Bike Sharing
Case Study

General Subjective Questions

Topic: Linear Regression
By: Ramesh Velivela

Batch: ML61 EPGP in AI & ML

General Subjective Questions

1. Explain the linear regression algorithm in detail.
2. Explain the Anscombe's quartet in detail.
3. What is Pearson's R?
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

1. Explain the linear regression algorithm in detail.

- Linear Regression is an algorithm that belongs to supervised Machine Learning. It tries to apply relations that will predict the outcome of an event based on the independent variable data points. The relation is usually a straight line that best fits the different data points as close as possible. The output is of a continuous form, i.e., numerical value. For example, the output could be revenue or sales in currency, the number of products sold, etc. In the above example, the independent variable can be single or multiple.
- Linear regression can be expressed mathematically as: $y= \beta 0+ \beta 1x+ \varepsilon$

Here, Y= Dependent Variable; X= Independent Variable; β 0= intercept of the line; β1 = Linear regression coefficient (slope of the line); ε = random error

Note: The last parameter, random error ε, is required as the best fit line also doesn't include the data points perfectly.

- Since the Linear Regression algorithm represents a linear relationship between a dependent (y) and one or more independent (x) variables, it is known as Linear Regression. This means it finds how the value of the dependent variable changes according to the change in the value of the independent variable. The relation between independent and dependent variables is a straight line with a slope.
- Types of Linear Regression: Linear Regression can be broadly classified into two types of algorithms:

1. Simple Linear Regression: A simple straight-line equation involving slope (dy/dx) and intercept (an integer/continuous value) is utilized in simple Linear Regression. Here a simple form is:

   y=mx+c where y denotes the output x is the independent variable, and c is the intercept when x=0. With this equation, the algorithm trains the model of machine learning and gives the most accurate output

2. Multiple Linear Regression: When a number of independent variables more than one, the governing linear equation applicable to regression takes a different form like:

   y= c+m1x1+m2x2… mnxn where represents the coefficient responsible for impact of different independent variables x1, x2 etc. This machine learning algorithm, when applied, finds the values of coefficients m1, m2, etc., and gives the best fitting line.

3. Non-Linear Regression: When the best fitting line is not a straight line but a curve, it is referred to as Non-Linear Regression.

1. Explain the linear regression algorithm in detail. Continued.,

- Cost functions – The cost functions helps to identify the best possible values for the $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches – **Unconstrained and constrained**.
  - Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
    - The straight-line equation is $Y = \beta_0 + \beta_1 X$
    - The prediction line equation would be $Y_{pred} = \beta_0 + \beta_1 x_i$ and the actual Y is as Y_i.
    - $Now\ the\ cost\ function\ will\ be\ J(\beta_1, \beta_0) = \sum(y_i - \beta_1 x_i - \beta_0)^2$
  - The unconstrained minimization are solved using 2 methods
    - Closed form
    - Gradient descent
- While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.
  - $e_i = y_i - y_{pred}$ is provides the error for each of the data point.
  - OLS is used to minimize the total e^2 which is called as Residual sum of squares.
  - $\text{RSS} = = \sum_{i=1}^{n} (y_i - y_{pred})^2$
- Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

1. Explain the linear regression algorithm in detail. Continued.,

Assumptions of Linear Regression: Naturally, if these assumptions are not considered, the results will not be reliable. Linear Regression also comes under same consideration. There are some common assumptions to be considered while using Linear Regression:

1. **Linearity:** The models of Linear Regression models must be linear in the sense that the output must have a linear association with the input values, and it only suits data that has a linear relationship between the two entities.
2. **Homoscedasticity**: Homoscedasticity means the standard deviation and the variance of the residuals (difference of $(y-y\hat{})^2$) must be the same for any value of x. Multiple Linear Regression assumes that the amount of error in the residuals is similar at each point of the linear model. We can check the Homoscedasticity using Scatter plots.
3. **Non-multicollinearity**: The data should not have multicollinearity, which means the independent variables should not be highly correlated with each other. If this occurs, it will be difficult to identify those specific variables which actually contribute to the variance in the dependent variable. We can check the data for this using a correlation matrix.
4. **No Autocorrelation**: When data are obtained across time, we assume that successive values of the disturbance component are momentarily independent in the conventional Linear Regression model. When this assumption is not followed, the situation is referred to be autocorrelation.
5. **Not applicable to Outliers**: The value of the dependent variable cannot be estimated for a value of an independent variable which lies outside the range of values in the sample data.

Advantages of Linear Regression

- For linear datasets, Linear Regression performs well to find the nature of the relationship among different variables.
- Linear Regression algorithms are easy to train and the Linear Regression models are easy to implement.
- Although, the Linear Regression models are likely to over-fit, but can be avoided using dimensionality reduction techniques such as regularization (L1 and L2) and cross-validation.

- **Disadvantages of Linear Regression**
  - An important disadvantage of Linear Regression is that it assumes linearity between the dependent and independent variables, which is rarely represented in real-world data. It assumes a straight-line relationship between the dependent and independent variables, which is unlikely many times.
  - It is prone to noise and overfitting. In datasets where the number of observations is lesser than the attributes, Linear Regression might not be a good choice as it can lead to overfitting. This is because the algorithm can start considering the noise while building the model.
  - Sensitive to outliers, it is essential to pre-process the dataset and remove the outliers before applying Linear Regression to the data.
  - It does not assume multicollinearity. If there is any relationship between the independent variables, i.e., multicollinearity, then it needs to be removed using dimensionality reduction techniques before applying Linear Regression as the algorithm assumes that there is no relationship among independent variables.

- Linear Regression is "Easy to Implement", "Scalable", "Interpretability", "easy to train and do not require much computational power".

2. Explain the Anscombe's quartet in detail.

- Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.
- **ANSCOMBE'S QUARTET FOUR DATASETS**
- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.
- Anscombe's quartet helps us to understand the importance of data visualization. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.
- Anscombe's Quartet signifies that multiple data sets with many similar statistical properties could still be different from one another when plotted.
- The dangers of outliers in data sets are warned by the quartet. Check the bottom 2 graphs. If those outliers would have not been there the descriptive stats would have been completely different in that case.
- Important points:
  - Plotting the data is very important and a good practice before analyzing the data.
  - Outliers should be removed while analyzing the data.
  - Descriptive statistics do not fully depict the data set in its entirety.

3. What is Pearson's R?

- The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearon's R returns values between -1 and 1. The interpretation of the coefficients are:
  - -1 coefficient indicates strong inversely proportional relationship.
  - 0 coefficient indicates no relationship.
  - 1 coefficient indicates strong proportional relationship.

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

- *N = the number of pairs of scores*
- *Σxy = the sum of the products of paired scores*
- *Σx = the sum of x scores*
- *Σy = the sum of y scores*
- *Σx^2 = the sum of squared x scores*
- *Σy^2 = the sum of squared y scores*

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- **What is scaling**: The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.
- **Why Scaling performed**: Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results in to the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance.
  The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.
- **Difference between Normalization/Min-Max scaling and Standardized scaling**:
  - The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

  - Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- $VIF = \frac{1}{1-R^2}$
- The VIF formula clearly signifies when the VIF will be infinite. **If the R^2 is 1 then the VIF is infinite**.
- The reason for **R^2 to be 1 is that there is a perfect correlation between 2 independent variables**.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- Q-Q plots are the **quantile-quantile plots**.
- It is a **graphical tool to assess the 2 data sets are from common distribution**. The theoretical distributions could be of type normal, exponential or uniform. **The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions**. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below
- **Interpretations**
  - Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
  - Y values < X values: If y-values quantiles are lower than x-values quantiles.
  - X values < Y values: If x-values quantiles are lower than y-values quantiles.
  - Different distributions: If all the data points are lying away from the straight line.

Thanks