UNIT – 2

PARAMETRIC MODEL AND MULTIVARIATE REGRESSION

In machine learning, a parametric model is any model that captures all the information about its predictions within a finite set of parameters. Sometimes the model must be trained to select its parameters, as in the case of neural networks. Sometimes the parameters are selected by hand or through a simple calculation process. Given the multiple definitions of the word "model," a parametric model can output either a probability or a value (in some cases a classification).

The vast majority of machine learning models one deals with on a practical basis are parametric, because relying on non-parametric models generally adds an assumption of too much simplicity in the underlying data.

Examples of parametric models

The following are five different examples of parametric models: exponential distributions, poisson distributions, normal distributions, the Weibull distribution, and linear regressions.

An exponential distribution is given by the function

$$P_\lambda(j) = \frac{\lambda^j}{j!} e^{-\lambda}$$

Exponential distributions are useful in the modeling of the time to failure of equipment, where the rate of failure is constant and the mean time to failure is the inverse of λ.

A poisson distribution is given by the function

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

The poisson distribution models the probability of an event occurring in a fixed amount of time or space when it is known to occur at a constant fixed rate λ (the parameter), and each event is independent.

The normal distribution is an example of a model with multiple parameters. It is given by

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

The parameter μ represents the mean of the data while the parameter σ represents the standard deviation of the data. σ is an example of a parameter with a restricted value range – it can only be positive. The normal distribution can be thought of as the simplest or smoothest distribution with a given mean and variance.

The two-parameter Weibull distribution is a parametric model useful in the modeling of the probability distribution of time to failure for equipment, and is a generalization of the exponential distribution. The distribution is given by

$$P(x) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} e^{-(\frac{x}{\lambda})^k}$$

for $x \geq 0$ and $P(x)=0$ for $x<0$. λ, the scale parameter, and k, the shape parameter are both restricted to be a positive.

Finally, we have linear regression. Linear regression models are given by

$$y = X\beta + \alpha,$$

where X is a matrix and the parameters β and α are vectors. β and α are chosen so as to minimize the sum of the squares of the residuals from the linear regression line to the actual input data points. The assumption is that the statistics of the residuals are such that they are normally distributed around the linear regression line.

Multivariate Regression is a supervised machine learning algorithm involving multiple data variables for analysis. Multivariate regression is an extension of multiple regression with one dependent variable and multiple independent variables. Based on the number of independent variables, we try to predict the output.

Multivariate regression tries to find out a formula that can explain how factors in variables respond simultaneously to changes in others.

There are numerous areas where multivariate regression can be used. Let's look at some examples to understand multivariate regression better.

1. Praneeta wants to estimate the price of a house. She will collect details such as the location of the house, number of bedrooms, size in square feet, amenities available, or not. Basis these details price of the house can be predicted and how each variables are interrelated.
2. An agriculture scientist wants to predict the total crop yield expected for the summer. He collected details of the expected amount of rainfall, fertilizers to be used, and soil conditions. By building a Multivariate regression model scientists can predict his crop yield. With the crop yield, the scientist also tries to understand the relationship among the variables.
3. If an organization wants to know how much it has to pay to a new hire, they will take into account many details such as education level, number of experience, job location, has niche skill or not. Basis this information salary of an employee can be predicted, how these variables help in estimating the salary.

4. Economists can use Multivariate regression to predict the GDP growth of a state or a country based on parameters like total amount spent by consumers, import expenditure, total gains from exports, total savings, etc.
5. A company wants to predict the electricity bill of an apartment, the details needed here are the number of flats, the number of appliances in usage, the number of people at home, etc. With the help of these variables, the electricity bill can be predicted.

The above example uses Multivariate regression, where we have many independent variables and a single dependent variable.

Mathematical equation

The simple regression linear model represents a straight line meaning y is a function of x. When we have an extra dimension (z), the straight line becomes a plane.

Here, the plane is the function that expresses y as a function of x and z. The linear regression equation can now be expressed as:

y = m1.x + m2.z+ c

y is the dependent variable, that is, the variable that needs to be predicted.
x is the first independent variable. It is the first input.

m1 is the slope of x1. It lets us know the angle of the line (x).
z is the second independent variable. It is the second input.
m2 is the slope of z. It helps us to know the angle of the line (z).
c is the intercept. A constant that finds the value of y when x and z are 0.

The equation for a model with two input variables can be written as:

y = β0 + β1.x1 + β2.x2

What if there are three variables as inputs? Human visualizations can be only three dimensions. In the machine learning world, there can be n number of dimensions. The equation for a model with three input variables can be written as:

y = β0 + β1.x1 + β2.x2 + β3.x3

Below is the generalized equation for the multivariate regression model-

y = β0 + β1.x1 + β2.x2 +….. + βn.xn

Where n represents the number of independent variables, β0~ βn represents the coefficients, and x1~xn is the independent variable.

The multivariate model helps us in understanding and comparing coefficients across the output. Here, the small cost function makes Multivariate linear regression a better model.

What is Cost Function?

The cost function is a function that allows a cost to samples when the model differs from observed data. This equation is the sum of the square of the difference between the predicted value and the actual value divided by twice the length of the dataset. A smaller mean squared error implies better performance. Here, the cost is the sum of squared errors.

Cost of Multiple Linear regression:

$$MSE=\frac{1}{2m}\Sigma\left(h_\theta(x)^{(i)}-y^i\right)^2$$

Steps of Multivariate Regression analysis

Steps involved for Multivariate regression analysis are feature selection and feature engineering, normalizing the features, selecting the loss function and hypothesis, setting hypothesis parameters, minimizing the loss function, testing the hypothesis, and generating the regression model.

- **Featureselection-**
  The selection of features is an important step in multivariate regression. Feature selection also known as variable selection. It becomes important for us to pick significant variables for better model building.
- **NormalizingFeatures-**
  We need to scale the features as it maintains general distribution and ratios in data. This will lead to an efficient analysis. The value of each feature can also be changed.
- **SelectLossfunctionandHypothesis-**
  The loss function predicts whenever there is an error. Meaning, when the hypothesis

prediction deviates from actual values. Here, the hypothesis is the predicted value from the feature/variable.

- **SetHypothesisParameters-**
  The hypothesis parameter needs to be set in such a way that it reduces the loss function and predicts well.
- **MinimizetheLossFunction-**
  The loss function needs to be minimized by using a loss minimization algorithm on the dataset, which will help in adjusting hypothesis parameters. After the loss is minimized, it can be used for further action. Gradient descent is one of the algorithms commonly used for loss minimization.
- **Testthehypothesisfunction-**
  The hypothesis function needs to be checked on as well, as it is predicting values. Once this is done, it has to be tested on test data.

Advantages of Multivariate Regression

The most important advantage of Multivariate regression is it helps us to understand the relationships among variables present in the dataset. This will further help in understanding the correlation between dependent and independent variables. Multivariate linear regression is a widely used machine learning algorithm.

Disadvantages of Multivariate Regression

- Multivariate techniques are a bit complex and require a high-levels of mathematical calculation.
- The multivariate regression model's output is not easy to interpret sometimes, because it has some loss and error output which are not identical.
- This model does not have much scope for smaller datasets. Hence, the same cannot be applied to them. The results are better for larger datasets.

Conclusion

Multivariate regression comes into the picture when we have more than one independent variable, and simple linear regression does not work. Real-world data involves multiple variables or features and when these are present in data, we would require Multivariate regression for better analysis.

Reference:

https://www.mygreatlearning.com/blog/introduction-to-multivariate-regression/#:~:text=driven%20decision%20maker!-,What%20is%20Multivariate%20Regression%3F,variable%20and%20multiple%20independent%20variables.