International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-4, Issue-1, January 2016

Bayesian Linear Regression in Data Mining

K.Sathyanarayana Sharma, Dr.S.Rajagopal

***Abstract*— This paper is Bayesian Linear Regression in Data Mining and methods substantially differ from the new trend of Data Mining From a Statistical perspective Data Mining can be viewed as computer automated exploratory data analysis of large complex data sets. Despite the obvious connections between data mining and statistical data analysis most of the methodologies used in Data Mining have so far originated in fields other than Statistics.**

***Index Terms*— Ordinary Least Square Method, Bayesian Linear Regression Method, Ordinary Least Square Estimation , Conjugate Prior Distribution, Posterior Predictive Distribution.**

I. INTRODUCTION

Data Mining (DM) is at best a vaguely defined field; its definition largely depends on the background and views of the definer. This also represents a main characteristic of it: From Pattern Recognition Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data; From Data Base, Data Mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions; From machine learning, Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data. There are many different techniques for data mining. Often which technique you choose to use is determined by the type of data you have and the type of information you are trying to determine from the data. The most popular data mining methods in current use are classification, clustering, neural networks, association, sequence-based analysis, estimation, and visualization. Demonstrating that statistics, like data mining, is concerned with turning data into information and knowledge, even though the terminology may differ, in this section we present a major statistical approach being used in data mining, namely regression analysis. In the late 1990s, statistical methodologies such as regression analysis were not included in commercial data mining packages. Nowadays, most commercial data mining software includes many statistical tools and in particular regression analysis. Although regression analysis may seem simple and anachronistic, it is a very powerful tool in DM with large data sets, especially in the form of the generalized linear models (GLMs). We emphasize the assumptions of the models being used and how the underlying approach differs from that of machine learning.

K.Sathyanarayana Sharma, M.Phil., Research Scholar, Department of Statistics, Salem Sowdeswari College, Salem-10
Dr.S.Rajagopal, Associate Professor , PG & Research Department of Statistics, Salem Sowdeswari College, Salem-10

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as evidence is acquired. Bayesian inference is an important technique in statistics, and especially in mathematical statistics. Bayesian updating is particularly important in the dynamic analysis of a sequence of data.

Bayesian inference has found application in a wide range of activities, including science, engineering, philosophy, medicine, and law. In the philosophy of decision theory, Bayesian inference is closely related to subjective probability, often called "Bayesian probability". The components of Bayesian statistical inference consist of the prior information, the sample data, calculation of the posterior density of the parameters and sometimes calculation of the predictive distribution of future observations.

From the posterior density one may make inferences for β by examining the posterior density. Some prefer to give estimates of β, either point or interval estimates which are computed from the posterior distribution. If β is one-dimensional, a plot of its posterior density tells one the story about β, but if β multidimensional one must be able to isolate those components of β one is interested in.The following diagram is method of Bayesian inference.

Algorithm and Procedure of OLS and Bayesian Linear Regression method:

Ordinary Least Square Method:

Step 1: Arranging the given observation in ascending order.

Step 2: Find the Normal Equation.

Step 3: Solve the Normal Equation and get the unknown parameter value β.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Step 4: Substitute the $\hat{\beta}$ value and form the regression equation.

Step 5: For Prediction, substitute the different values for X and y

Bayesian Linear Regression Method:

Step 1: Find the Joint Posterior from the Prior and Likelihood function.

Step 2: From the Joint Posterior, find the Marginal Posterior distribution.

Step 3: Find the conditional probability of unknown parameter β for the given variable

Step 4: Find the mean of the posterior distribution. Then the mean of the posterior distribution is unbiased for the unknown parameter $\hat{\beta}$.

Step 5: For prediction, substitute the different values of x and y in the Posterior predictive distribution.

Linear Regression Model

We are considering a random variable y as a function of a (typically non-random) vector valued variable $x \in R^k$. This is modeled as a linear relationship, with coefficients β_j and i.i.d Gaussian random noise,

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \dots\dots\dots\dots\dots\dots + x_{ik}\beta_k + \varepsilon_i$$

Where $\varepsilon_i \sim N(0, \sigma^2)$ [∴ Mean vector Zero and covariance matrix vector σ^2]

In Matrix form, this looks like

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & \dots & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & \dots & \dots & x_{2k} \\ \dots & \dots & & & & \dots \\ \dots & \dots & & & & \dots \\ \dots & \dots & & & & \dots \\ x_{n1} & x_{n2} & \dots & \dots & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

Simply,

$$y = X\beta + \varepsilon \quad \dots\dots\dots\dots\dots (1)$$

It is common to set the first column of X to a constant of 1's, so that β_1 is an intercept term. In the equation (1), the β is the unknown parameter. So, we find the unknown parameter value β using the method of "Ordinary Least Square" (OLS).

Ordinary Least Square Estimation

Our aim is to estimate the unknown parameter β. The MLE of β is based on the Gaussian likelihood,

$$P(y/x, \beta : \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \|y - x\beta\|^2\right]$$

This is the product of likelihood for each of the individual components [$y=y_1, y_2, \dots\dots y_n$].

Therefore, we take log of this likelihood,

$$logP(y/x, \beta : \sigma^2) = log\left[\frac{1}{(2\pi\sigma^2)^{n/2}}\right] + log\left[\exp\left[-\frac{1}{2\sigma^2}\|y - x\beta\|^2\right]\right]$$

$$logP(y/x, \beta : \sigma^2) = log1 - log\left[(2\pi\sigma^2)^{n/2}\right] - \frac{1}{2\sigma^2}(y - x\beta)^T(y - x\beta)$$

$$logP(y/x, \beta : \sigma^2) = 0 - {}^{n}/_{2}\, log[2\pi\sigma^2] - \frac{1}{2\sigma^2}[y^T y - y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta]$$

$$logP(y/x, \beta : \sigma^2) = -{}^{n}/_{2}\, log[2\pi\sigma^2] - \frac{1}{2\sigma^2}[y^T y - 2\beta^T X^T Y + \beta^T X^T X\beta] \dots (2)$$

Differentiate w.r.to β in equation (2),

$$\frac{\partial logP}{\partial \beta} = 0$$

$$-0 - \frac{1}{2\sigma^2}[-2X^T y + 2X^T X\beta] = 0$$

$$-\frac{1}{2\sigma^2} 2[-X^T y + X^T X\beta] = 0$$

$$-\frac{1}{\sigma^2}[-X^T y + X^T X\beta] = 0$$

$$-X^T y + X^T X\beta = 0$$

$$X^T X\beta = X^T y$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad \dots\dots\dots (3)$$

Where the inverse here is the Moore-Penrose pseudoinverse (same as the inverse when it exists). The MLE for β is Gaussian distributed and unbiased.

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1}).$$

Bayesian Linear Regression

In statistics, Bayesian linear regression is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference. When the regression model has errors that have a normal distribution, and if a particular form of prior distribution is assumed, explicit results are available for the posterior probability distributions of the model's parameters.

The β can be overinflated for higher-order coefficients, as the model tries to over fit the data with a "wiggly" curve. To counteract this, we may inject our prior belief that these coefficients should not be so large. So, we introduce a conjugate Gaussian prior, $\beta \sim N(0, \Lambda^{-1})$. Here we are parameterizing the Gaussian using the inverse covariance, or precision matrix Λ, which will make computations easier. A common choice is $\Lambda = \lambda I$, for a positive scalar parameter λ.

Conjugate Prior Distribution

For an arbitrary prior distribution, there may be no analytical solution for the posterior distribution. In this section, we will consider a so called conjugate prior for which the posterior distribution can be derived analytically.

$$P\left(\beta/_{\Lambda}\right) = \prod\left(\frac{1}{2\pi\Lambda^{-1}}\right)^{1/2} \exp\left[-\frac{1}{2\Lambda^{-1}}\|\beta\|^2\right]$$

$$P\left(\beta/_{\Lambda}\right) = \prod\left(\frac{1}{2\pi\Lambda^{-1}}\right)^{1/2} \exp\left[-\frac{1}{2}\Lambda\|\beta\|^2\right]$$

$$P\left(\beta/_{\Lambda}\right) = \prod\left(\frac{1}{2\pi\Lambda^{-1}}\right)^{1/2} \exp\left[-\frac{1}{2}\beta^T\Lambda\beta\right] \quad \dots\dots\dots\dots (4)$$

Posterior Distribution

Then, the Posterior for β is,

$$P(\beta/y : X, \sigma^2) = \frac{P(y/X, \beta : \sigma^2)\, P(\beta/\Lambda)}{P(y/\Lambda, \sigma^2)}$$

$$P(\beta/y : X, \sigma^2) \;\alpha\; \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\|y - x\beta\|^2\right] \prod\left(\frac{1}{2\pi\Lambda^{-1}}\right)^{1/2} \exp\left[-\frac{1}{2}\beta^T\Lambda\beta\right]$$

$$P(\beta/y : X, \sigma^2) \;\alpha\; \exp\left[-\frac{1}{2\sigma^2}\|y - x\beta\|^2\right] \exp\left[-\frac{1}{2}\beta^T\Lambda\beta\right]$$

$$P(\beta/y : X, \sigma^2) \;\alpha\; \exp\left[-\frac{1}{2\sigma^2}\|y - x\beta\|^2 - \frac{1}{2}\beta^T\Lambda\beta\right] \dots\dots\dots\dots (5)$$

Just as we worked out in the univariate case, the conjugate prior for β results in the posterior also being a multivariate Gaussian. Completing the square inside the exponent, we see that the posterior for β has the following distribution.

$$\beta \sim N(\mu_n, \Sigma_n)$$

Where, then the Estimate of $\beta = \hat{\beta}$ = Mean of the Posterior distribution (μ_n)

$$\mu_n = (X^T X + \sigma^2 \Lambda)^{-1}$$
$$\Sigma_n = \sigma^2 (X^T X + \sigma^2 \Lambda)^{-1}$$

Posterior Predictive Distribution

In statistics, and especially Bayesian statistics, the posterior predictive distribution is the distribution of unobserved observations (prediction) conditional on the observed data. Described as the distribution that a new i.i.d. data point would have, given a set of N existing i.i.d. observations. In a frequentist context, this might be derived by computing the maximum likelihood estimate (or some other estimate) of the parameter(s) given the observed data, and then plugging them into the distribution function of the new observations. However, the concept of posterior predictive distribution is normally used in a Bayesian context, where it makes use of the entire posterior distribution of the parameter(s) given the observed data to yield a probability distribution over an interval rather than simply a point estimate. Specifically, it is computed by marginalizing over the parameters, using the posterior distribution.

Now, say we are given a new independent data point $\tilde{x}$, and we would like to predict the corresponding unseen dependent value, $\tilde{y}$. The posterior predictive distribution of $\tilde{y}$ is given by,

$$P(\tilde{y}/y;\, \tilde{x}, x, \sigma^2, \Lambda) = \int P(\tilde{y}/\beta;\, \tilde{x}, \sigma^2)\; P(\beta/y; x, \sigma^2, \Lambda)\; d\beta$$

This is now a univariate Gaussian;

$$\tilde{y}/y \sim N\left(\tilde{x}^T \mu_n,\, \sigma_n^2(\tilde{x})\right)$$
$$\sigma_n^2(\tilde{x}) = \sigma^2 + \tilde{x}^T \Sigma_n \tilde{x}$$

The first term on the right is due to the noise (the additive $\in$), and the second term is due to the posterior variance of β, which represents our uncertainty in the parameters.

Interpretation:

Bayesian Statistics takes into account prior information in sample set up. It can update information through the Bayes formula to modify according to the latest results and also having less error. So we conclude that the Bayesian Linear Regression is most appropriate method for prediction.

Demonstrating that statistics, like data mining is concerned with turning data into information and knowledge, even though the terminology may differ, in this project we present a major statistical approach being used in Data Mining, namely Bayesian Linear Regression. In Practical, the Bayesian Linear Regression is most advantage method for calculating prediction in Data Mining.

REFERENCE

[1] A.K.Bansal ; Bayesian Parametric Inference, Alpha Science International Ltd.
[2] Box G.E.P; Tiao, G.C. (1973) : Bayesian Inference in Statistical Analysis.
[3] Carlin , Bradley P. and Louis, Thomas . A (2008) ; Bayesian Methods for Data Analysis
[4] Dr.S.Asif Alisha & T. Srinivas (2014) ; Multiple Linear Regression in Data Mining.
[5] Douglas . C. Montgomery ; Introduction to linear regression analysis, John wiley publication, New York.
[6] Dhaval Sanghavi, Hitarth Patel and Sindhu Nair (2014) ; Logistic Regression in Data Mining and its application in identification of Discase.