**REGRESSION MODELS IN DATA MINING AND ANALYSIS**

Regression is the data mining technique used to predict a range of continuous values given a certain data set. The dependent (response) variable is what we want to predic while the independent (predictor) variables are the attributes of interest which are known.

Regression analysis model relationship between dependent (response) variable and one or more independent ( predictor) variables.It is used when independent (predictor) variables are continuous values.

The Regression analysis methods

1.Linear regression

a. Straight-line linear regression

The Straight-line linear regression gives the estimate of relationship between dependent variable and independent variable and uses mathematical formula of a straight line y = mx + b.

y-intercept Regression Analysis

m1 : slope

m and b are regression coefficients

From this relationship between X and Y is a straight line with few outliers. e.g There is a relationship between income and level of education the X value would increase at the same rate, making the relationship between them a straight line..

Method of least squares estimates the best-fitting straight line as the one that minimizes the error between the actual data and the estimate of the line

b.Multiple linear regression

Multiple linear regression estimate the relationship between dependent variable and more than one predictor variables.

the larger the number of predictor attributes the slower the performance.

Example of multiple linear regression There is a relationship between savings and salary and spending .

2. Non-linear regression

Non-linear regression is used to come up with accurate model when the data does not show a linear dependence . eg

y = w + w x + w x2 + w x 3

3.Generalized linear model Regression

Generalized linear model is used in categorical response variables

Types of G.L.M includes:

a. Logistic regression- models the probability of event occurring as a linear function of a set of independent (predictor) variables

b. Poisson regression - models the data that exhibit Poisson distribution

4.Log-linear models

In Log-linear models attributes must be categorical and continuous-valued attributes must first be discretized.

5.Regression trees and Model trees

Trees is used to predict continuous values rather than class labels Regression and model trees tend to be more accurate than linear regression when the data are not represented well by a simple linear model.

**Log Linear MODELS IN DATA MINING AND ANALYSIS**

loglinear modeling a technique used which involves fitting models to the observed frequencies in the cross-tabulation of categoric variables that is cell counts in contingency table. It is a specialized cases of generalized linear models for Poisson distributed data.

Loglinear models are mostly used to evaluate multiway contingency tables that involve three or more variables. Also used to analyze the relationship between two categorical variables (2-way contingency tables).

loglinear models demonstrate association between variables and these variables investigated by log linear models are all treated as response variables. That is no distinction is made between dependent and independent variables.

When one or more variables are treated as explicitly dependent and others as independent, then logit or logistic regression should be used instead. When the variables investigated are continuous and cannot be broken down into discrete categories then logit or logistic regression would again be the appropriate analysis

The loglinear model is one of the specialized cases of generalized linear models for Poisson-distributed data. Loglinear analysis is an extension of the two-way contingency table where the conditional relationship between two or more discrete, categorical variables is analyzed by taking the natural logarithm of the cell frequencies within a contingency table. Although loglinear models can be used to analyze the relationship between two categorical variables (two-way contingency tables), they are more commonly used to evaluate multiway contingency tables that involve three or more variables. The variables investigated by log linear models are all treated as “response variables”. In other words, no distinction is made between independent and dependent variables. Therefore, loglinear models only demonstrate association between variables. If one or more variables are treated as explicitly dependent and others as independent, then logit or logistic regression should be used instead.

Logit or logistic regression would be used when variables being investigated are continuous and can not be broken down into discrete categories,Also Log linear analysis uses likelihood statistic which approximate to chi-square distribution in a large sample size.

**Method Concept hierarchy generation for categorical data**

The Concept hierarchies are used to reduce the data by collecting and replacing low level concepts with higher level concepts.categorical variable is a variable type with two or more categories. Sometimes called a discrete variable, it is mainly classified into two (nominal and ordinal).

1.Categorical data: no ordering among values

2.Specification of a partial ordering of attributes explicitly at the schema level by users or experts.

3.Specification of a portion of a hierarchy by explicit data grouping

4.Specification of a set of attributes, but not of their partial ordering

5.Specification of only a partial set of attributes

6.Specification of a set of attributes, but not of their partial ordering

7.Specification of only a partial set of attributes.

References

J. Han, M. Kamber, Data Mining: Concepts and Techniques, Elsevier Inc. (2006). (Chapter 6)

I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Regression Analysis Edition, Elsevier Inc., 2005. (Chapter 6)