Machine Learning 1 – Module 2 – Bike Sharing Assignment
by Sankalp Gupta

Assignment based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: There are 7 categorical variables: season, yr, holiday, weekday, workingday, weathersit and mnth.

(i) Season has a very high effect on the target/dependent variable. Max rentals are in fall, followed by summer. Rentals drop in winter and are minimum in spring

(ii) Mnth has a very similar effect to season. Max rentals are between Jun to Sep, which also corresponds with Fall. Similarly minimum rentals are between Dec and Mar, when the season is spring.

(iii) Weathersit also has a big effect, and also follows season and mnth. Clear days are good for rentals. Light snow is bad. There are no rentals during heavy snow.

(iv) There are fewer rentals on holidays and there are higher rentals on workingdays

(v) There are fewer rentals on weekends compared to higher rentals during working weekdays

(vi) yr – There is a significant increase in rentals in the 2nd year compared to the 1st year, which can be clearly understood as 1st year was the introduction of business. In fact registered users are markedly high in the 2nd year, which indicates business is booming

2. Why is it important to use drop_first=True during dummy variable creation?

Ans: drop_first=True is required to remove redundancy in dummy variables. Its important to remove redundancy, as otherwise the linear regression model will show infinite VIF between redundant variables, as they are 100% correlated.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: From the pair-plots of numerical variables, it is observed that the variables *temp* and *atemp* have the highest correlation with the target variable, *cnt*.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: There are 3 assumptions tested in this linear regression model:

(i) Linear Relationship: The relationship between predictor and target variables is linear. This is ascertained from an observation of the scatter-plot of the most significant variable with respect to the target variable

(ii) Errors are normally distributed: The second assumption is that all errors are normally distributed. Plotting a distribution plot of residuals confirms that this assumption is valid too

(iii) Errors are independent: By a visual inspection of the scatter plot of the residuals against the record index, it is ascertained that the mean of errors is zero.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: The top 3 features contributing significantly are: temperature (temp), year (yr) and working day (workingday).

Higher the temperature, higher are the chances of rentals

Here, the meaning of yr should be considered as not-the-first-year. It means, that for the next year, the value of yr should not be 2, but continue to remain as 1, at least for the purpose of calculating the projection of the target variable. Rentals are high for not-first-year.

If a day is a working day, rentals are higher, compared to a non-working-day

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: The linear regression model is based on the following formula:

$y = a_0 + a_1 * x_1 + a_2 * x_2 + a_3 * x_3 + ...$

where:

y is the target /dependent / predicted variable

x_is are one or more predictor / independent variables

a_is are the coefficients of the corresponding predictor variables. It means, for example, for x_2, that if all other x_is are kept constant, then y will change by a_2 times x_2.

a_0 is a constant, which indicates that even if all x_is are = 0, y is still not zero, but is equal to a_0.

The above model assumes that there is a linear relationship between the predictor and independent variables.

It also assumes that all the independent variables are independent of each other.

The objective of the linear regression algorithm is to find the best fit, that is the best set of values of a_is. The algorithm uses the least-squares method to find the best fit.

The concept of least-squares, is that the sum of the distances between y-predicted and y-actual of all points, should be minimum for the right fit.

(i) For any single point:

(ii) Error e = y-actual – y-predicted

= y-actual – $a_0 + a_1 * x_1 + a_2 * x_2 + a_3 * x_3 + ...$

(iii) Sum = Sum of square of all error terms (also called the Cost function

(iv) For different choice of the set of a_is, the above sum / cost function will be different. It will be minimum for a certain set, and we need to find that set. In order to do so, we use the concept of finding a minima using calculus. Thus we differentiate the above sum with respect to a_is.

(v) By doing partial differentiation with respect to each a_i, one at a time, we will arrive at a set of simultaneous equations, which when solved (using matrix algebra), will give the best combination of values of a_is.

(vi) The above process of solving a matrix can be time and resources consuming when the number of features become large. Sometimes it may not even yield any result, if the matrix turns out to be singular. In such a case, it is best to use an iterative approach.

Instead of solving analytically, it is better to solve using an iterative process. Gradient Descent is the process that is used for minimising the Cost function.

Here are the steps involved in the algorithm:

(i) Assume a set of random initial values for a_is.

(ii) Calculate the Cost function: Sum of squares of (y-actual - y-predicted) for each data-point

(iv) Calculate the gradient.

(v) Adjust the a_is a little, by an amount called the learning rate

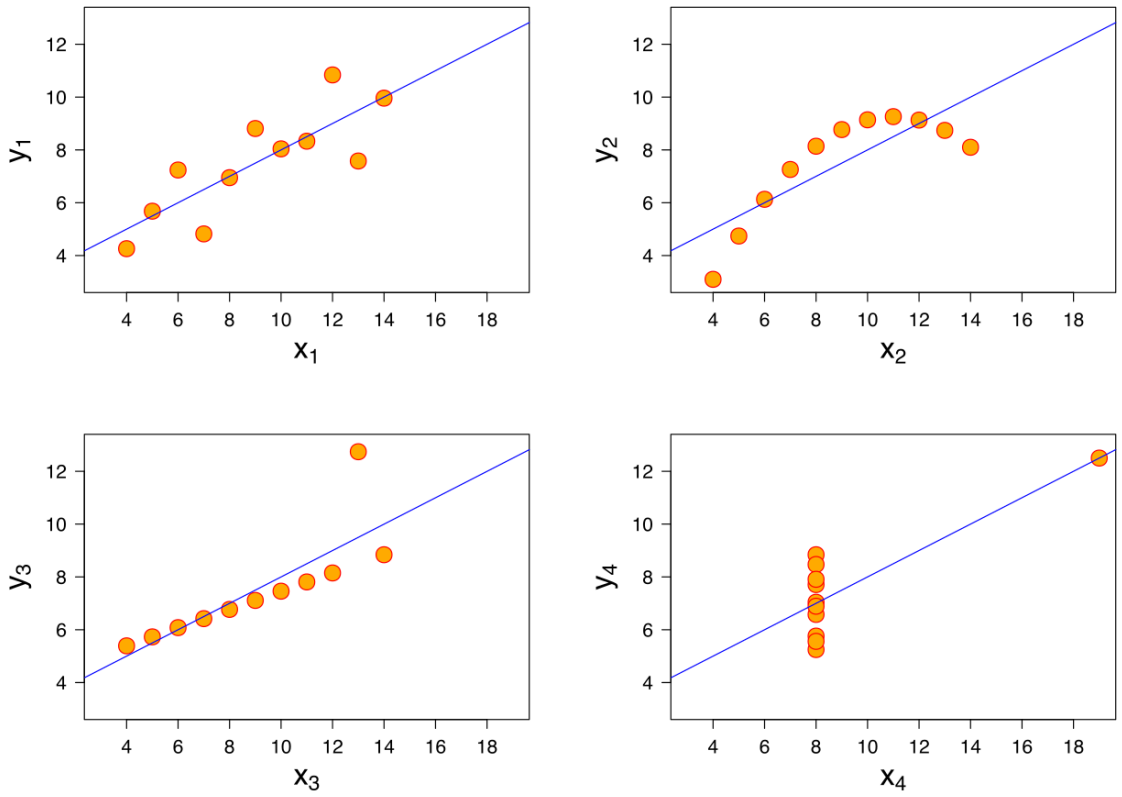
(v) Perform steps (ii) and (v) again, till the cost function is minimised.

(ix) Note the final value of a_is

2. Explain the Anscombe's quartet in detail.

Ans: In 1973, the French mathematician Francis Anscombe came up with 4 data sets of x-y pairs that had almost the same statistical properties like mean (for both x and y), standard-deviation (for both x and y), x-y correlation, linear regression line and R^2.

However, the four datasets, plotted on a graph, showed a marked difference among each other. Shown below are the four datasets.



The first dataset-x1,y1 showed a simple linear relationship, but the second-x2,y2 is a convex curve. The third-x3,y3 had a single outlier that changed the linear relationship. In the fourth-x4,y4 a single outlier, changed a constant value of x to a varying relationship.

With this quartet, Anscombe demonstrated that visualization is a very important tool in data understanding.

3. What is Pearson's R?

Ans: Pearson's R, also called the correlation coefficient, is a measure of the linear correlation between two variables.

Mathematically, it is defined as the ratio of the covariance of two variables and the product of their standard deviations.

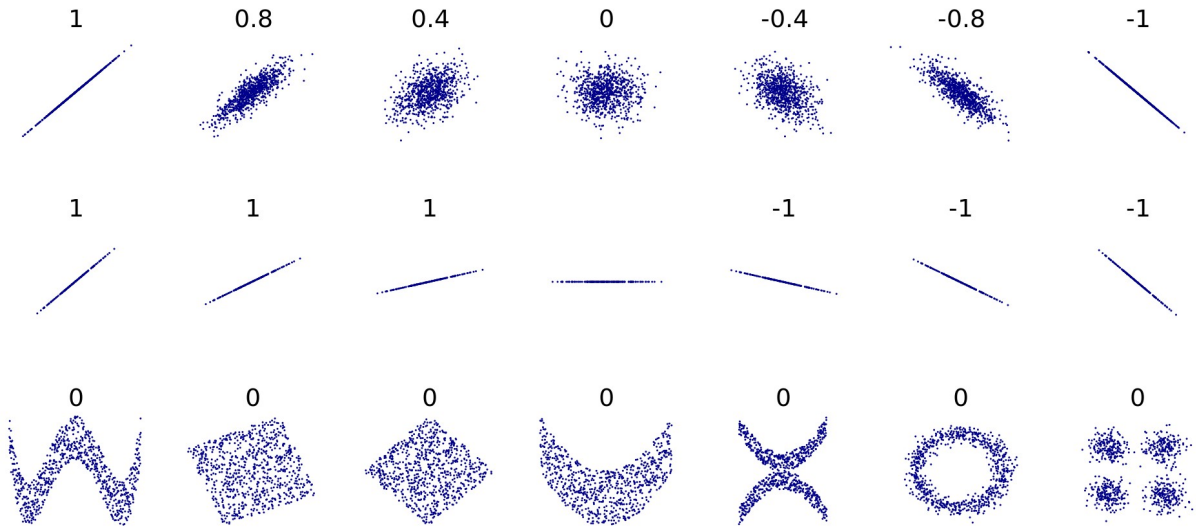
Its value ranges from -1 to 1. Higher the absolute value of Pearson's R, higher is the correlation between the two variables.

A positive value indicates a proportional correlation, i.e, an increase in x, will mean an increase in y. Whereas, a negative value indicates an inverse

correlation, meaning, an increase in x will mean a decrease in y.

A value of 0, indicates there is no correlation between x and y.

The diagram below illustrates the meaning of the various values of Pearson's R.



As noticed in the above diagram, a value of 0 indicates no **linear** relationship, even though there might be other complex relationships.

A value of 1 (whether negative or positive) indicates a perfect linear relationship. Even if the slope is different, the relationship is still 1.

In simple terms, the correlation can also be viewed as the spread of the data. Higher the correlation, less the spread.

In the middle of the figure is a horizontal line, which has no value. This is because here the value of y is a constant irrespective of the value of x. So although it is a perfect straight line, there is no correlation between x and y.

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

Ans: Scaling is the process of re-sizing a data-set, such that when two variables are compared, then their range of values is similar to each other.

It is a process similar to converting numbers to percentages.

Scaling helps in:

(i) Easier comparison of two variables, both numerically and graphically

(ii) Retaining the significance of all the variables in an equation. Else, sometimes, a variable with a value range orders-of-magnitude lower than another, can be completely overshadowed, and may even be lost due to numerical precision errors

(iii) Easier calculations, as for instance in using z-tables for normal distribution.

(iv) Faster calculations: When all the variables are in similar range, and the results are expected after an iterative process, then calculations converge faster.

The two types of scaling: normalized and standardized are defined as follows:

Normalized scaling: for any x in a data-set: $x' = (x - x^{min}) / (x^{max} - x^{min})$

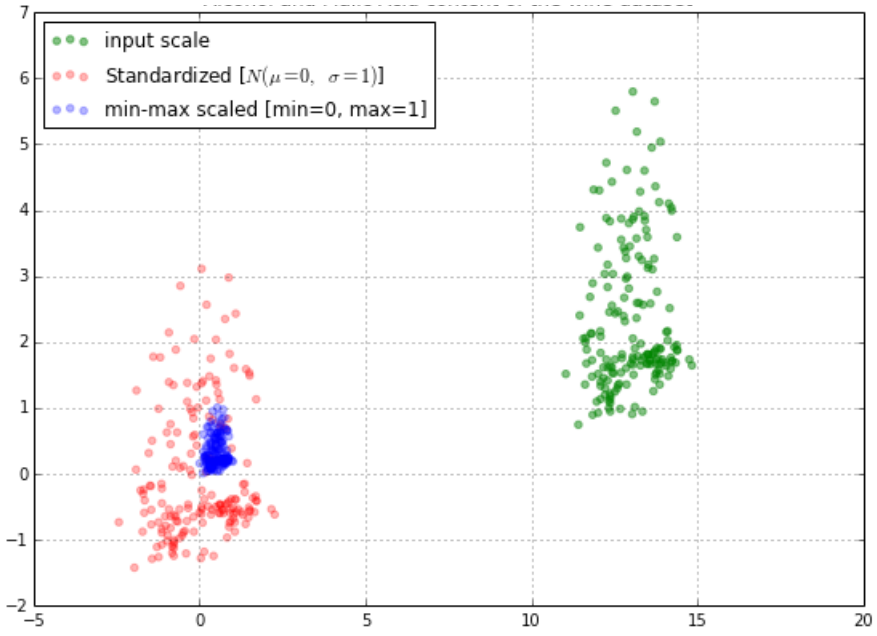
Standardized scaling: for any x in a data-set: $x' = (x - x^{mean}) / \sigma$

where σ is the standard deviation

Normalized scaling converts all the values in a data-set to a range between 0 to 1, and

Standardized converts it to a range of numbers 95% of the data points are roughly between -3 to 3, the mean is 0, and standard deviation is equal to 1.

The diagram below illustrates both types of scaling:



The original data is in green colour on the right hand side. Scaled data is in red and blue on the left hand side. Red coloured data is standardized scale and blue is normalized scale.

You will notice that irrespective of the scale, the spread (shape of the scatter image) of the data is similar in all 3 cases.

You will also notice that in the normalized scale, the date is within 0 and 1, and in normalized scale, most of it is within -3 to +3.

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

Ans: This means that the corresponding features are 100% correlated, i.e, their correlation coefficient is exactly equal to +1 or -1.

In such a situation, we need to drop one of the two variables, in order for the model to work properly.

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

Ans: Q-Q plot, or a quantile-quantile plot is a great visual tool to compare actual values against predicted values of a target variable. Its a 2D plot where we expect the plot of actuals vs predicted values to be a 45° line, since we expect the predicted values to match the actual values.

It can also be used to determine the kind of distribution a given dataset possesses. Whether its a normal distribution or a uniform distribution, etc.

In this case, we used the Q-Q plot to check the following:

1. Whether the residuals follow normal distribution

2. Compare the actual vs predicted values of the target variable