(v) How will you calculate the cosine distance, cosine similarity and angle between the vectors Jspi and Zspi?

(vi) How will you calculate Euclidean similarity using six neighbour distances DEu starting from Zspi for Yb= 0 and Zspi values in columns 6 for Yb= 1 to 5?

SOLUTION

(i) 8{2012, 0, (0,5), (1,3)}, {2013, 1, (0,4), (1,4)}, {2014, 2, (0,8), (1,8)}, {2015, 3, (0,9), (1,9)}, {2016, 4, (0,6), (1,6)}, {2017, 5, (0,10), (1,10)}, {2018, 6, (0,8), (1,8)}

(ii) Y = {2012, 2013, 2014, 2015, 2016, 2017, 2018}, Yb= {0, 1, 2, 3, 4, 5, 6}, (M, Jspi = {{0,5), (0, 4), (0, 8), (0, 9), (0. 6), (0, 10), (0, 8)}, and (M, Zspi) = {(1, 3), (1, 4), (1, 8), (1, 9), (1, 6), (1, 4), (1, 7)}

(iii) (Yb, Jspi) = {(0,5), (0,4), (0,8), (0,9), (0,6), (0,10), (0,8)} and (Yb, Zspi) = {{0,3), (0,4), (0,8), (0,9), (0,6), (0,4), (0,8)}

(iv) (Yb,Jspi' Zspi) = {{(0,5, 3), (1,4, 4), (2,8, 8), (3,9, 9), (4,6, 6), (5,10, 4), (6,8, 8)}.

(v) Dcos Ospi, Zspi) = {(5 x 3) + (4 x 4) + (8 x 8) + (9 x 9) + (6 x 6) + (10 x 4) + *(Bx* S)}/V {5_2 + 4_2+ S_{2+} 9_2 + 6_2+ 10_2 + S_{2}} x v{3_2 + 4_2+ S_{2+} 9_{2+} 6_2+ 4_2 + S_{2}} = o.951

Cosine similarity= 1 - Dcos Ospi, Zspi) = 1- 0.951 = 0.049

Angle betweenjspi, Zspi = $\cos^{-1}$(Dcos) = 87.191

(vi) Use Equation (6.20b) for the computations.

1. DEu (Yb= 0, Yb= 1) = v{$(0-1)_2$ + (3 - 4)2};

2. DEu (Yb= 1, Yb= 2) = v{(l - 2)2 + (4 - 8)2};

3. DEu (Yb= 2, Yb= 3) = v{$(2 - 3)_2$ + (8- 9)2}

4. $DEu(Yb= 3, Yb= 4) = v\{(3 - 4)2 + (9 - 6)2\}$

5. $DEu(Yb= 4, Yb= 5) = v\{(4 - 5)^2 + (6 - 4)^{2}\}$

6. $DEu(Yb= 5, Yb= 6) = v\{(5 - 6)2 + (4 - 8)2\}$

Euclidean similarity coefficient= 1 - v'[Sum of all square of all six DEu values using 1-NN} divided by {v(6^2 + Sum of square of all six Jspi values)}

(vii) Euclidean similarity= $1 - v\{2^2 + 17^2 + 2^2 + 10^2 + 5^2 + 17^{2}\}/ \{v(6^2 + 386\}$ = -0.305

Differing Similarity Coefficients for SPis Calculated from Cosine distances and Euclidean Distances The following section explains the use of cosine distance and the situations in which Dcos does not find similarity correctly.

Consider a comparison between the cosine and Euclidean similarities when finding similar items. Several situations exist in which predictions from two computational approaches differ. The reason is that triangular inequality holds true for Euclidean distances, while does not hold true for cosine distances.

Certain dimensions have widely different values. For example, let us compare sales JLRS and ZS in column 3 and 5 of Table 6.2. ZS values are nearly ten times the value of JLRS values. A solution is normalizing the values in all dimensions by dividing with the mean values using Equation (6.21). However, that also may give differing and incorrect results using Dcos·

Cosine singularity is found to exhibit correct results for similarities in text documents. Cosine similarity is very efficient to evaluate situations of sparse vectors and those where one needs to consider non-zero values in the dimensions.

Concept of Sparse and Dense Vectors Sparse vector uses a hash-map and consists of non-zero values. Hash-map is a collection, which stores data in (key• value) format (Section 3.3.1). Format is also called random access. Hashing means to convert a large value or string into shorter value or string so that indexing for searching is fast.

For example, assume a vector, which consists of array elements, (subject,

number of students opting, average GPA).

1. Dense vectors have elements (Hive, 40, 8.0), (lava, 30, 8.5), (FORTRAN, 0, 0), (Pascal, 0, 0). Dense vector consists of all elements, whether the element value is 0 or not 0.
2. Sparse vectors will be two only with elements (4, 40, 8.0) and (3, 30, 8.5). Random access Sparse vector means access to elements (key, value pairs) using key. Sparse vector consists of elements for which key is such that value is not 0 (Section 3.3.1).
3. Sparse vector has an associated hash-map in form of a hash-table. First row- Pascal, 1, second row- FORTRAN, 2, third row- Java, 3 and fourth row-Hive.
4. Hashing is a process of assigning a small number or small-sized string indexing, searching and memory saving purposes. Hash process uses a hash function, which results into not-colliding values. In case of two colliding numbers, the process assigns a new number. Sequential access sparse vectors mean two parallel accessing vectors, i.e., one to access keys and the other for values.

6.4.4.5 Edit Distance

Edit distance DEd is a distance measure for dissimilarity between two set of strings or words. DEd equals the minimum number of inserts and deletes of characters needed to transform one set into another. Applications of edit distances are in text analytics and natural language processing, similarities in DNA sequences etc. DNA sequences are strings of characters.

Levenshtein suggested a method for finding edit distance, minimum number of operations of deletion, insertion or substitution of a character in a set of strings to transform one into another. The cost of substitution is taken as 2. Thus, edit distance from computation using that method is also called the Levenshtein

method.3

6.4.4.6 Hamming Distance

If both **U** and V are vectors, Hamming distance DHa is equal to the number of

different elements between these two vectors. Recall Example 6.5 (iv) for Hamming distance between Jspi and Zspi. Hamming similarity-coefficient between car models Jaguar Land Rover and Zest is (1- 2/7) = 0.7. [70%]

If M is a matrix, then DH_a is equal to the number of different elements between the rows of M ignoring the columns.

DH_a between two strings of equal length is the number of positions at which the corresponding characters differ. DH_a is also equal to the minimum number of substitutions required to transform one string into the other. DH_a is also equal to the minimum number of errors that need correction using transformation or substitution.

Hamming distance is therefore another distance measure for measuring the edit distance between two sets of strings, words or sequences.

Self-Assessment Exercise linked to LO 6.3

1. Why is triangular inequality in a distance measure important?
2. How will you compute Jaccard similarity coefficients between datasets for Jspi and Zspi. Use data Table 6.2 as the training dataset.
3. Why does similarity in documents computed?
4. Write applications of Euclidean, Jaccard, Cosine, Edit and Hamming distance measures.
5. Explain how Euclidean, Jaccard, Cosine and Hamming distance measures can be applied for analyzing the dataset given in Table 6.2?

6.5 | FREQUENT ITEMSETS AND ASSOCIATION RULE MINING

The following subsections describes frequent itemset mining, market basket model, association rules mining, and their applications.

6.5.1 Frequent Itemset Mining

Extracting knowledge from a dataset is the main goal of data analytics and data

mining. Data mining mainly deals with the type of patterns that can be mined. A method of mining is Frequent Patterns (FPs) mining method. Frequent patterns occur frequently in transactional data.

LO 6.4

Frequent ItelilTI-setiS ñilig fl 1,. applications of IFI market DaSk@t rn oo~l, if11Tiiiil@ tlirte a ssociatiom rules, U5.e! of Apiiori algoiriliih m, e ahill:rtdicm of candid ate rulas, app'lica:t:i:ons. off associ31Diolill rules, and filiildi n g the JSSOGiatiOlll and Siliflilarit'

Frequent itemset refers to a set of items that frequently appear together, for example, Python and Big Data Analytics. Students of computer science frequently choose these subjects for in-depth studies. *Frequent itemset* refers to a frequent iternset, which is a subset of items that appears frequently in a dataset.

Frequent Itemset Mining (FIM) refers to a data mining method which helps in discovering the itemsets that appear frequently in a dataset. For example, finding a set of students who frequently show poor performance in semester examinations. *Frequent subsequence* is a sequence of patterns that occurs frequently. For example, purchasing a football follows purchasing of sports kit. *Frequent* substructure refers to different structural forms, such as graphs, trees or lattices, which may be combined with itemsets or subsequences.

FIM is one of the popular techniques to extract knowledge from data. The technique has been an essential part of data analysis and data mining. The extraction is based on frequently occurring events. An algorithm specifies a given minimum frequency threshold for considering an itemset as frequent. The extraction generally depends on the specified threshold.

FIM finds the regularities in data. Frequent itemset mining is the preceding step to the association rule learning algorithm. Most often the algorithm is used for analyzing a business. For example, customers of supermarkets, mail order companies and online shops use FIM to find a set of products that are frequently bought together. This provides the knowledge of important pairs of items that occur much more frequently than the items bought independently. A sales person can learn the pattern of what should be bought together for sales.

The analysis results in:

- Improvement of arrangement of products in shelves and on catalog pages
- Marketing and sales promotion
- Planning of products that a store should stock up

- Support cross-selling (suggestion of other products) and product bundling.

6.5.2 Association Rule- Overview

An important method of data mining is association rule mining or association analysis. The method has been widely used in many application areas for discovering interesting relationships which are present in large datasets. The objective is to find uncovered relationships using some strong rules. The rules are termed as association rules for frequent itemsets. Mahout includes a 'parallel frequent pattern growth' algorithm. The method analyzes the items in a group and then identifies which items typically appear together (association) (Section 6.8). A formal statement of the association rule problem is:

Let I = {I_1, I_2, ..., I_d} be a set of d distinct attributes, also called literals. Let **T** = {t_1, t_2, ..., *t)* be set of *n* transactions and contain a set of items such that **T c** I. An association rule is an implication of the form, *X* _.. *Y,* where X, *Y* belong to sets of items called itemsets (X, *Y* c I), and X and Y are disjoint itemsets (X ∩ *Y* = 0). Here, *X* is called antecedent, and *Y* consequent.

Explanation:

1. c means 'subset of, c means 'proper (strict) subset of, ∩ means intersection and 0 means disjoint, no commonality in members.
2. Consider an If() then () form of a rule. The *Ifpart* of the rule *(A)* is known as *antecedent* and the *THEN* part of the rule (B) is known as *consequent.* The condition is *antecedent.* Result is *consequent.*

6.5.3 AprioriAlgorithm

Apriori algorithm is used for frequent itemset mining and association rule mining. Apriori algorithm is considered as one of the most well-known association rule algorithms. The algorithm simply follows a basis that any subset of a large itemset must be a large itemset. This basis can be formally given as the Apriori principle. The Apriori principle can reduce the number of itemsets needed to be examined. Apriori principle suggests if an itemset is frequent, then all of its subsets must also be frequent. For example, if itemset {A, B, C} is a frequent itemset, then all of its subsets {A}, {B}, {C}, {A, B}, {B, C} and {A, C} must be frequent. On the contrary, if an itemset is not frequent, then none of its

supersets can be frequent. This results into a smaller list of potential frequent itemsets as the mining progresses.

Support is an indication of how popular an itemset is. That is the frequency of the itemset for appearing in a database.

Assume X and Y are two itemsets. Apriori principle holds due to the following property of support measure:

$$\forall X, Y: (X \subset Y) \Rightarrow s(X) \geq s(Y) \quad (6.24)$$

Explanation: $\forall$ means for all, and $\subset$ means 'subset of and can be 'equal to or included in'. Support of an itemset never exceeds the support of its subsets. This is known as the *anti-monotone property* of support.

The algorithm uses k-itemsets (An itemset which contains k items is known as a k-itemset) to explore (k-1)-itemsets in order to mine frequent itemsets from transactional database for the Boolean association rules (If Then rule is a Boolean association rule, as it checks if true or false).

The frequent itemset algorithm uses candidate generation process. The groups of candidates are then tested against the dataset. Apriori uses breadth-first search method and a hash tree structure to count candidate itemsets. Also, it is assumed that items within an itemset are kept in lexicographic order. The algorithm identifies the frequent individual items in the database and extends them to larger and larger itemsets as long as those itemsets are found in the database. The frequent itemsets provide the general trends in the database as well.

6.5.4 Evaluationof CandidateRules

Apriori algorithm evaluates candidates for association as follows:

C_k: Set of candidate-itemsets of size *k*

F_l: Set of frequent itemsets of size *k*

f_1 = {large items}

for (k=l; F_k != 0; k-«) *do* {

C_{k+i} = New candidates generated from F_k

for each transaction t in the database *do*

Increment the count of all candidates in C_{k+1} that are contained in t

F_{k+1} = Candidates in C_{k+1} with minimum support

}

Steps of the algorithm can be stated in the following manner:

1. Candidate itemsets are generated using only large itemsets of the previous iteration. The transactions in the database are not considered while generating candidate itemsets.
2. The large itemset of the previous iteration is joined with itself to generate all itemsets having size higher by 1.
3. Each generated itemset that does not have a large subset is discarded. The remaining itemsets are candidate itemsets.

Figure 6.8 shows Apriori algorithm process for adopting the subset of frequent itemsets as a frequent itemset.

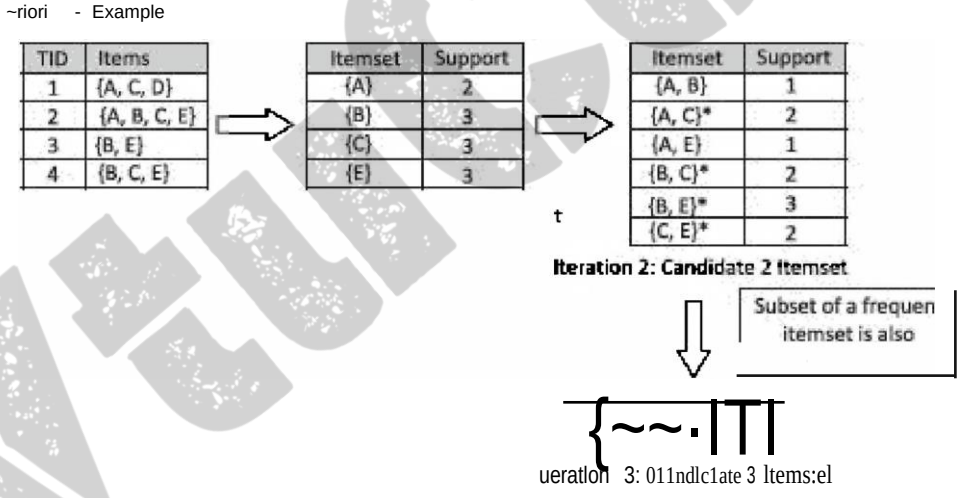


Figure 6.8 Apriori algorithm process for adopting the subset of frequent itemsets as a frequent itemset

It is observed in the Apriori example that every subset of a frequent itemset is also frequent. Thus, a candidate itemset in C_{k+1} can be pruned even if one of its subsets is not contained in F_k.

The Apriori algorithm adopts the fact that the subset of a frequent itemset is

also a frequent itemset. The algorithm thus reduces the number of candidates being considered by only considering the itemsets whose support count is greater than the minimum support count. All infrequent itemsets are pruned if they have an infrequent subset.

Apriori algorithm also possesses certain disadvantages. The algorithm requires multiple scans of a database. The process for generation of a complex candidate exploits more time, space and memory. Therefore, Big Data analytics need alternatives to Apriori algorithm to cut down on the size of candidate pairs. Section 7.4 will describe Park, Chen and Yu (PCY), multistage and multihash algorithms.

6.5.5 Applications of Association Rules

FIM is a popular technique for market basket analysis.

6.5.5.1 Market Basket Model

Market basket analysis is a tool for knowledge discovery about co-occurrence of items. A co-occurrence means two or more things occur together. It can also be defined as a data mining technique to derive the strength of association between pairs of product items. If people tend to buy two products (say A and B) together, then the buyer of product A is a potential customer for an advertisement of product B.

The concept is similar to the real market basket where we select an item (product) and put it in a basket (itemset). The basket symbolizes the transactions. The number of baskets is very high as compared to the items in a basket. A set of items that is present in many baskets is termed as a frequent itemset. Frequency is the proportion of baskets that contain the items of interest.

Market basket analysis can be applied to many areas. The following example explains the market basket model using application examples.

EXAMPLE 6.8

Suggest application examples of the market basket model.

SOLUTION

Application 1:

1. Iterns = Products

   Baskets = Sets of products a customer purchases at one time from a store.

   Example of an application: Given that, many people buy chocolates and flowers together:

   · Run sales on flowers; raise price of chocolates.

   The knowledge is useful when many buy chocolates and flowers together.

Application 2:

2. Items= Words

   Baskets = Web pages

   Unusual words appearing together in a large number of documents, for example, 'research' and 'plastic' may provide interesting information.

Market basket analysis generates If-Then scenario rules. For example, if X occurs then Y is likely to occur too. If item A is purchased, then item B is likely to be purchased too. The rules are derived from the experience. This may be the result of frequencies of co-occurrence of items in past transactions.

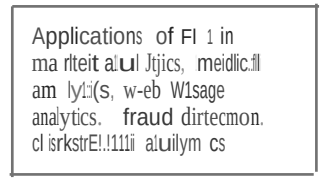
The rules can be used in several analytical strategies. The rules can be written in format If {A} Then {B}. The *If* part of the rule (A) is known as *antecedent* and the *THEN* part of the rule (B) is known as *consequent.* The condition is *antecedent* and the result is *consequent.*

If-then rules about the contents of baskets: {p_1, P_2, ..., P_k} ~ q means, "If a basket contains all of P_1, P_2, ..., P_k then it is *likely* to contain *q,"*

Scale of analysis:

- Amazon sells more than 12 million products and can store hundreds of millions of baskets.
- www has 1000 million words and several billion pages.

- 75 million credit card transactions in a month in India (RBI statistics of June, July 2016) at Point of Sales (POS) terminals.



Market basket analysis signifies shopping carts and supermarket shoppers at once. The analysis is the mining of transaction data to identify relations between different products. This is normally performed to identify products that a customer is likely to buy, given the products that they have already bought (or added to basket). The approach behind Amazon's users who bought a particular product also reviewed or bought other list of items is a well-known example of market basket analysis.

The applications of market basket analysis in various domains other than retail are:

- Medical analytics: Market basket analysis can be used for conditions and symptom analysis. This helps in identifying a profile of illness in a better way. The analysis is also useful in genome analysis, molecular fragment mining, drug design and studying the role of biomarkers in medicine. The analysis can also help to reveal biologically relevant associations between different genes. Further, it can also help to find the effect of environment on gene expressions.
- Web usage analytics: FIM approaches can be used with viewing data on websites. The information contained in association rules can be exploited to learn about website browsing of visitor's behavior, developing website structure by making it more effective for visitors, or improving web marketing promotions. The results of this type of analysis can be used to inform website design (how items are grouped together) and to power recommendation engines (Section 6.8). Results are helpful in targeted marketing. For example, advertising content that people are probably interested in, based on past behavior of users.
- Fraud detection and technical dependence analysis: Extract knowledge so that normal behavior patterns may be obtained in illegal transactions from a credit card database in order to detect and prevent fraud. Another example can be to find frequently occurring relationships or FIM rules

between the various parties involved in the handling of the financial claim. Some examples are:

- Financial institutions to analyze credit card purchases of customers to build profiles for fraud detection purposes and cross-selling opportunities.
- Insurance institution builds the profiles to detect insurance claim fraud. The profiles of claims help to determine if more than one claim belongs to a particular victim within a specified period of time.

- Click stream analysis or web link analysis: Click stream refers to a sequence of web pages viewed by a user. Analysis of clicks is the process of extracting knowledge from web logs. This helps to discover the unknown and potentially interesting patterns useful in the future. It facilitates an understanding of the behavior of website visitors. This knowledge can be used to enhance the way that web pages are interconnected or for increasing the sales of the commercial websites.
- Telecommunication services analysis: Market basket analysis can be used to determine the type of services being utilized and the packages customers are purchasing. This knowledge can be used to plan marketing strategies for customers who are interested in similar services. For example, telecommunication companies can offer TV Internet, and web• services by creating combined offers. The analysis might also be useful to determine capacity requirements.
- Plagiarism detection: It is the process of locating instances of similar content or idea within a work or a document. Plagiarism detection can find similarities among statements that may lead to similar paragraphs if all statements are similar and that possibly lead to similar documents. Formation of relevant word and sentence sequences for detection of plagiarism using association rule mining technique is also very popular technique.

6.5.5.2 Finding Association

Association rules intend to tell how items of a dataset are associated with each

other. The concept of association rules was introduced in 1993 for discovering relations between items in sales data of a large retailing company.

The following examples give rules between items found associated in the sales data of a retailer.

EXAMPLE 6.9

Suggest association rules between items found in the sales data of a retailer, and rules for course choice for a computer science student in college.

SOLUTION

1. {Bread} ~{Butter}

   The rule suggests a relationship between the sales of bread and butter. A customer who buys bread also buys butter.

2. {Chocolates}~ {a Gift Box}

   The rule suggests a that relationship between the sales of chocolates and empty gift boxes exists. A customer who buys chocolates also buys a gift box.

3. {Java programmingl r+ {advanced web technology} and
   {Python programming} ~ {Big Data Analytics}

   The rules suggest relationships between Java and advanced web technology, and Python programming and data analytics. Students who opt for Java programming also want to learn advanced web technology, and those who opt for Python programming also opt for Big Data Analytics.

4. {DataMiningj r+ {DataVisualization}

   The rule may be that 90% of students who select data mining as a major subject will opt for the data visualization course as well.

5. {Computer Graphics, Modeling Techniques} ~ {Animation}

   The rule may be that students who study computer graphics and modeling techniques courses are likely to choose the course on animation in higher semesters.

Association analysis is applicable to several domains. Some of them are marketing, bioinforrnatics, web mining, scientific data analysis, and intrusion detection systems.

The applications might be to find: products that are often purchased together, types of DNA sensitive to a new drug, the possibility of classifying web documents automatically, geophysical trends or patterns in seismicity to predict earthquakes and automate the malicious detecting characteristics.

In medical diagnosis, for example, considering the co-morbid (co-occur) conditions can help in treating the patient in better way. This helps in improving patient care and medicine prescription.

6.5.5.3 ***Finding Similarity***

Section 6.4 describes finding similarity of an item attribute, such as sales percentage increase using Euclidean or cosine similarity coefficients. Section 6.4.2 describes Jaccard similarity of sets. The similarity of sets applies to recommenders and collaborative filtering.

Let A and B be two itemsets. Jaccard similarity index of two itemsets is measured in terms of set theory using the following equation:

$$\text{Jaccard iternsets similarity index} = 1 - \frac{|A \cap B|}{|A \cup B|} \times 100\%.$$

Explanation: $\cap$ means intersection, number of those elements or items which are the same in set A and B. $\cup$ means union, number of elements or items present in union of A and B.

EXAMPLE 6.10

(i) How will you define similarity in purchase of a car model?

(ii) How will you specify frequent threshold for FIM? How will you use association rule to find and count the cities where more than threshold numbers buy a specific car model?

SOLUTION

(i) Assume two sets of car customers, youth Y and family F. Assume in set

Y, 40 out of 100 youths and F 50 out of 200 families opted for the Tata Zest car model. Jaccard similarity index Jzest (Y, F) = 40/ (100 + 200).100% = 13%. Two sets are sharing 13% of the members who purchased a Zest.

(ii) FIM involves finding similarity index in large number of sets after specifying the similarity index threshold which defines an itemset as frequent. Assume N sets of car customers, youth $Y_1, Y_2, \ldots, YN'$ and N¢ sets of families $F_1, F_2, \ldots, FN¢$ in N¢¢ cities. Assume that meaning of frequent is that 10% or more of Yi+ Fi buying Zest among the various car models. Assume all other models sell less than that in the cities. Here i = 1, 2, ..., N¢¢.

Let set X is a set, which has Xi as member if youth buy or if family buy the Zest car model frequently in the ith City. Initialize value, j = 0 for frequent item sets. Then association rule for the FIM in the present case is:

If (Jzest (Y, i, F)>10%) Then (City Xi is a member of X and j =j + 1)}(6.26)

The rule is used for all cities for i = 1, 2, ..., N¢; Here j is the number of cities where frequent item set {Youth, Zest). FIM gives a set of j cities and youth, where youth buy Zest more than 10% of all car buyers.

Text, Web Content, Link, and Social Network Analytics

LEARNING OBJECTIVES

After studying this chapter, you will be able to:

LO 9.1 Use the methods of text mining and machine learning (ML)- Naive-Bayes classifier, and support vector machines for text analytics

LO 9.2 Get knowledge and use the methods of mining the web-links, web-structure and web-contents, and analyzing the web graphs

LO 9.3 Get knowledge and use methods of PageRanking, analysis of web-structure, and discovering hubs, authorities and communities in web-structure

LO 9.4 Get concepts representing social networks as graphs, social network analysis methods, finding the clustering in social network graphs, evaluating the SimRank, counting triangles (cliques) and discovering the communities

RECALL FROM EARLIER CHAPTERS

Graph Data Stores consist of various interconnected data nodes (Section 3.3.5). Models of graph and graph network organization describe the entities and objects, along with their relationships, associations and properties (Sections 8.2 and 8.3). Web and social network graphs are examples of Graph network organization.

Graph structure analytics discovers the degree of interactions, closeness, betweenness, ranks, influences, probabilities distribution, beliefs and potentials. Analysis of the community and network discovers the *close-by* entities and fully mesh like connected sets. Network graph analyzes centralities, and computes the PageRank of the links (Section 8.4 and 8.5).

9.1 ! INTRODUCTION

Text Analytics often termed as 'text mining' refers to analyzing and extracting the meanings, patterns, correlations and structure hidden in unstructured and semi-structured textual data. Text data stores consist of strong temporal dimensions, have modularity over time and sources, such as topics and sentiments.

Methods of machine-learning are prevalent in text analytics also. For example, when a user books an air-flight ticket using a tablet or desktop, the user receives an SMS on the mobile about details of the booking and flight timings. An ML algorithm, such as *Windows Crotona* at the mobile reads and learns by itself from the SMSs received at the phone. *Crotona* uses the ML for the SMS text analysis .. Learning results in SMS alerts to the user. An alert is reminder a day before the flight. Another alert is two hours before the flight, about the need to reach the airport. Those alerts are system-generated without prior request from the user.

The reader is required to know the meaning of the following select key terms:

Vector refers to an entity with number of interrelated elements. For example, a data point consists of n-elements in an n-dimensional space, and represents a vector to that point from the origin in the space. A word is a vector of characters as the elements. Consider a vector representation of word 'McGraw-Hill', then vector vMH = [M, c, G, r, a, w, -, H, i, l, l]. vMH is vector of 11 elements (characters) that refers to word McGraw-Hill.

Feature refers to a set of properties associated with an entity, object or category. For example, feature of properties, such as description of data analysis, data cleaning, data visualization and other topics in a book on data analytics.

Category refers to a classification on the basis of set of distinct features (for example, a category of text, document, cars, toys,

students, news or fruits).

Label refers to a name assigned to a category, for example to sports-news, latest data analytics books.

Dimension refers to a number of associated values, features or states, along the distinct spaces (dimensions). For example, a sentence has a number of words, each word has a number of characters, each word may have a feature, and so the sentences are in a three-dimensional space. Two dimensions are in metric spaces, which mean values in quantifiable spaces, such as the number of words, probability of occurrences in sentences, etc. Third dimension is in feature space, measured by a feature such as noun, verb, adverb, preposition, punctuation marks and stop word.

Another example is *apples.* Metric space variables are values, such as variables n, *number of apples* of specific properties, and P the *probability of preferring* apples of specific properties. Assume, feature space variables are four properties, *colour, shape, type* and *freshness.* Metric parameters and properties of apples are said to be in six-dimensional space, two are metric space (n and P) and four are feature space.

Graph data model refers to the data modelled by a set of entities. The entities identify by vertices V. A set of relations or associations identifies by edges E. An edge *e* represents a relation or association between two entities. Nodes represent the entities in the graph. The model also represents a hierarchy between the parent and children nodes.

Graph data network organization refers to a structure created by organizing entities or objects in a network, such as social network, business network and student network. A network organization means where persons or entities interconnect with each other, and have areas of common interest, business or study. A graph enables ease in traversing from one entity, person or web page link to another in the network by following a path. Web graph and social network graph enable such analysis. A graph network organization models the web and social networks. Examples of social networks are SlideShare, Linkedln, Facebook and Twitter. The analytics of social networks finds the link ranks, clusters and correlations. The analytics discovers hubs and communities.

Web content mining refers to the discovery of useful information from web documents and services. Search engines use web content mining. A search provides the links of the required information to the user.

Hyperlinks refer to links mentioned in the contents that enable the retrieval of contents at web, file, object or resources repository.

Link analytics means web structure mining of hyperlinks between web documents. The analytics of links and analyzing them for metrics such as page ranks, clusters, correlations, hubs and communities.

Count triangles Algorithm is an algorithm that finds a number of triangular relationships among the nodes. Triangular relationships mean interrelations between each other.

Graph node centrality metric means the centrality of a node in reference to other nodes using certain metrics. Metrics used for centrality of a node are degree, closeness, betweenness or other characteristics of the node, such as rank, belief, potential, expectation, evidence, reputation or status.

Degree centrality of a node refers to the number of direct connections. Having more number of direct connections is not always a better metric. Better measure is the fact that the connection directs to significant results and tell how the nodes connect to the isolated node.

Betweenness centrality is a measure that provides the extent to which a node lies on paths between other nodes. A node with high betweenness signifies high influence over what flows in the network indicating importance of link and single point of failure.

Closeness centrality is the degree to which a node is near all other nodes in a network (directly or indirectly). It reflects the ability to access information through the network.

The present Chapter focuses on text, web, contents, structure and social network graph analytics. Section 9.2 describes text mining and usage of ML techniques=Naive-Bayes analysis and support vector machines (SVMs) for analyzing text. Section 9.3 describes web mining, methods to implement the system, and analyzing the web graphs.

Section 9.4 describes PageRank methods, web structure analytics and finding the hubs and communities. Section 9.5 describes social network analysis, representation of social networks as graphs and computational methods of finding the clustering in social network graphs, evaluating the SimRank, counting triangles (cliques) and discovering the communities.

This chapter follows a method of notations as mentioned earlier in Section 6.1 for fonts when absolute value, mean value, function value, vector element or set member, entity or variable when these denote by a character or character-set. This chapter follows the notations for the probabilities, earlier specified in
Section 8.3.2. Condition probability P specifies as P (xilck) which means probability of variable x = xi at condition c = ck.

9.2 ! TEXT MINING

LO 9.1

i.thods of text mining and machinele;uning(L)• Naive-Bayesdassifier,and support vectormachines for text analytics

Today, large amounts of textual data is generated in computing applications. Text stream arriving continuously over time generates text data. For example, news articles, news reports, online comments on news, online traffic reports, corporate reports, web searches, and contents at social media discussion• forums (such as Linkedin, Twitter and Facebook), short messages on phones, chat messages, transcripts of phone conversations, biogs and e-mails.

The abundance of textual data leads to problems which relate to their collection, exploration and ways of leveraging data. Textual data presents challenges for computing and storage requirements, consists of a strong temporal dimension, has modularity over time and have sources such as topics and sentiments. Examples of text processing techniques are clustering analysis, classifications, evolution analysis and event detection. Following subsections describe text mining in details:

9.2.1 Text Mining

Four definitions are:

1. "Text mining refers to the process of deriving high-quality information from text." (Wikipedia)
2. "Text mining is the process of discovering and extracting knowledge from unstructured data." (National Center of Text Mining -The University of Manchester+)
3. "Text mining is the process of analyzing collections of textual contents in order to capture key concepts themes, uncover hidden relationships, and discover the trends without requiring that you know the precise words or terms that authors have used to express those concepts." (IBM2)
4. "Text mining is a technique which helps in revealing the patterns and relationships in large volumes of textual content that are not visible to the naked eye, leading to new business opportunities and improvements in processes." (Amazon BigData Official Blog3)

Applications of text mining in business domains are predicting stock movements from analysis of company results, decision making for product and innovations developed at the company and contextual advertising. Some other applications are (i) mail filtering (spam), (ii) drug action reports (iii) fraud detection (iv) knowledge management, and (iv) social media data analysis.

The applications provide innovative and insightful results. The results when combined with other data sources, find the answers to the following:

(i) Two terms which occur together

(ii) Information linkage with another information

(iii) Different categories that can be created from extracted information

(iv) Prediction of information or categories.

9.2.1.1 Text Mining Overview

Text mining includes extraction of high-quality information, discovering and extracting knowledge, and revealing patterns and relationships from unstructured data available in the form of text.

The term *text analytics* evolves from provisioning of strong integration with the already existing database technology, artificial intelligence, machine learning, data mining and text Data Store techniques. Information retrieval, natural language processing (NLP), classification, clustering and knowledge management are some of such useful techniques. Figure 9.1 shows process-pipeline in text• analytics.

9.2.1.2 Areas and Applications of Text Mining

Natural Language Processing (NLP) is a technique for analyzing, understanding and deriving meaning from human language. NLP involves the computer's understanding and manipulation of human language. NLP algorithms are typically based on ML algorithms. They automatically learn the rules. First, they analyze set of examples from a large collection of sentences in a book. Then, they make the statistical inferences.

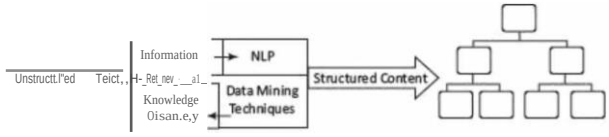


Figure 9.1 Text analytics process pipeline

NLP contributes to the field of human computer interaction by enabling several real-world applications such as automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction and stemming. The common uses of NLP include text mining, machine translation and automated question answering.

Information Retrieval (IR) is a process of searching and retrieving a subset of documents from the abundant collection of documents. IR can also be defined as extraction of information required by a user. IR is an area derived fundamentally from database technology. One of the most popular applications of IR is searching the information on the web. Search engines provide IR using various advance techniques. For example, the crawler program is capable of retrieving information from a wide variety of data sources. Search methods use metadata or full-text indexing.

Information Extraction (IE) is a process in which the software extracts structured information from unstructured and/or semi• structured documents. IE finds the relationship within text or desired contents from text. IE ideally derives from machine learning, more specifically from the NLP domain. Content extraction from the images, audio or video is an example of information extraction.

IE requires a dictionary of extraction patterns (For example, "Citizen of <x>, or "Located in -oc-") and a semantic lexicon (dictionary of words with semantic category labels).

Document Clustering is an application which groups text documents into clusters. Automating document organization, topic extraction and fast information retrieval or filtering use the document clustering method. For example, web document clustering facilitates easy search by users.

Document Classification is an application to classify text documents into classes or categories. The application is useful for publishers, news sites, biogs or areas where lot of contents are present.

Web Mining is an application of data mining techniques. They discover patterns from the web Data Store. The patterns facilitate understanding. They improve the services of web-based applications. Data mining of web usage provides the browsing behavior of a website.

Concept Extraction is an application that deals with the extraction of concept from textual data. Concept extraction is an area of text classification in which words and phrases are classified into a semantically similar group.

9.2.1.3 Text Mining Process

Text is most commonly used for information exchange. Unlike data stored in databases, text is unstructured, ambiguous and difficult to process. Text mining is the process that analyzes a text to extract information useful for a specific purpose.

Syntactically, a text document comprises characters that form words, which can be further combined to generate phrases or sentences. Text mining steps are (i) recognizing, extracting and using the information present in words. Along with searching of words, mining involves search for semantic patterns as well.

Text mining process consists of a process-pipeline. The pipeline processes execute in several phases. Mining uses the iterative and interactive processes. The processing in pipeline does text mining efficiently and mines the new information. Figure 9.2 shows five phases of the process pipeline.

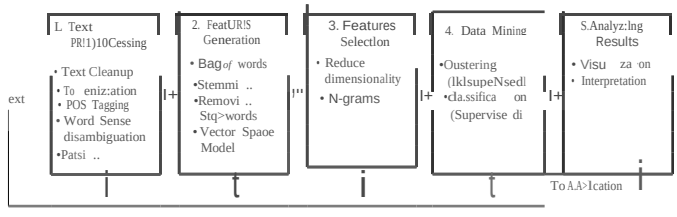


Figure 9.2 Five phases in a process pipeline

The following subsection describes these phases:

9.2.1.4 Text Mining Process Phases

The five phases for processing text are as follows:

Phase 1: Text pre-processing enables Syntactic/Semantic text-analysis and does the followings:

1. Text *cleanup* is a process of removing unnecessary or unwanted information. Text cleanup converts the raw data by filling up the missing values, identifies and removes outliers, and resolves the inconsistencies. For example, removing comments, removing or escaping "%20" from URL for the web pages or cleanup the typing error, such as teh (the), don't (do not) [%20 specifies space in a URL].
2. *Tokenization* is a process of splitting the cleanup text into tokens (words) using white spaces and punctuation marks as delimiters.
3. *Part of Speech (POS) tagging* is a method that attempts labeling of each token (word) with an appropriate POS. Tagging helps in recognizing names of people, places, organizations and titles. English language set includes the noun, verb, adverb, adjective, prepositions and conjunctions. Part of Speech encoded in the annotation system of the Penn Treebank Project has 36 POS tags. [4]
4. *Word sense disambiguation* is a method, which identifies the sense of a word used in a sentence; that gives meaning in case the word has multiple meanings. The methods, which resolve the ambiguity of words can be context or proximity based. Some examples of such words are bear, bank, cell and bass.
5. *Parsing* is a method, which generates a parse-tree for each sentence. Parsing attempts and infers the precise grammatical relationships between different words in a given sentence.

Phase 2: Features Generation is a process which first defines features (variables, predictors). Some of the ways of feature generations are:

1. *Bag of* words-Order of words is not that important for certain applications.

   Text document is represented by the words it contains (and their occurrences). Document classification methods commonly use the bag-of-words model. The pre-processing of a document first provides a document with a bag of words. Document classification methods then use the occurrence (frequency) of each word as a feature for training a classifier. Algorithms do not directly apply on the bag of words, but use the frequencies.
2. Stemming-identifies a word by its root.

   (i) Normalizes or unifies variations of the same concept, such as *speak* for three variations, i.e., speaking, speaks, speakers denoted by [speaking, speaks, speaker —+ speak]

   (ii) Removes plurals, normalizes verb tenses and remove affixes.

   Stemming reduces the word to its most basic element. For example, impuriflcation r+ pure.
3. *Removing stop words* from the feature space-they are the common words, unlikely to help text mining. The search program tries to ignore stop words. For example, ignores *a, at, for,* it, *in* and *are.*
4. *Vector Space Model* (VSM)-is an algebraic model for representing text documents as vector of identifiers, word frequencies or terms in the document index. VSM uses the method of term frequency-inverse document frequency (TF-IDF) and evaluates how important is a word in a document.

   When used in document classification, VSM also refers to the bag-of-words model. This bag of words is required to be converted into a term-vector in VSM. The term vector provides the numeric values corresponding to each term appearing in a document. The term vector is very helpful in feature generation and selection.

Term frequency and inverse document frequency (IDF) are important metrics in text analysis. TF-IDF weighting is most common• Instead of the simple TF, IDF is used to weight the importance of word in the document.

TF-IDF stands for the 'term frequency-inverse document frequency'. It is a numeric measure used to score the importance of a word in a document based on how often the word appears in that document and in a given collection of documents. It suggests that if a word appears frequently in a document, then it is an important word, and should therefore be high in score. But if a word appears in many more other documents, it is probably not a unique identifier, and therefore should be assigned a lower score. The TF-IDF is measured as:

$$\text{TF-IDF<t>} \quad \frac{\text{No. of times t appears in a document}}{\text{Total No. of terms in the document}} \times \log \frac{\text{No. of documents in the collection}}{\text{No. of documents that contain t}} \qquad (9.1)$$

where t denotes the term vector.

Following example suggests method of calculating TF-IDF (t):

EXAMPLE 9.1

Consider a document containing 1000 words wherein the word *toys* appears 16 times. How will the TF-IDF weight be calculated?

SOLUTION

The term frequency (TF) for *toys* is then (16/1000) = 0.016. Let, there are 10 million documents and the word *toys* appear in 1000 of them. Then, the inverse document frequency (IDF) is calculated as $\log_{10}$ (10,000,000/1,000) = 4.

TF-IDF weight= 0.016 x 4 = 0.064

Additional weight is assigned to terms appearing as keywords or in titles. Documents are usually represented as a sparse vector of terms weights and extra weights are added to the terms appearing in title or keywords.

Pre-processing of web data succeeds the conversion of bag of words into vector space model (VSM) or simply by vector creation.

Common Information Retrieval Technique - Vector space model (VSM) is an algebraic model for representing textual information as vectors of identifiers, such as, index terms. Information retrieval methods use VSM.

Each document or HTML page represents by a sparse vector of term weights. The sparse matrices represent the term frequencies (TFs).

(Sparse vector and sparse-matrix have many elements as zero or null. An associated metadata enables data storing of them in a form which does not include zeros in case of large datasets. The metadata then includes indices map with the positions in the list of elements of the vector or matrix.)

The following example gives the conversion method for evaluating TFs and matrix in pre-processing phase.

EXAMPLE 9.2

Assume that the documents below define the document space with five documents di, dz, d3, d4 and ds:

Train Document Set:

dt: Children like the toys.

dz: The toys are precious.

Test Document Set:

d3: There are many toys in the shop.

d4: Some toys are precious and some toys are costly as well.

ds: The toys shop is one of the famous shops.

How will be the documents term vector and matrix be calculated for features generation/selection?

SOLUTION

First, create an index vocabulary of the words of the train document set using the documents dt and dz from the document set. The index vocabulary E(t) where tis the term will be:

$$Eu) = \begin{cases} \text{I. when I= "children"} \\ \text{2. when I = "toys"} \\ \text{3. when I= "precious"} \\ \text{4. when t = "shop"} \\ \text{5. when t = "costly"} \\ \text{6. when t = "famous"} \end{cases}$$

Note that the stop words are already not considered during the pre-processing step. The term frequency (TF) is a measure of how many times the terms present in vocabulary E(t) are present in the documents ds, d4 and ds.

$$TF(l.d) = \sum_{1EJ} \text{I,count } (x.t) \qquad (9.2)$$

where the count (x, t), is a simple function defined as:

$$\text{coumt.r.r} = \begin{cases} 1. & \text{if } x = t \\ 0. & \text{otherwise} \end{cases} \quad (9.3)$$

For example, TF ("toys", ds) = 2.

Create the document vector as:

$$v_{1,f} = <TFul.dn).TFU2.dn) \quad TF(tn.dnJ) \quad (9.4)$$

Thus, the documents d3, d4 and ds are represented as vector as:

$$v_{43} = <TFUl.d3J.TF(r2.d3) \quad TFUn.d3))$$

$$v_{1,14} = (TF<tl.d4).TF(t2.d4J \quad TF(tn.d4))$$

$$v_{45} = (TFU\,1.d5).\ TF<t2.d5)\bullet\ldots\ldots\ TFUn.\ d5>1 \quad (9.5)$$

This gives:

$$I'_{43} = (I.I.0.1.0.0)$$

$$v_{44} = (0,2.L0.1.01$$

$$I'_{45} = (0.1.0.2.0.1) \quad (9.6)$$

The resulting vector $\dot{I}\cdot_{,3}$ shows 1 occurrence of the term "children", 1 occurrence of the term "toys" and so on. In the $i\text{-}14,$ there is 0 occurrence of the term "children", 2 occurrences of the term "toys" and so on.

A collection of web documents requires representation as vectors. Another representation is a matrix with IDI x F shape, where IDI is the cardinality of the document space (total number of documents) and the Fis the number of features. F represents the vocabulary size in the example. Matrix representation of the vectors described above is by a 3 x 6 matrix as follows:

$$MIDIXF = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 & 0 & 1 \end{bmatrix} \quad (9.7)$$

Example 9.2 shows that the matrices representing term frequencies tend to be very sparse (with majority of terms zeroed). A common representation of such matrix is thus the sparse matrices.

Phase 3: Features Selection is the process that selects a subset of features by rejecting irrelevant and/or redundant features (variables, predictors or dimension) according to defined criteria. Feature selection process does the following:

1. *Dimensionality* reduction-Feature selection is one of the methods of division and therefore, dimension reduction. The basic objective is to eliminate irrelevant and redundant data. Redundant features are those, which provide no extra information. Irrelevant features provide no useful or relevant information in any context.

   Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) are dimension reduction methods. Discrimination ability of a feature measures relevancy of features. Correlation helps in finding the redundancy of the feature. Two features are redundant to each other if their values correlate with each other.

2. *N-gram* evaluation-finding the number of consecutive words of interest and extract them. For example, 2-gram is a two words sequence, ["tasty food", "Good one"]. 3-gram is a three words sequence, ["Crime Investigation Department"].

3. Noise *detection and evaluation of* outliers methods do the identification of unusual or suspicious items, events or observations from the data set. This step helps in cleaning the data.

The feature selection algorithm reduces dimensionality that not only improves the performance of learning algorithm but also reduces the storage requirement for a dataset. The process enhances data understanding and its visualization.

Phase 4: Data mining techniques enable insights about the structured database that resulted from the previous phases. Examples of techniques are:

1. Unsupervised learning (for example, clustering)

   (i) The class labels (categories) of training data are unknown

   (ii) Establish the existence of groups or clusters in the data

   Good clustering methods use high intra-cluster similarity and low inter-cluster similarity. Examples of uses - biogs, patterns

and trends.

2. *Supervised learning (for example, classification)*

   (i) The training data is labeled indicating the class

   (ii) New data is classified based on the training set

   Classification is correct when the known label of test sample is identical with the resulting class computed from the classification model.

   Examples of uses are *news filtering application,* where it is required to automatically assign incoming documents to pre-defined categories; *email spam filtering,* where it is identified whether incoming email messages are spam or not.

   Example of text classification methods are *Naive Bayes Classifier* and *SVMs.*

3. *Identifying evolutionary patterns* in temporal text streams-the method is useful in a wide range of applications, such as summarizing of events in news articles and extracting the research trends in the scientific literature.

Phase 5: Analysing results

(i) Evaluate the outcome of the complete process.

(ii) Interpretation of Result- If acceptable then results obtained can be used as an input for next set of sequences. Else, the result can be discarded, and try to understand what and why the process failed.

(iii) Visualization - Prepare visuals from data, and build a prototype.

(iv) Use the results for further improvement in activities at the enterprise, industry or institution.

Open source tools, such as *nltk* are available for text analytics. Online contents accompanying book describe how text analytics tasks can be performed using Python library *nltk* in the solution of Practice Exercise 9.2.

9.2.1.5 Text Mining Challenges

The challenges in the area of text mining can be classified on the basis of documents area-characteristics. Some of the classifications are as follows:

1. NLP issues:
   (i) POS Tagging
   (ii) Ambiguity
   (iii) Tokenization
   (iv) Parsing
   (v) Stemming
   (vi) Synonymy and polysemy
2. Mining techniques:
   (i) Identification of the suitable algorithm(s)
   (ii) Massive amount of data and annotated corpora
   (iii) Concepts and semantic relations extraction
   (iv) When no training data is available
3. Variety of data:
   (i) Different data sources require different approaches and different areas of expertise
   (ii) Unstructured and language independency
4. Information visualization
5. Efficiency when processing real-time text stream
6. Scalability

9.2.1.6 Supervised Text Classification

The categorization of text documents requires information retrieval, ML and NLP techniques. Some important approaches to

automatic text categorization are based on ML techniques.

The *supervised text classification* requires labeled documents and additional knowledge from experts. The algorithms exploit the training data (where zero or more categories) to learn a classifier, which classifies new text documents and labels each document. A document is considered as a positive example for all categories with which it is labeled, and as a negative example to all others. The task of a training algorithm for a text classifier is to find a weight vector which best classifies new text documents.

The different approaches for supervised text classification are:

(i) K-Nearest Neighbour Method

(ii) Support Vector Machine

(iii) Naive Bayes Method

(iv) Decision Tree

(v) Decision Rule

K-Nearest Neighbours (KNN) method makes use of training text document. The training documents are the previously categorized set of documents. They train the system to understand each category. The classifier uses the training 'model' to classify new incoming documents. KNN assumes that close-by objects are more probable in the same category. KNN finds k objects in the large number of text documents, which have most similar query responses. Thus, in KNN, predictions are based on a method that is used to predict new (not observed earlier) text data. The predictions are by (i) majority vote method (for classification tasks) and (ii) averaging (for regression) method over a set of K-nearest examples.

The decision trees or decision rules are built to predict the category for an input document. A decision tree or rule represents a set of nested logical if-then conditions on the observed values of the text features that enable the prediction of the target variable. The decision tree and decision rules are also used to classify (categorize) the document. Classification is done by recursively splitting the text features into a set of non-overlapping regions (Refer Section 6.8). (Section 6.7.4)

The following subsections describe Naive Bayes Method and Support Vector Machines in detail.

9.2.2 Naive Bayes Analysis

Naive Bayes classifier is a simple, probabilistic and statistical classifier. It is one of the most basic text classification techniques, also known as multivariate Bernoulli method. Naive Bayes classifies using Bayes theorem along with the Naive independence assumptions (conditional independence). The classifier computes condition probabilities for the conditional independence (Refer Section 8.3.2).

Probability that a bag-of-words ~ belong to kth class equals the product of individual probabilities of those words. P (~lck) = II P (xilck), where xi is a discrete random variable (word), i = 1, 2, ..., n, where n is number of words in the bag. II is sign for the product of n terms. [P~lck) means probability of condition that state the value= xi and of c = ck (Example 8.6)].

The P <xlck) is normalized as all distributed probabilities equals 1. P <xlck) is normalized by dividing the product on right hand side by *Li* P (~lck) P (ck).

The following example gives the method of deciding the most likely class.

EXAMPLE 9.3

How is "maximum a posteriori (MAP)" used to obtain the most likely class and take a decision?

SOLUTION

Text classification problem uses the words (or tokens) of the document in order to classify it on the appropriate class. Bayes' rule is applied to documents and classes. For a document (d) and class (c), we get:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (9.8)$$

The "maximum a posteriori (MAP)" (to obtain most likely class) decision rule is applied to documents and classes:

$$c_{MAP} = \arg\max_{c} \hat{P}(c|d) = \arg\max_{c} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c) \quad (9.9)$$

(MAP is "maximum posteriori" = most likely class)

$$c_{MAP} = \arg\max_{c \in C} \left(\frac{P(d|c)P(c)}{P(d)} \right) \quad \text{(Bayes Rule)} \tag{9.10}$$

LO 3.1

(…opping the denominator) (9.11)

(9.12)

where $t_1, t_2, \ldots, t_n$ are tokens of document.

Multinomial Naive Bayes independence assumptions

$$P(t_1, t_2, \ldots, t_n | c) \tag{9.13}$$

Bag-of-words assumption: Assume the position of the word does not matter.

Conditional independence: Assume the feature probabilities $P(t_i|h)$, are independent given the class(c):

$$P(t_1, t_2, \ldots, t_n | c) = P(t_1|c) \cdot P(t_2|c) \cdot P(t_3|c) \cdot \ldots \cdot P(t_n|c) \tag{9.14}$$

and thus, conditional independences are given by

$$c_{MAP} = \arg\max_{c \in C} P(t_1, t_2, \ldots, t_n | c) P(c) \tag{9.15a}$$

$$c_{NB} = \arg\max_{c \in C} \left(P(c_j) \prod_{t \in T} P(t|c) \right) \tag{9.15b}$$

Applying multinomial Naive Bayes classifier to text classification where positions = all word positions in the text document,

$$c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{i \in positions} P(t_i|c_j) \tag{9.16}$$

The equation estimates the product of the probability of each word of the document given a particular class (likelihood), multiplied by the probability of the particular class (prior) to find in which class one should classify a new document.

Select the one with the highest probability among all the classes of set C. Calculation of product of the probabilities leads to float point underflow when handling numbers with specific decimal point accuracy by computing devices. Such small numbers will be rounded to zero, implying the analysis is of no use at all. In order to avoid this, instead of maximizing the product of the probabilities, the maximization of the sum of their logarithms is done:

$$c_{NB} = \arg\max_{c_j \in C} \left(\log P(c_j) \prod_{i \in Positions} \log P(t_i|c_j) \right) \tag{9.17}$$

Here, choose the one with the highest log score rather than choosing the class with the highest probability. Given that the logarithm function is monotonic, the decision of MAP remains the same.

When compared with other techniques, such as Random Forest, Max Entropy and SVM, the Naive Bayes classifier performs efficiently in terms of less CPU and memory consumption. Naive Bayes classifier requires a small amount of training data to estimate the parameters. The classifier is not sensitive to irrelevant features as well. Furthermore, the training time is significantly smaller with Naive Bayes as opposed to other techniques.

The classifier is popularly used in a variety of applications, such as email spam detection, personal email sorting, document categorization, language detection, authorship identification, age/gender identification and sentiment detection.

9.2.3 Support Vector Machines

Support vector machines (SVM) is a set of related *supervised learning methods* (the presence of training data) that analyze data, recognize patterns, classify text, recognize hand-written characters, classify images, as well as bioinformatics and bio sequence analysis.

A vector has in general n components, x2, x3, ..., xn. A datapoint represents by (Xl, X2, ..., Xn) in n-dimensional space. Assume for the sake of simplicity, that a vector has two components, Xl and X2 (Two sets of words in text analysis).

Section 6.7.6 described the use of the concept of hyperplanes for classification. A *hyperplane* is a subspace of one dimension less than its ambient space in geometry (Figure 6.18). If a space is 3-dimensional then its hyperplanes are 2-dimensional planes, while if the space is 2-dimensional, its hyperplanes are I-dimensional, which means lines.

The hyperplane which separates the two classes most appropriately has maximum distance from closest data points of the distinct classes. This distance is termed as margin. Figure 9.3 shows the concept of support vectors, separating hyperplane and margins when using Bas a classifier. The margin for hyperplane Bin Figure 9.3 is more as compared to two hyperplanes, A and C shown by dotted lines. The margin of the data points from B is maximum. Therefore, the hyperplane B is the maximum margin classifier.

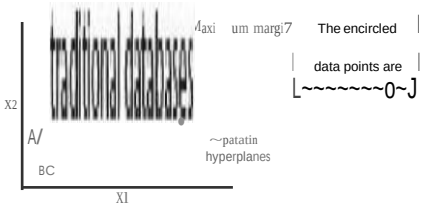


Figure 9. 3 Support vectors, separating hyperplane (B) and margins

A and Care closest (least margins) to the data points. These are called the *support vectors.* They support the classifications of the star and dotted data points. [Remember that with n-dimensional datapoints space, a hyperplane has the vectors along (n - 1) axes.]

The support vectors are such that a set of data points lies closest to the decision (classification) surface (or hyperplane). Those points are most difficult to classify. They have direct bearing on the optimum location of the classification surface. Support vectors along maximum margin classification surface are thus gives the best results.

Thus, a SVM classifier is a *discriminative classifier* formally defined by a separating hyperplane. The concept applies extensively in number of application areas of ML. Applications of SVMs are as follows:

1. classification based on the outputs taking discrete values in a set of possible categories, SVM can be used to separate or predict if something belongs to a particular class or category. SVM helps in finding a decision boundary between two categories.
2. Regression analysis, if learning problem has continuous real-valued output (continuous values of x, in place discrete n values, (X1, X2, X3, ... , x.)
3. Pattern recognition
4. Outliers detection.

The following example illustrates the discriminative classifier method, formally defined by a separating a hyperplane for taking the decision for effective elements (entities, set of words, itemsets) in a training set.

EXAMPLE 9.4

How is the discriminative classifier used?

SOLUTION

Consider a mapping function (f) used for linear separation in the feature space (H). An optimal separating hyperplane depends on the data through dot products in H (f(xJf(l|_)),

A kernel function k is required that acts as a measure of similarity. Let us use k where

LO 4.3 (9.1)

The objective is to select the hyperplane which separates the two classes most appropriately. This helps in identifying the right or appropriate hyperplane. Figure 9.3 showed three hyperplanes. A, B and C. A and C are least margin planes and B is maximum margin plane.

A hyperplane,maximum margin classifier is the right hyperplane. It has maximum distances from the nearest data points (of either classes). An important reason for selecting the maximum margin classification surface is robustness. A hyperplane having low margin has considerably high chance of misclassifying.

Binary Classification

For a given training data [xi, y(xi)] for i = 1 ...N, with xi E !Rd and Yi E {-1,1},learn a classifier f(x) such that:

$$/(X;) \begin{cases} >O & \text{for } y';= +I \\ -<0 & \text{for } Y;=-1 \end{cases} \quad (9.19)$$

The above equation implies that $yJ\{.x) > 0$ for correct classification.

Figure 9.4 shows a two-class classification. The method is using one hyperplane B for separating two-class classification of data points.

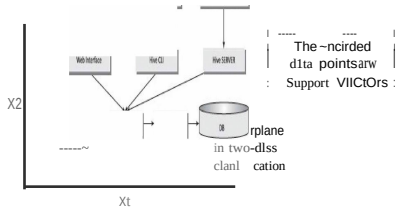


Figure 9. 4 Concept of training set data using support vectors

Let us take the simplest case of two-class classification. Suppose there are two features X1 and X2 and it is required to classify objects as shown in the Figure 9.4. Stars and dots represent the objects (itemsets, sets of words, entities) of two classes. The goal is to design a hyperplane (B) that classify all training vectors in two classes for linearly separable binary set.

The following example gives the method to design a hyperplane (B) that classify all training vectors:

EXAMPLE 9.5

How will you select an appropriate hyperplane that classifies all training vectors?

SOLUTION

Plane Bis f(x) = wx + b where w is weight vector. Figure 9.5 shows the method of selecting the right hyperplane.

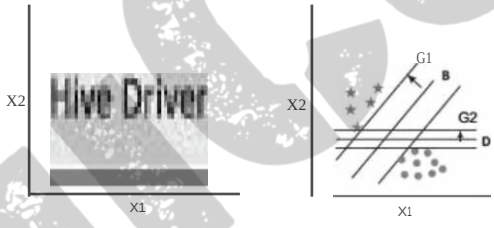


Figure 9.5 Method of selecting the right hyperplane

The best choice is the hyperplane that leaves the maximum margin from both the classes. Thus, Bis the right choice, since Gl > G2.

Thus, for line segment B,

$$ft.x) \sim 1 \text{ "}// e\ x -+ \text{ class } 1$$
$$ft.x) \sim -1 \text{ "}// e\ \text{ll} -+ \text{ class } 2 \quad (9.20)$$

Self-Assessment Exercise linked to LO 9.1

1. Define text analytics.
2. List the steps in text pre-processing phase. Why are tokenization and POS tagging needed? Give an example of each step.
3. How is bag-of-word used in text analysis? Give 5 examples of stemming the affix wordforms to its root word.
4. How do the TF-IDF weighting and sparse matrices represent the term frequencies (TFs) for use in text analysis?
5. How does maximum a posteriori (MAP) in Naive Bayes classifier enable the decision about classification?
6. How does support vectors in SVMs classify the data points in n-dimensional space.

LO 9.2

1inIng the veb<Ilnks, veb-structure and web• contents. and analyzing the vebgr.iphs

Web is a collection of interrelated files at web servers. Web data refers to (i) web content–text, image and records, (ii) web structure–hyperlinks and tags, and (iii) web usage–http logs and application server logs.

Features of web data are:

1. Volume of information and its ready availability
2. Heterogeneity
3. Variety and diversity (Information on almost every topic is available using different forms, such as text, structured tables and lists, images, audio and video.)
4. Mostly semi-structured due to the nested structure of HTMLcode
5. Hyperlinks among pages within a website, and across different websites
6. Redundant or similar information may be present in several pages
7. Mostly, the web page has multiple sections (divisions), such as main contents of the page, advertisements, navigation panels, common menu for all the pages of a website and copyright notices
8. A web form or HTMLform on a web page enables a user to enter data that is sent to a server for processing
9. Website contents are dynamic in nature where information on the web pages constantly changes, and fast information growth takes place such as conversations between users, social media, etc.

The following subsections describe web data mining and analysis methods:

9.3.1 Web Mining

Data Mining is a process of discovering patterns in large datasets to gain knowledge. The process can be shown as [Raw Data - Patterns - Knowledge]. Web data mining is the mining of web data. Web mining methods are in multidisciplinary domains: (i) data mining, ML, natural language, (ii) processing, statistics, databases, information retrieval, and (iii) multimedia and visualization.

Web consists of rich features and patterns. A challenging task is retrieving interesting content and discovering knowledge from web data. Web offers several opportunities and challenges to data mining.

Definition of Web Mining

Web mining refers to the use of techniques and algorithms that extract knowledge from the web data available in the form of web documents and services. Weh mining applications are as follows:

(i) Extracting the fragment from a web document that represents the full web document

(ii) Identifying interesting graph patterns or pre-processing the whole web graph to come up with metrics, such as PageRank

(iii) User identification, session creation, malicious activity detection and filtering, and extracting usage path patterns

Web Mining Taxonomy

Web mining can broadly be classified into three categories, based on the types of web data to be mined. Three ways are web content mining, web structure mining and web usage mining. Figure 9.6 shows the taxonomy of web mining.

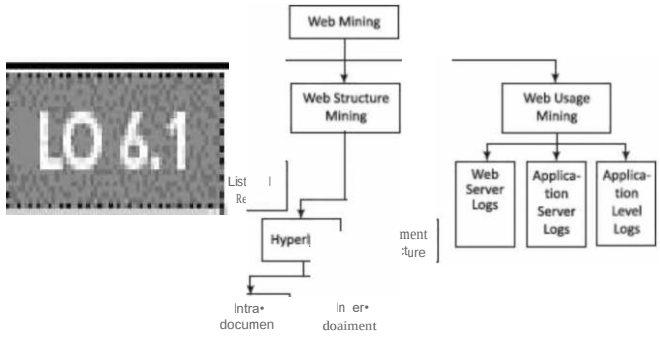


Figure 9.6 Web mining taxonomy

Web content mining is the process of extracting useful information from the contents of web documents. The content may consist of text, images, audio, video or structured records, such as lists and tables.

Web structure mining is the process of discovering structure information from the web. Based on the kind of structure-information present in the web resources, web structure mining can be divided into:

1. Hyperlinks: the structure that connects a location at a web page to a different location, either within the same web page (intra• document hyperlink) or on a different web page (inter-document hyperlink)
2. Document Structure: The structure of a typical web graph consists of web pages as nodes, and hyper links as edges connecting the related pages.

Web usage mining is the application of data mining techniques which discover interesting usage patterns from web usage data. The data contains the identity or origin of web users along with their browsing behavior at a web site. Web usage mining can be classified as:

(i) Web Server logs: Collected by the web server and typically include IP address, page reference and access time.

(ii) Application Server Logs: Application servers typically maintain their own logging and these logs can be helpful in troubleshooting problems with services.

(iii) Application Level Logs: Recording events usually by application software in a certain scope in order to provide an audit trail that can be used to understand the activity of the system and to diagnose problems.

9.3.2 Web Content Mining

Web Content Mining is the process of information or resource discovery from the content of web documents across the World Wide Web. Web content mining can be (i) direct mining of the contents of documents or (ii) mining through search engines. They search fast compared to direct method.

Web content mining relates to both, data mining as well as text mining. Following are the reasons:

(i) The content from web is similar to the contents obtained from database, file system or through any other mean. Thus, available data mining techniques can be applied to the web.

(ii) Content mining relates to text mining because much of the web content comprises texts.

(iii) Web data are mainly semi-structured and/or unstructured, while data mining is structured and the text is unstructured.

Applications

Following are the applications of content mining from web documents:

1. Classifying the web documents into categories
2. Identifying topics of web documents
3. Finding similar web pages across the different web servers
4. Applications related to relevance:

(a) Recommendations - List of top "n" relevant documents in a collection or portion of a collection

(b) Filters - Show/Hide documents based on some criterion

(c) Queries - Enhance standard query relevance with user, role, and/or task-based relevance.

9.3.2.1 Common Web Content Mining Techniques

Pre-processing of contents The pre-processing steps are quite similar to the pre-processing for text mining. The content preparation involves:

1. Extraction of text from HTML
2. Data cleaning by filling up the missing values and smoothing the noisy data
3. *Tokenizing:* Generates the tokens of words from the cleaned up text
4. *Stemming:* Reduce the words to their roots. The different grammatical forms or declinations of verbs identify and index (count) as the same word. For example, stemming will ensure that both "closed" and "closing" are derived from the same word "close". Stemming algorithm, *Porter,* can be used here. The java code for *Porter* stemming algorithm can be obtained from *https://tartarus.org/martin/PorterStemmer /.java.bct,*
5. *Removing the stop* words: The common words unlikely to help in the mining process such as articles (a, an, the), or prepositions (such as, to, in, for) are removed.
6. *Calculate collection wide-word frequencies:* The distinct-word stem obtained after stemming process and removing the stop words results into a list of significant words (or terms). Calculating the occurrence of a significant term (t) in a collection is called collection frequency (CFt). CF counts the multiple occurrences.)

   Now, find the number of documents in the collection that contains the specific term (t). This numeric measure is the document frequency (DF t).
7. *Calculate per Document Term Frequencies* (TF). TF is a numeric measure that is used to score the importance of a word in a document based on how often it appeared in that document (Refer Example 9.1).
8. *Bag of* words: Web document is represented by the words it contains (and their occurrences.

   The following example explains the concept of CF and DF using the data of toy sales collection.

EXAMPLE 9.6

Using the table below on collection and document frequencies, which is prepared from the toys sales collection, analyze the

dJscoaot issss sso

Collection frequency (CF) and document frequency (DF)

numeric values of CFs and DFs.

Ille 1,'67 1230

SOLUTION

The table suggests that the collection frequency (CF) and document frequency (DF) can behave differently. The CF values for both *discount* and *sale* are nearly equal, but their DF values differ significantly. The reason is that the word sale is present in a large number of documents and the word discount in a less number of documents. Thus, when a query related to *discount* is generated, it must be searched in the concerned documents only.

Mining Tasks for Web Content Analytics

Following are the tasks for web content analytics:

1. classification - A supervised technique which:

(i) Identifies the class or category a new web documents belongs to from the set of predefined classes or categories

(ii) Categories in the form of a term vector that are produced during a "training" phase

(iii) Employs algorithms using term vector to categorize the new data according to the observations at the training set.

2. *Clustering* - An unsupervised technique:

(i) Groups the web documents (clustered) with similar features using some similarity measure

(ii) Uses no pre-defined perception of what the groups should be

(iii) Measures most common similarity using the dot product between two web document vectors.

3. *Identifying the association* between web documents - Association rules help to identify correlation between web pages that occur mostly together.

The other significant mining tasks are:

1. *Topic identification, tracking and drift analysis* - A way of organizing the large amount of information retrieved from the web is categorizing the web pages into distinct topics. The categorization can be based on a similarity metric, which includes textual information and co-citation relations. Clustering or classification techniques can automatically and effectively identify relevant topics and add them in a topic-wise collection library.

   Adding a new document to a collection library includes:

   (i) Assigning each document to an existing topic (category)

   (ii) Re-checking of collection for the emergence of new topics

   (iii) Tracking the number of views to a collection

   (iv) Identifying the drift in a topic(s)

2. *Concept hierarchy creation* - Concept hierarchy is an important tool for capturing the general relationship among web documents. Creation of concept hierarchies is important to understand a category and sub-categories to which a document belongs. The clustering algorithms leverage more than two clusters, which merge into a cluster. That is merging the sub• clusters into a cluster.

   Important factors for creation of concept hierarchy include:

   (i) Identifying the organization of categories, such as flat, tree or network

   (ii) Planning the maximum number of categories per document

   (iii) Building category dimensions, such as domain, location, time, application and privileges.

3. *Relevance of content* - Relevance or the applicability of web content can be measured with respect to any of the following basis:

   (i) Document relevance describes the usefulness of a given document in a specified situation.

   (ii) Query-based relevance is the most useful method to assess the relevance of web pages. Query-based relevance is used in information retrieval tools such as search engines. The method calculates the similarity between query (search) keywords and document. Similarity, results can be refined through additional information such as popularity metric as seen in Google or the term positions in AltaVista.

   (iii) User-based relevance is useful in personal aspects. User profiles are maintained, and similarity between the user profile and document is calculated. The relevance is often used in push notification services.

   (iv) Role/task-based relevance is quite similar to user-based relevance. Instead of a user, here the profile is based on a particular role or task. Multiple users can provide input to profile.

9.3.3 Web Usage Mining

Web usage mining discovers and analyses the patterns in click streams. Web usage mining also includes associated data generated and collected as a consequence of user interactions with web resources.

Figure 9.7 shows three phases for web usage mining.

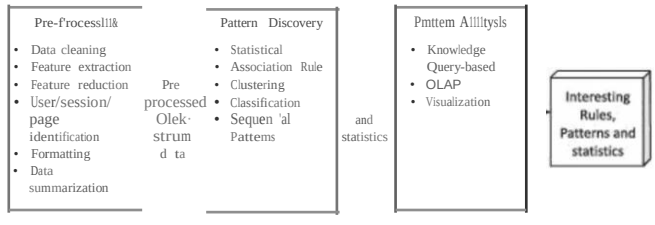


Figure 9.7 Process of web usage mining

The phases are:

1. Pre-processing - Converts the usage information collected from the various data sources into the data abstractions necessary for pattern discovery.
2. Pattern discovery - Exploits methods and algorithms developed from fields, such as statistics, data mining, ML and pattern recognition.
3. Pattern analysis - Filter outs uninteresting rules or patterns from the set found during the pattern discovery phase.

Usage data are collected at server, client and proxy levels. The usage data collected at the different sources represent the navigation patterns of the overall web traffic. This includes single-user, multi-user, single-site access and multi-site access patterns.

9.3.3.1 Pre-processing

The common data mining techniques apply on the results of pre-processing using vector space model (Refer Example 9.2).

Pre-processing is the data preparation task, which is required to identify:

(i) User through cookies, logins or URL information

(ii) Session of a single user using all the web pages of an application

(iii) Content from server logs to obtain state variables for each active session

(iv) Page references.

The subsequent phases of web usage mining are closely related to the smooth execution of data preparation task in pre-processing phase. The process deals with (i) extracting of the data, (ii) finding the accuracy of data, (iii) putting the data together from different sources, (iv) transforming the data into the required format and (iv) structure the data as per the input requirements of pattern discovery algorithm.

Pre-processing involves several steps, such as data cleaning, feature extraction, feature reduction, user identification, session identification, page identification, formatting and finally data summarization.

9.3.3.2 Pattern Discovery

The pre-processed data enable the application of knowledge extraction algorithms based on statistics, ML and data mining algorithms. Mining algorithms, such as path analysis, association rules, sequential patterns, clustering and classification enable effective processing of web usages. The choice of mining techniques depends on the requirement of the analyst. Pre-processed data of the web access logs transform into knowledge to uncover the potential patterns and are further provided to pattern analysis phase.

Some of the techniques used for pattern discovery of web usage mining are:

Statistical techniques They are the most common methods which extract the knowledge about users. They perform different kinds of descriptive statistical analysis (frequency, mean, median) on variables such as page views, viewing time and length of path for navigational.

Statistical techniques enable discovering:

(i) The most frequently accessed pages

(ii) Average view time of a page or average length of a path through a site

(iii) Providing support for marketing decisions

Association rule The rules enable relating the pages, which are most often referenced together in a single server session. These pages may not be directly connected to one another using the hyperlinks.

Other uses of association rule mining are:

(i) Reveal a correlation between users who visited a page containing similar information. For example, a user visited a web page related to admission in an undergraduate course to those who search an eBook related to any subject.

(ii) Provide recommendations to purchase other products. For example, recommend to user who visited a web page related to a book on data analytics, the books on ML and Big Data analytics also.

(iii) Provide help to web designers to restructure their websites.

(iv) Retrieve the documents in prior in order to reduce the access time when loading a page from a remote site.

Clustering is the technique that groups together a set of items having similar features. Clustering can be used to:

(i) Establish groups of users showing similar browsing behaviors

(ii) Acquire customer sub-groups in e-commerce applications

(iii) Provide personalized web content to users

(iv) Discover groups of pages having related content. This information is valuable for search engines and web assistance providers.

Thus, user clusters and web-page clusters are two cases in the context of web usage mining. Web page clustering is obtained by grouping pages having similar content. User clustering is obtained by grouping users by their similarity in browsing behavior.

Model-based or distance-based clustering can be applied on web usage logs. The model type is often specified theoretically with model-based clustering. The model selection techniques and parameters estimate using maximum likelihood algorithms, such as Expectation Maximization (EM) determines the structure of model. Distance-based clustering measures the distance between pairs of web pages or users, and then groups the similar ones together into clusters. The most popular distance-based clustering techniques include partitional clustering and hierarchical clustering (Section 6.6.3).

Classification The method classifies data items into predefined classes. Classification is useful for:

(i) Developing a profile of users belonging to a particular class or category

(ii) Discovery of interesting rules from server logs. For example, 3750 users watched a certain movie, out of which 2000 are between age 18 to 23 and 1500 out of these lives in metro cities.

Classification can be done by using supervised inductive learning algorithms, such as decision tree classifiers, Naive Bayesian classifiers, k-nearest neighbour classifiers, support vector machines.

Sequential pattern discovery User navigation patterns in web usage data gather web page trails that are often visited by users in the order in which pages are visited. Markov Model can be used to model navigational activities in the website. Every page view in this model can be represented as a state. Transition probability between two states can represent the probability that a user will navigate from one state to the other. This representation allows for the computation of a number of significant user or site metrics that can lead to useful rules, pattern, or statistics.

9.3.3.3 Pattern Analysis

The objective of pattern analysis is to filter out uninteresting rules or patterns from the rules, patterns or statistics obtained in the pattern discovery phase.

The most common form of pattern analysis consists of:

(i) A knowledge query mechanism such as SQL

(ii) Another method is to load usage data into a data cube in order to perform Online Analytical Processing (OLAP) operations

(iii) Visualization techniques, such as graphing patterns or assigning the colors to different values, can often highlight overall patterns or trends in the data

(iv) Content and structure information can filter out patterns containing pages of a certain usage type, content type or pages that match a certain hyperlink structure.

Data cube enables visualizing data from different angles. For example, *toys* data visualization using category, colour and children preferences. Another example, news from category, such as sports, success stories, films or targeted readers (children, college students, etc).

Self-Assessment Exercise linked to LO 9.2

1. Defineweb mining. Discussthe broad classifications of web mining and their applications.
2. Listthe tasks in pre-processing of web contents.
3. Howare web-content mining tasks performed using machine learning algorithms?
4. Howare topic identification, tracking and drift analysis done?

5. List and explain three phases of web-usage mining.
6. Highlight the techniques used for pattern discovery in web-usage mining giving an example of each.

9.4 PAGE RANK, STRUCTURE OF WEB AND ANALYZING A WEB GRAPH

Page Ranking, analysis of web-structure and discovering hubs, authorities and communities in web-structure

Sections 9.2 and 9.3 described text data and web contents analysis. Hyperlinks links exist between the web contents. Link analysis finds the answers to the following:

1. Can a linked (web) page rank them higher or lower?
2. Can the links be modeled as edges of graphs, structure of web as graph network, and applied the tools same as for graph analytics?
3. Can web graph mining method analyze and find a link sending spams?
4. Does a set of links correspond to a hub? Do the links correspond to an authority?
5. Does a linked page has higher authority compared to others?

Links analysis applies to domains of social networks and e-mail. The following sub-sections describe the applications of link analysis:

9.4.1 Page Rank Definition

The in-degree (visibility) of a link is the measure of number of in-links from other links. The out-degree (luminosity) of a link is number of other links to which that link points.

PageRank definition according to earlier approaches

Assume a web structure of hyperlinks. Each hyperlink in-links to a number of hyperlinks and out-links to a number of pages. A page commanding higher authority (rank) has greater number of in-degrees than out-degrees. Therefore, one measure of a page authority can be in-degrees with respect to out-degrees.

PageRank refers to the authority of the page measured in terms of number of times a link is sought after.

PageRank definition according to the new approach

Earlier approach of page ranking based on in-links and out-links does not capture the relative authority (importance) of the parents. Page and co-authors (1998) defined a page ranking method,5 which considers the entire web in place of local neighbourhood of the pages and considers the relative authority of the parent links (over children).

9.4.2 Web Structure

Web structure models as directed-graphs network-organization. Vertex of the directed graph models an anchor. Let n = number of hyperlinks at the page **U.** Assume **u** is a vector with elements $u_1, u_2, \cdots u_n$. Each page Pg **(u)** has anchors, called hyperlinks. Page Pg (v) consists of text document with m number of hyper links. v is a vector with elements $v_i, v_2, \cdots v_m$. The m is number of hyper links at Pg (v). A vertex u directs to another Page V. A page Pg (v) may have number of hyperlinks directed by out-edges to other page Pg **(w).** Consider the following hypotheses:

1. Text at the hyperlink represents the property of a vertex u that describes the destination V of the out-going edge.
2. A hyperlink in-between the pages represents the conferring of the authority.

Pages **U** and **U¢** hyperlinks u and ue out-linking to Page V. Let Page U has three hyperlinks parenting three Pages, V one, W two, X two, U' one, and Y two, respectively. Figure 9.8 shows a web structure consisting of pages and hyperlinks.

Figure 9.8 Web structure with hyperlinks from a parent to one or more pages

9.4.2.1 Dead Ends

Dead-end web pages refer to pages with no out-links. When a web page links to such pages, its page rank gets reduced. Dead ends are on a website having poor linking structure.

The web structure of service pages may have pages with a dead end. The end causes no further flows for further action and no internal links. Good website structures have the pages designed such that they specifically gently guide the visitors toward actions and towards next step. For example, if one searches for a book title on Amazon, then visitor gets links of other books also on a similar topic.

9.4.2.2 Analyzing and Implementing a System with Web Graph Mining

Number of metrics analyze a system using web graph mining. Following are the examples:

1. In-degrees and out-degrees
2. Closeness is centrality metric. Closeness, Cc(v) = I/ I', gdist(v,u), where gdist is the geodesic distance between vertex v with u and sum is over all u linked with V. Geodesic distance means the number of edges in a shortest path connecting two vertices. Assume v has an edge with w, and w has an edge with u. Assume u does not have direct edge from v. Then, geodesic distance= 2 (two edges between v and u in shortest path).
3. Betweenness
4. PageRank and LineRank
5. Hubs and authorities
6. Communities parameters, triangle count, clustering coefficient, K-neighbourhood
7. Top K-shortest paths

9.4.3 Computation of PageRank and PageRank Iteration

Assume that a web graph models the web pages. Page hyperlinks are the property of the graph node (vertex). Assume a Page, Pg (v) in-links from Pg (u), and Pg (u) out-linking similar to Pg (v), to total Nout [(Pg (u)] pages. Figure 9.9 shows Pg (v) in-links from Pg (u) and other pages.

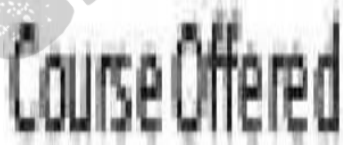


Figure 9.9 Page Pg (v) in-links from Pg (u) and other pages

Nout for page U is 7 and for V is 1 in the figure. Number of in-linking Nin for page V is 4. Two algorithms to compute page rank are as follows:

1. PageRank algorithm using the in-degrees as conferring authority

Assume that the page U, when out-linking to Page V "considers" an equal fraction of its authority to all the pages it points to, such as Pgv. The following equation gives the initially suggested page rank, PR (based on in-degrees) of a page Pgv:

$$PR(Pgv) = DC \cdot \sum_{Pp:Pto-+P,.} [PR(Pga)/N(Pgu)) \quad (9.21)$$

where N(Pgu) is the total number of out-links from U. Sum is over all Pgv in-links. Normalization constant denotes by nc, such that PR of all pages sums equal to 1.

However, just measuring the in-degree does not account for the authority of the source of a link. Rank is flowing among the multiple sets of the links. When Pgv in-links to a page Pgu, its rank increases and when page Pgu out-links to other new links, it means that N (Pgu) increases, then rank PR(Pgv) sinks (decreases). Eventually, the PR (Pgv) converges to a value.

Therefore, rank computation algorithm iterates the rank flowing computations as shown below:

EXAMPLE 9.7

Assume S corresponds to a set of pages. Initialize 'V Pg E S. Symbols mean that initialize all pages Pg contained in the S and initialize Page Rank (Pgv) for each page as follows:

$$PRinit\ (Pgv) = 1/|S| \quad (9.22)$$

How are the page ranks of the pages in a given set of pages iterated and computed till the ranks do not change (within specified margin, that means untill converge)?

SOLUTION

Iterate and compute PR (Pgv) for each page as follows:

Until ranks do not change (within specified margin) (that means converge)

for each Pgv E S compute,

$$PR(Pgv> = \sum_{P,u:Pgu-+\sim} [PR(PguJ/N(PguJ] \quad (9.22)$$

and normalization constant,

$$nc = \sum_{JI,YES} [PRnew(Pgv)] \quad (9.23a)$$

$$\text{for each Pgv e S } PR(Pgv) = nc\ PR(Pgv) \quad (9.23b)$$

2. PageRank algorithm using the relative authority of the parents over linked children

A method of PageRank considers the entire web in place of local neighbourhood of the pages and considers the relative authority of the parents (children). The algorithm uses the relative authority of the parents (children) and adds a rank for each page from a rank source.

The PageRank method considers assigning weight according to the rank of the parents. Page rank is proportional to the weight of the parent and inversely proportional to the out-links of the parent.

Assume that (i) Page v (Pgv) has in-links with parent Page u (Pgu) and other pages in set PA (v) of parent pages to v that means E PA(v), (ii) R(v) is PageRank of Pgv, (iii) R (u) is weight (importance/rank) of Pgu, and (iv) ch (u) is weight of child (out-links) of Pgu. Then the following equation gives PageRank R (v) of link v:

$$R(v) = \sum_{\bullet e\ PA(v)} [R(11)\sim ch(u)l] \quad (9.25)$$

where PA(v) is a set of links who are parents (in-links) of link v. Sum is over all parents of v. nc is normalization constant whose sum of weights is 1.

Assume that a rank source E exists that is addition to the rank of each page R (v) by a fixed rank value E(v) for Pgv. E(v) is fraction a of[l/lPA(v)I].

An alternative equation is as follows:

$$R(v) = nc\cdot\{(1-a) \sum_{\bullet ePA(v)} \left[\frac{\sim(\omega)}{\sim h(II)}\right]^{+} a\cdot E(v)\}. \quad (9.26)$$

where nc = [1/R(v)]. R(v) is iterated and computed for each parent in the set PA(v) till new value of R(v) does not change within the defined margin, say 0.001 in the succeeding iterations.

Significance of *a* PageRank can be seen as modeling a "random surfer" that starts on a random page and then at each point: E(v) models the probability that a random link jumps (surfs) and connect with out-link to Pgv. R(v) models the probability that the random link connects (surf) to Pgv at any given time. The addition of E(v) solves the problem of Pgv by chance out-linking to a link with dead end (no outgoing links).

Therefore, rank computation algorithm iterates the rank flowing computations as shown in Example 9.8.

EXAMPLE 9.8

Assume PA corresponds to a set of parent pages to a page v. Initialize V Pg E PA (v). Symbols mean that initialize all pages u contained in the set of parent pages of PA (v) and initialize Page Rank R(v) for each page as follows:

R(v) = [1/IPA(v)I]

How are the page ranks of the pages in a given set of pages iterated and computed till the ranks do not change (within specified margin, that means untill converges)?

SOLUTION

Iterate and compute R (v) for each page as follows:

Until ranks do not change that means converges (within specified margin, say 0.001)

for each v E PA (v) compute,

$$R(v) = n \cdot \{(1-ex \; L \; [\frac{R(u)}{b \; ll} + aE \; V)$$

and normalization constant,

$$n = L \; [R \; v)$$

$$f \; r \quad h \; v \in PA \; v: \quad R \quad = n \; R \; v$$

PageRank Iteration using MapReduce functions in Spark Graph

The computation of PageRank using SparkGraph method (Section 8.5),

```
graph.pageRank(0.0001) .vertices

ranksByUsername = users.join(ranks) .map{case id, (username, rank)) => (username, rank).
```

The method includes conversions to MapReduce functions and using HDFS compatible files. Functions PageRank (), ranksByUsername () do the computations using the PageRankObject. GraphX consists of these functions (GraphOps). Graphx Operators includes the functions (Section 8.5).

Static PageRank algorithm runs for a fixed number of iterations, while dynamic PageRank runs until the computed rank converges. Convergence means that after certain iterations, the rank does not change significantly and any change remains within a pre• specified tolerance. Thereafter the iterations stop.

Assume specified tolerance at the start of iterations is 0.0001 (1 in 10000). When the rank does not change beyond that tolerance, it means rank value will converge and then the iterative process will stop.

9.4.4 Topic Sensitive PageRank and Link Spam

Number of methods have been suggested for computations of topic-sensitive page ranking, RTs· The RTs (v) of a page P (v) may be higher for a specific topic compared to other topics. A topic associates with a distinct bag of words for which the page has higher probability of surfing than other bags for that topic.

Topic-sensitive PageRank method uses surfing weights (probabilities) for the pages containing the topic or bag of words corresponding to a topic. Method for creating topic-sensitive PageRank is to compute the bias to rank R(v) and thus increase the effect of certain pages containing that topic or bag of words.

Refer equation (9.25) for computations of R (v), and equation (9.26) for computations after introducing additional influence to the page. A method of introducing biasing is simple. It assumes that a rank source E exists that is additional having in-links from other pages, and thus adds to the rank of each page R (v) by a fixed (uniform) or non-uniform weight factor a. The factor a is a multiplication factor to actual in-links without the bias.

Recapitulate equation (9.26). Probability of random jump to page v is E(v). An alternative equation for topic sensitive PageRank, R(v) computation for page P (v) is as follows:

$$R(v) = \text{n.:-\{o-a,)-P(v)} \sum_{\text{IIE P.-'(,·)}} [1\sim(U)j]+a, .E(v)\}.$$

Probability of random jump to page v is E (v). at = 0 for page unrelated to a topic a is not 0 for page related to a topic. at = surfing probability for in-links for a topic t. Further, coefficient (1- a) is considered as biasing factor depending on the web page P (v) selected for a queried topic *t.*

The page is having in-links from other pages. Assume N, is number of topics to which a page is sensitive to surfing those topics. Effect of topics on PageRanks increases by using a non-uniform N, x 1 personalization vector for surfing probability p. Higher a means higher p.

Assume that the topics are tl, t2, ..., tn. Fix the number Nt. RTs is to be computed for each of them. Therefore, compute for each topic tj' the PageRank scores of page v as a function of tj, which means compute R (v, j), where j = 1, 2, ..., n. That also means compute the n elements of a non-uniform N, x 1 personalization vector Rrs (v) for tl, t2, ..., tn.

Link Spam

Effects of a *link spam* can be nullified using the topic-sensitive PageRank algorithm. Link Spam tries to mislead the PageRank algorithm. A link spam attempts to make PageRank algorithm ineffective. The spam assisting pages connects to the page repeatedly and increases the in-degree of a page, thereby enhancing the rank to a large value.

A link spam creator website ws also has a page ls for whom *ws* attempts to enhance the PageRank. The w_5 has a large number of assisting pages al_5 which out-links to *ls* only. The *als* pages also prevent the PageRank of l_5 from being lost. A spam mass consists of w_5, ls and its *als* pages.

Methods nullify the effect by introducing a trust rank for a page u used in equation (9.29) and tracing spam mass of in-link pages to the page v.

Following are the steps for finding spam mass:

1. A distant topic sensitive page has unusually high in-degrees compared to the other pages of the same topic. A plot known as power-law plot is drawn between the log of number of web pages on the y-axis out-linking to the page v and logs of them in• degrees of v on the x-axis.
2. Plot is nearly linear as the number exponential decays is within degrees. N is proportional to exp (-d), where d is decay constant.
3. An unusual pattern with marked deviation from near linearity identifies the distant link spam mass.

9.4.5 Hubs and Authorities

A hub is an index page that out-links to a number of content pages. A content page is topic authority. An authority is a page that has recognition due to its useful, reliable and significant information.

Figure 9.lO(a) shows hubs (shaded circles) with the number of out-links associated with each hub. Figure 9.IO(b) shows authorities (dotted circles) with the number of in-links and out-links associated with each link.

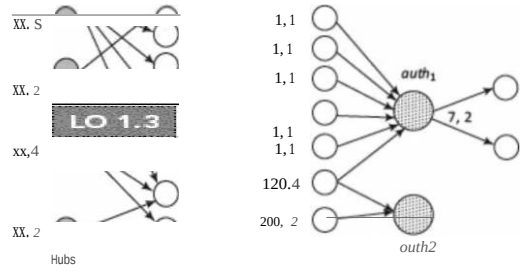


Figure 9.10 (a) Hubs (shaded circles) and (b) Authorities (dotted circles)

In-degrees (number of in-edges from other vertices) can be one of the measures for the authority. However, in-degrees do not distinguish between an in-link from a greater authority or lesser authority.

Authority, $auth_1$ in Figure 9.lO(b) has in-links from 6 vertices (in-degrees = 6) and $auth_2$ has in-links to just 2 (in-degree = 2). However, $auth_1$ has link with six vertices with in-degrees = 1, 1, 1, 1, 1 and 120 (total = 125). Authority, $auth_2$ has links with two vertices with in-degrees= 120 and 200 (total= 220). $Auth_2$has association with greater authorities. Therefore, in-degrees may not be a good measure as compared to authority.

Kleinberg (1998) developed the Hypertext-Induced Topic Selection (HITS) algorithm.6 The algorithm computes the hubs and authorities on a specific topic t. The HITS analyses a sub-graph of web, which is relevant to t. Basis of computation is (i) hubs are the ones, which out-link to number of authorities, and
(ii) authorities are the ones, which in-link to number of hubs. A bipartite graph exists for the hubs and authorities.

Consider a specifically queried topic t. Following are the steps:

1. Let a set of pages discover a root set R using standard search engine. Root pages may limit to top 200 for ***t.***
2. Find a sub-graph of pages S, using a query that provides relevant pages for t and pointed by pages at R. Sub-graph S pages form Set for computations as it includes the children of parent R and limit to a random set of maximum 50 pages returned by a "reverse link" query.
3. Eliminate purely navigational links and links between two pages on the same host.
4. Consider only u (Illull = 4-8) pages from a given hyperlink as pointer to any individual page. (Section 9.4.2)

Sub-graph for HITS consisting of root set R of pages and children of parents in the sub-graph S. Figure 9.11 shows subgraph S for HITS consisting of root set R of pages and all the pages pointed to by any page of R.

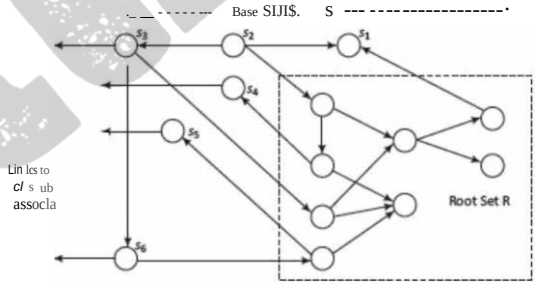


Figure 9.11 Sub-graph for HITS consisting of root set R of pages and base sub-graph *S* including all the pages pointed to by any page of *R.*

The left directed leftmost arrows from s3, s4 , s5 and s6 are pointing to nodes in sub-graph(s) associated to S. The following example explains the algorithm steps to compute hub score and authority score.

EXAMPLE 9.9

Assume that v has number of in-links and v has number of out-links. Assume S corresponds to base set of pages and R corresponds to root set. (i) Initialize S to R. (ii) lnitialize \:/ u E S. Symbols mean that initialize all pages u, contained in the S.

(iii) Normalization constant is nc. The (i) hub (v) hub score and (ii) auth authority score of page v for each page is as follows:

For each v E S, auth (v) = 1; hub(v) = 1; nc = 1; (9.30)

How are the hub and authority of pages in a given set of pages iterated and computed till the ranks do not change (within specified margin, that means untill converges)? Usually 20 iterations converge the result within margin, usually set to 0.001.

SOLUTION

Iterate and compute auth (v) and hub (v) for each page as follows:

Until ranks do not change (within specified margin) (that means converges)

for each v E S compute,

$$\text{authrvj}_{e}\ \text{nc-}\ L\ [\text{hub(ul}]. \qquad (9.\text{Jla})$$

$$\text{hub(v)=nc2}\cdot\ \underset{\text{uv~a}}{L}\ [\text{auth(u)}]. \qquad (9.\text{Jlb})$$

and normalization constant,

$$\text{nd} = \underset{\text{P,VES}}{L}\ \text{ll[auth(v)f.} \qquad (9.32.1)$$

$$\text{nc2} = \underset{\text{P,VES}}{L}\ \text{ll[authMf.} \qquad (9.\text{J2b})$$

$$\text{for each v e ~: authtv) = nc Lauthtv]: authtv) = nc l .autht v): authtv):} \qquad (9.\text{J2c})$$

Difference between HITS and PageRank

HITS considers mutual reinforcement between authority and hub pages. PageRank ranks the pages just by authority and does not take into account distinctions between hubs and authorities. HITS considers the local neighbourhood between 4 to 8 pages surrounding the results of a query, whereas PageRank is applied to the entire web. HITS depends on topic t, while PageRank is topic• independent. PageRank effects by
dead-ends.

9.4.6 Web Communities

Web communities are web sites or collections of websites, which limit the contents view and links to members. Examples of web communities are social networks, such as Linkedln, SlideShare, Twitter and Facebook.

The communities consist of sites for do-it-yourself sites, social networks, blogs or bulletin boards. The issues are privacy and reliability of information.

Metric for analysis of web-community sites are web graph parameters, such as triangle count, clustering coefficient and K• neighbourhood.

K-neighbourhood analysis means the number of 1st neighbour nodes, 2nd neighbour nodes, and so on
(K = 1, 2, 3, 4 and so on).

K-core analysis means the number of cores within a marked area. A core may consist of a triangle of connected vertices. A core may consist of a rectangle with interconnected edges and diagonals. A core may also be a group of cores.

Spark Graphx (Section 8.5) described functions for degree centralities, degree distribution, separation of degree, betweenness centralities, closeness centralities, neighbourhoods, strongly connected components, triangle counts, PageRank, shortest path, Breadth First Search (BFS), minimum spanning tree (forest), spectral clustering and cluster coefficient.

9.4.7 Limitations of Link, Rank and Web Graph Analysis

Following are the limitations oflink and web graph analysis:

1. Search engines rely on metatags or metadata of the documents. That enhances the rank if metadata has biased information.
2. Search engines themselves may introduce bias while ranking the pages of clients higher as the pages of advertising companies

may provide higher searches and hence lead to biased ranks.

3. A top authority may be a hub of pages on a different topic resulting in increased rank of the authority page.
4. Topic drift and content evolution can affect the rank. Off-topic pages may return the authorities.
5. Mutually reinforcing affiliates or affiliated pages/sites can enhance each other's rank and authorities.
6. The ranks may be unstable as adding additional nodes may have greater influence in rank changes.

Self-Assessment Exercise linked to LO 9.3

1. Write and explain the equations for computing PageRank using relative authority of parent nodes.
2. Show diagramatically network organization model of directed graphs for the structure of the web. How are the page hub and page authority computed?
3. What are the metrics which Spark GraphX compute?
4. How does the equation for computing the hub of a page differ from the computing authority of a page?
5. How does link spam function? How is the link spam discovered from the plot between the number of web pages and in• degrees?

9.51 SOCIAL NETWORKS AS GRAPHS AND SOCIAL NETWORK ANALYTICS

A social network is a social structure made of individuals (or organizations) called "nodes," which are tied (connected) by one or more specific types of inter-dependency, such as friendship, kinship, financial exchange, dislike or relationships of beliefs, knowledge or prestige. *(Wikipedia)*

Representation of social networks as graphs, methods of social network analysis, finding *the* clustering in social netviork graphs, evaluating the Sim Ran counting triangles (cliques) and discovering the communities

Social networking is the grouping of individuals into specific groups, like small rural communities or some other neighbourhoods based on a requirement. The following subsections describe social networks as graph, uses, characteristics and metrics.

9.5.1 Social Network as Graphs

Social network as graphs provide a number of metrics for analysis. The metrics enable the application of the graphs in a number of fields. Network topological analysis tools compute the degree, closeness, betweenness, egonet, K-neighbourhood, top-K shortest paths, PageRank, clustering, SimRank, connected components, K-cores, triangle count, graph matches and clustering coefficient. Bipartite weighted graph matching does collaborative filtering.

Apache Spark Graphx and IBM System G Graph Analytics tools are the tools for social network analysis.

Centralities, Ranking and Anomaly Detection

Important metrics are degree (centrality), closeness (centrality), betweenness (centrality) and eigenvector (centrality). Eigenvector consists of elements such as status, rank and other properties. Social graph-network analytics discovers the degree of interactions, closeness, betweenness, ranks, probabilities, beliefs and potentials.

Social network analysis of closeness and sparseness enables detection of abnormality in persons. Abnormality is found from properties of vertices and edges in network graph. Analysis enables summarization and find attributes for *anomaly.*

Social network characteristics from observations in the organizations are as follows:

1. Three-step neighbourhoods show positive correlation between a person and high performance. Betweenness between vertices and bridges between numbers of structures are not helpful to the organization. Too many strong links of a person may have a negative correlation with the performance.
2. Social network of a person shows high performance outcome when the network exhibits structural diversity. Person with a social network with an abundant number of structural holes exhibits higher performance. This is because having diverse relations help an organization.

Social network analysis enables detection of an anomaly. An example is detection of one dominant edge which other sub-graphs are follow (succeed). *Ego network* is another example. The network structure is such that a given vertex corresponds to a sub-graph where only its adjacent neighbours and their mutual links are included.

The analysis enables spam detection. Spam is discovered by observation of a near star structure.
Figure 9.12 shows discovering anomaly, ego-net and spam from the analysis.

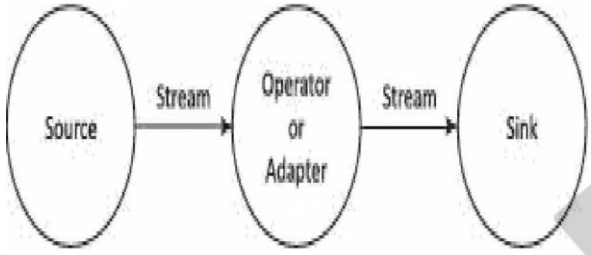


Figure 9.12 Discovering anomaly, ego-net and spam (using near star) from the analysis

Social network has concerns of privacy, security and falsehood dissimentation. Security issues are phsishing attacks and malwares.

9.5.2 Social Graph Network Topological Analysis using Centralities and PageRank

Social graph network can be topologically analyzed. The centralities (degree, closeness, effective closeness and betweenness) and PageRank (vertexRank similar to PageRank in web graph network) are the parameters analyzed.

Degree

Degree of a graph vertex means the total number of edges linked to that. *In-degree of a vertex* means the number of in-edges from the other vertices. *Out-degree of a vertex* means the number of out-edges to other vertices to which that vertex directs. *Degree distribution function* means the distribution function for the degrees of vertices (Section 6.2.5 described the common distribution functions).

Closeness

Graph vertex closeness $C_C(v)$ is a way of defining the centrality of a vertex in reference to other vertices. Sum is the overall vertices connected to other vertices u. The u is a subset of vertices in set V.

The centrality (closeness index), c is function of distances of vertices.

$$C_C(v) = \left[\sum_{u \in V} d(u, v)\right]$$

where d (u, v) is distance between u and v for path traversal.

Effective Closeness

Effective closeness $C_{EC}(v)$ can also be analyzed. Use approximate average distance from v to all other vertices in place of the shortest paths. C_{EC} reduces run time for cases with a large number of edges and near linear scalability in computations.

Betweenness

Graph vertices betweenness means the number of times a vertex exists between the shortest path and the extent to which a vertex is located 'between' other pairs of vertices. Betweenness $C_B(v)$ of a vertex v requires calculating the lengths of shortest paths among all pairs of vertices and computations of the summation for each pairing vertex in V.

PageRank

PageRank is a metric for the importance of each vertex in a graph, assuming an edge from *vl* to *v2* represents endorsement of importance of *v2* by *vl* by connecting, following, interacting, opting for relationship, sharing belief or some other means.

Contacts Size

Contacts size means a vertex connection to many vertices. The size of each vertex does not convey any meaningful information. A big social graph network will also require high maintenance cost.

Indirect Contacts

Indirect contacts metric means betweenness, which is the sum of the shortest paths within geodesic distances from all other pairing vertices. Three-step contact metric means a number of edges to other vertices plus the number of edges from other vertices within geodesic distances = < 3.

Both metrics convey meaningful information. The indirect contacts metric has meaning in terms of magnitude of betweenness centrality.

Structure Diversity

Structure diversity metric means that social graph has access to diverse sub-graphs (knowledge).

9.5.3 Social GraphNetwork Analysis using K-core and NeighbourhoodMetrics

K-core is a sub-graph in a graph network structure. *Graph Vertex* Kth neighbourhood is number of 1st neighbour vertices, 2nd neighbour vertices and so on to a querying vertex that are correlated, linked, and have weighted correlations or the associations.

K-nearest neighbourhood (KNN) finds K-similar objects, items, or entities, which are nearest neighbours after computing the similarities. For example, KNN is K-documents (or books) in the large number of text documents (books) that are most similar to the queried document.

Collaborative filtering for frequent itemsets uses weighted bipartite graph matching.

Figure 9.13 shows the K-cores and K-neighbourhood metrics for a social network graph. The figure also shows frequent itemsets obtained from collaborative filtering algorithm (Sections 6.4 and 6.8.1).

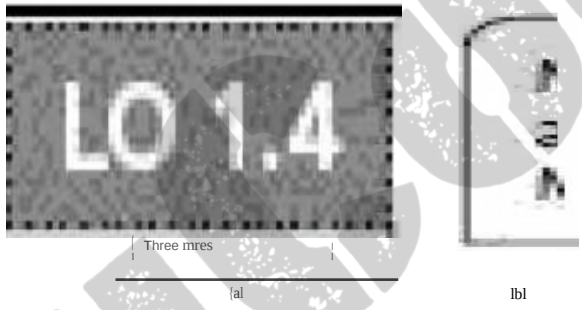


Figure 9.13 (a) K-cores and K-neighbourhoods with K = 1, 2, 3 and 4 and (b) Frequent itemsets from collaborative filtering algorithm (weighted bipartite graph matching)

Figure 9.13(a) shows three cores of two triangles, one quadrilateral, two cores of one pentagon and one traingle) in K• neighbourhoods. K = 1, 2, 3 and 4. Figure 9.13(b) shows frequent itemsets from collaborative filtering algorithm (weighted bipartite graph matching).

9.5.4 Clusteringin Social Network Graphs

One of the methods of detecting communities from social graph analysis is finding clustering and cluster coefficients. A clustering coefficient is a metric for the likelihood that two associated vertices of a vertex are also associated with other vertices. A higher clustering coefficient indicates a greater association and cohesiveness.

Connected components mean components of the datasets (represented by properties of vertices) connected together. For example, finding student-teacher datasets, social network datasets, etc.

9.5.5 SimRank

Similarity can be defined by properties of graph vertices. For example course, subject, student, scientist, Java programmer, status, values, or any other salient characteristic. Social network analysis of graphs computes *SimRank.*

SimRank is the metric for measuring similarity between vertices of the same type. The computation starts from a vertex possessing specific property and path traversals through the edges search the similarities. The vertices having similar properties are counted to the SimRank. The counting continues till counts per unit traversals converge within a prefixed margin, say .001. SimRank converges to a value which is applicable for path traversals within, say geodesic distance, say up to 200. The computations are analogous to ones for PageRank as in Example 9.7

9.5.6 Counting Triangles and Graph Matches

One of the methods of detecting communities is counting of triangles. A triangle means three vertices forming a triangle with edges interconnecting them.

Triangle count refers to the number of triangles passing through each vertex. The count is a measure of clustering. A vertex is part of a triangle when it has two adjacent vertices with an edge between them.

Graph matches are computed using filtering or search algorithm, which uses the properties, labels of vertices, edges or the geographic locations.

Figure 9.14 shows triangles and triangles between similar graph properties found from graph matches. Edge labels show the GPAs of students socially connected.

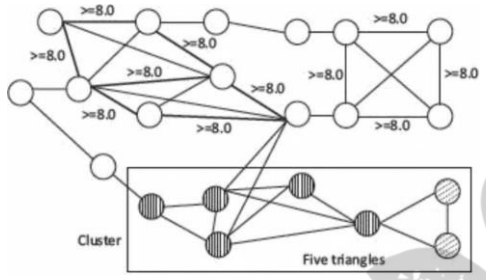


Figure 9.14 Clustering of five triangles and three matches of graphs

9.5.7 Using SparkGraph (Map-Reduce) for Network Graphs

Section 8.5 describes Spark GraphX algorithms for analyzing graphs. Connected components compute by graph. connectedComponents () . vertices method in SparkGraph. Connected Components Algorithm labels each connected component of the graph with an ID. Each connected component ID is ID of the lowest-numbered vertex. For example, in a social network, connected component objects can approximate clusters. GraphX contains an implementation of the algorithm in the ConnectedComponentsObject. The clusters are found by discovering close-by connected components using closeness centrality metric.

SparkGraphX triangle-count algorithm computes the number of triangles passing through each vertex. The count is a measure of clustering. TriangleCount requires the edges to be in canonical orientation (srcld < dstld). Source vertex ID is srcld and Destination vertex ID is dstID. Graph is partitioned using *Graph.partitionBy* operator.

9.5.8 Direct Discovery of Communities

Three metrics identify groups and communities from a social graph:

1. Cliques - A clique forms by a set of vertices when each of the vertices directly connects to every other individual vertex through the edges. Detecting the cliques leads to direct discovery of communities.
2. Structurally cohesive blocks.
3. Social circles from connections and neighbourhoods

A bridge enables the link between two groups. Application of analyzing communities, SimRanks and bridges are finding a set of experts, specific areas of expertise, and ranking the expertise in an organization.

Experience in social science fields shows that the social network of a person is the key indicator of the stature of the person and his/her success potential. Social graph analysis enables finding key bridges and persons with most connections.

Figure 9.15 shows a social graph with two cliques and a bridge.

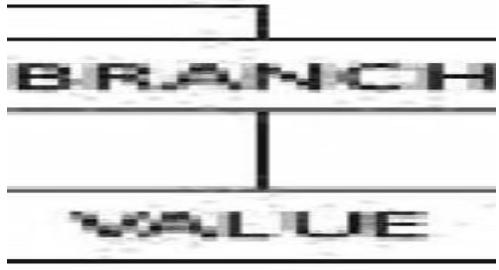


Figure 9.15 Two cliques in a social graph network and a bridge between the cliques

Clique 1 has set of four vertices, each connected with three edges to three others. Clique 2 has five vertices, each connected by edges to other four. Two edges in the figure provide the bridge between two sub graphs, on left and right sides.

Self-Assessment Exercise linked to LO 9.4

1. How do the metrics analyze a social network graph of persons in an organization? How do they relate to inter-dependency, performance, groups, expertises, beliefs, knowledge or prestige?
2. Define the terms degree, closeness, betweenness, egonet, K-neighbourhood, top-K shortest paths, PageRank, clustering, SimRank, connected components, K-cores,triangle count, graph matches and clustering coefficient.
3. How the cliques discover communities from social network analysis?
4. What are the uses of Apache Graphx Connectedcomponents and triangles count methods in social graph anaysis?

KEY CONCEPTS

anomaly detection

authority

bag of words

betweenness

centralities

clique

closeness

collaborative filtering

collection wide-word frequency

concept extraction

content relevance

document frequency

documents classification

documents clustering

effective closeness

ego net

feature selection

HITS algorithm

hub

hyperplane

in-degrees

KNN

link analysis

marginalization

Naive Bayes classifier

out-degrees

outliers

PageRank

part-of-speech tagging

pattern analysis

pattern discovery

sequential patterns

SimRank

social network

social network graph

spam detection

structured text

SVM classifier

term frequency

text analytics process pipeline

text cleanup

text features generation

text mining

text pre-processing

TF-IDF

Top K shortest paths

triangles count

unstructured text

vector space model

web community

web content analytics

web graph

web structure

web usage mining

,

L09.1

1. Text mining techniques help revealing the patterns and relationships in large volumes of textual content that are not directly

visible. The mining leads to new business opportunities and improvements in processes.

2. Five phases in text mining are (i) text pre-processing, (ii) feature generation, (iii) feature selection, (iv) text data mining, and (v) analysing the results. Text analytics involves provisions of strong integration with the already existing database, artificial intelligence, machine learning, and text mining techniques such as, information retrieval, natural language processing, classification, clustering and knowledge management, respectively.
3. Machine learning based text classification methods are (i) K nearest neighbour classifier, (ii) Naive Bayes method, (iii) decision trees, (iv) decision rules classification, and (v) support vector machines.
4. Naive Bayes classifier is a simple, probabilistic and statistical classifier. The classifier computes the conditional probability tables.
5. SVMs based classifier is a discriminative classifier formally defined by a separating hyperplane. SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that select the only effective elements in the training set.

L09.2

1. Web data mining is a process of discovering patterns in large datasets to gain knowledge. The process can be shown as Raw Data ↦ Patterns ↦ Knowledge. Web data refers to (i) web content - text, image, records, (ii) web structure - hyperlinks and tags, and (iii) web usage - http logs and application server logs.
2. Steps for pre-processing of web-data are quite similar to pre-processing for text mining. The steps include collection of wide• word frequencies and document frequencies. Machine learning techniques for web content analytics are clustering, classification and association rule mining.
3. Web usage mining discovers and analyses the patterns in click stream and associated data generation and collection as a consequence of user interactions with web resources on the World Wide Web.
4. A link spam creator website ws also has a page *ls.* ws attempts to enhance the PageRank of *ls.*
5. A hub is an index page that out-links to number of content pages. A content page is topic authority. Authority is a page that has recognition due to provisioning useful, reliable and significant information. HITS algorithm computes the hubs and authorities on a specific topic t.
6. Web community is website or collection of the websites that limits the view of contents, and that links the members (for example, Linkedln). Metric for analysis of web community sites are web graph parameters, such as triangle count, clustering coefficient and K-neighbourhood.

L09.3

1. Link analysis enables finding the PageRank, centralities, hubs, and authorities. Page ranking method considers the entire web in place of local neighbourhoods of the pages. PageRank of a page refers to relative authority of the parents out-linking to the page.
2. Web structure models as directed-graphs network-organization. A page may have a number of hyperlinks directed by out• edges to other pages. Text at the hyperlink represents the property of vertex that describes the destination of the out-going edge. A hyperlink in-between the pages represents the conferring of the authority.
3. SparkGraph includes conversions to MapReduce functions and use HDFS compatible files. Page rank() and ranksByUsername() are static and dynamic methods compute on the PageRankObject

L09.4,

1. A social network is a social structure made of individuals (or organizations) called "nodes", which are tied (connected) by one or more specific types of interdependency, such as friendship, kinship, financial exchange, dislike or relationships of beliefs, knowledge or prestige
2. Social network topological analysis tools compute the degree, closeness, betweenness, egonet, K-neighbourhood, Top-K shortest paths, PageRank, clustering, SimRank, connected components, K-cores, triangle count, graph matches and clustering coefficient. Bipartite weighted graph matching does collaborative filtering.
3. Analysis enables summarization and find attributes for anomaly.
4. Apache Spark Graphx includes PageRank, ConnectedComponents and TrianglesCount algorithms, and fundamental operations

for social graph analytics.

5. Analysis of cliques discovers groups and communities. Analysis also finds the bridge between the cliques.

Objective Type Questions 11

Select one correct-answer option for each questions below:

9.1 The term *text analytics* evolves from (i) provisioning of strong integration with the already existing (ii) database, (iii) artificial intelligence, (iv) machine learning, and (v) text Data Store techniques such as (vi) information retrieval, (vii) natural language processing,
(viii) classification, (ix) clustering, and (x) knowledge management, respectively.

(a) all except ii and iv

(b) ii to ix

(c) all except ii, iii, iv and vii

(d) all

9.2 SVMs main uses are (i) classification based on the outputs taking discrete values in a set of possible categories, (ii) separation or prediction, if something belongs to a particular class or category. Other uses are (iii) finding a decision boundary between two categories, (iv) clustering, (v) regression analysis, and regression, if continuous real-valued output (continuous values of x, in place discrete n values, x2, x3, ..., xn), and (vi) discriminative classifier.

(a) all except iii

(b) i, ii and iv

(c) all except iv

(d) i to v

9.3 Applications of (i) web content mining, and (ii) web-structure mining of web documents are: (iii) Classifying the web documents into categories, (iv) identifying the topics of the web documents, (v) creation of tables and databases, (vi) finding similar web pages across different web servers, and (vii) relevance or the applicability of web content measured with respect to a basis, such as making recommendations, filtering or querying

(a) all except vii

(b) all except ii

(c) all except iv

(d) i tov

9.4 The HITS analyses a (i) subgraph of web, which is relevant to (ii) topic t, (iii) query q. The assumptions are (iv) authorities are the ones, which out-link to number of hubs, and (v) hubs are the ones, which in-link to number of authorities. (vi) A bipartite graph exists for the hubs and authorities. (vii) First *set* of pages discovers a root set R using standard search engine, then (viii) finds a sub-graph of pages S, using a query that provides relevant pages for t and pointed by pages at R.

(a) all except iii to v

(b) all except iii and vi

(c) all except vi

(d) all except iv and vi

9.5 Web contents mining tasks are: (i) finding clustering, (ii) classifying, (iii) mining association rules, and (iii) topic identification, tracking and drift analysis for adding new documents to a collection library. Other tasks are: (iv) assigning by rechecking for the emergence of new topics, (v) creation of concept hierarchy, building of category dimensions, such as domain, location, time, application, privileges, and (vi) measuring the relevance or the applicability of web content on basis of documents, queries, roles or tasks or user profiling.

(a) all except vii and viii

(b) all

(c) all except vii

(d) All except vi to viii

9.6 The most common form of pattern analysis consists of (i) a knowledge query mechanism such as SQL, (ii) loading usage data into a data cube in order to perform OLAP operations,
(iii) visualization techniques, such as graphing patterns or assigning colors to different values. The analysis also finds (iv) content and structure information which can filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

(a) all except iii

(b) all

(c) all except i and ii

(d) i to ii

9.7 PageRank method considers (i) the entire web in place of a local neighbourhood of the pages, (ii) queries top 10 pages, and (iii) considers the relative authority of the children pages with respect to parent page. *PageRank method* considers assigning weight (iv) as 1, and (v) according to the rank of the parents. (vi) PageRank is inversely proportional to the weight of the parent and proportional to out-links of the parent.

(a) i, iii, and v

(b) iandv

(c) all except ii and vi

(d) all

9.8 Social network graph analysis tools do the (i) clustering analysis which means the number of 1st neighbour nodes, 2nd neighbour nodes, and so on. (K = 1, 2, 3, 4 and so on), (ii) social network community and network analysis. The graph analysis finds the (iii) close-by entities, (iv) fully mesh-like connected sets, (v) network graph analysis beside centralities, (vi) also does computations of the *property* of the links, (vii) rectangle counts, and (viii) clustering coefficient.

(a) i to vi

(b) all except i, vi and vii

(c) ii to iv

(d) i to iii, v, viii

Part II

9.1 How are the features evaluated in the text documents? (LO 9.1)

9.2 Explain five phases and steps in the phases during text analytics. (LO 9.1)

9.3 When is the Naive Bayes conditional probabilities based classifier used? When is the support vectors based discriminative classifier used? Write details of each. (LO 9.1)

9.4 What are the tasks in web data analytics? Describe the pre-processing steps and mining tasks in web contents analytics. (LO 9.2)

9.5 How is the emergence of new topics discovered? How do concept hierarchy create and build from category dimensions, such as domain, location, time, application and privileges? (LO 9.2)

9.6 How does the web usage mining discover and analyze patterns in click stream, and generate and and collect associated data? (LO9.2)

9.7 Describe various link analysis metrics used for analytics. How is PageRank iterated and computed using relative authority of in linking pages? How does ranking algorithm compute topic-sensitive PageRank? (LO 9.3)

9.8 How does structure of web model as graph network? Draw a diagram for web graph nodes and edges. What are the metrics computed for a web graph? (LO 9.3)

9.9 Describe HITS algorithm to iterate and compute the hubs and authorities? (LO 9.3)

9.10 How does social graph analysis relate to positivity and negativity analysis about the persons? How does social graph network anomaly detection help an organization? (LO 9.4)

9.11 How are social graph analytics metrics, degree, closeness, betweenness, egonet, K-neighbourhood, Top-K shortest paths and SimRank computed by path traversals? (LO 9.4)

9.12 What are the operators provisioned in Apache Spark Graphx for social network graphs analysis? (LO 9.4)

Practical Exercises

9.1 List the steps in the methods used for grouping the text documents into clusters, automating the document organization, topic extraction. Take the example of HTML pages or your University or Company website. (LO 9.1)

9.2 Explain how text analytics tasks performs using Python library *nltk*. (LO 9.1)

9.3 List the steps in document clustering method. How do you use the clusters for the fast information retrieval or filtering? Take the example of student grade cards or Company annual reports.
(LO 9.1)

9.4 List the steps in classifying web documents into categories, identifying similar pages across different web documents to classify them as web pages of a university or company. (LO 9.2)

9.5 List the steps in recommendations for top N relevant documents in a collection or portion of a collection. List the steps in filtering- show/hide documents based on most/least relevancy.(LO 9.2)

9.6 Using Example 9.7, write algorithms for PopularityRank, SimRank and best student search. (LO 9.3)

9.7 Rewrite PageRank and HITS algorithms using vectors and matrices. (LO 9.3)

9.8 Write the steps in performing bipartite weighted graph matching in social network graph analysis. (LO 9.4)

9.9 Describe steps to compute the triangles, junction trees, shortest paths and top K-shortest paths and discover the communities in social network graphs of students. (LO 9.4)

1 http://www.nactem.ac.uk/brochure/NaCTeM_Brochure.pdf

2 https:/ /www .ibm.com/ support/knowledgecenter / en/ SS3RA7_18.1.1/ta_guide _ddita/textmining/ shared_entities/tm_intro _tm_defined.htn

3 https:/ /blogs.aws.amazon.com/bigdata/ post/Tx22THFQ9MI86F9/ Applying-Machine-Learning-to-Text-Mining-with-Amazon• S3-and-RapidMiner

4 https:/ /www.ling.upenn.edu/ courses/Fall_2003 /lingOOl/ penn_ treebank_pos.html

5 http://papers.cumincad.org/data/works/att/2873.content.pdf "The Anatomy of a Large-Scale Hypertextual Web Search Engine" Sergey Brin Lawrence Page, 1998

6 J. Kleinberg (1998), Authoritative sources in a hyperlinked environment, Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms. A longer version appears in the Journal of the ACM 46, 1999.Available from http://www.cis.hut.fi/Opinnot/T-61.6020/2008/pagerank_hits.pdf