**MACHINE LEARNING – WORKSHEET 11 (LINEAR REGRESSION) ANSWERSHEET**

**In Q1 to Q8, only one option is correct, Choose the correct option:**

1. A) remains same
2. B) SST = SSR + SSE
3. C) difference between the actual value and mean of dependent variable.
4. C) By its slope
5. C) can be any value between -1 to 1
6. A) Scatter plot
7. B) f-statistics
8. C) Ridge

**In Q9 to Q11, more than one options are correct, Choose all the correct options:**

1. A) It shows the causal relationship between dependent and independent variables

D) It is a straight line that is the best approximation of the given data sets

1. A) Reducing the training time B) Generalizing the test set
2. A) Normal Equation D) nodes

**Q12 to Q15 are subjective answer type questions, Answer them briefly.**

1. **R-squared** is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

**R-squared = Explained variation / Total variation**

R-squared is always between 0 and 100%:

0% indicates that the model explains none of the variability of the response data around its mean.

100% indicates that the model explains all the variability of the response data around its mean.

**The adjusted R-squared** is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it’s usually not. It is always lower than the R-squared.

1. **A cost function** is a measure of how wrong the model is in terms of its ability to estimate the relationship between X and y. This is typically expressed as a difference or distance between the predicted value and the actual value. The cost function (you may also see this referred to as loss or error.) can be estimated by iteratively running the model to compare estimated predictions against “ground truth” — the known values of y.
2. **Total sum of squares (SST)** is the sum of squared deviations of individual measurements from the mean. The total sum of squares is a sum of 2 portions:

(1**) Regression sum of squares (SSR**) which is the contribution of factors into the variance of the dependent variable, and

(2) **Error sum of squares (=residual sum of squares) (SSE)** which is the stochastic component of the variation of the dependent variable.

SSR is the sum of squared deviations of predicted values (predicted using regression) from the mean value, and SSE is the sum of squared deviations of actual values from predicted values.

The significance of regression is evaluated using F-statistics:

where

df(SSR)= g - 1 is the number of degrees of freedom for the regression sum of squares which is equal to the number of coefficients in the equation, g, minus 1;

df(SSE)= N - g is the number of degrees of freedom for the error sum of squares which is equal to the number of observations, N, minus the number of coefficients, g;

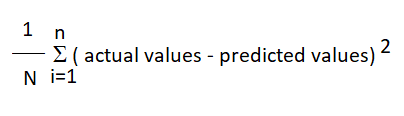
df(SST) = df(SSR) + df(SSE) = N - 1 is the number of degrees of freedom for the total sum of squares.

1. **1. Mean Absolute Error or MAE-**We know that an error basically is the absolute difference between the actual or true values and the values that are predicted. Absolute difference means that if the result has a negative sign, it is ignored.

Hence, MAE = True values – Predicted values

MAE takes the average of this error from every sample in a dataset and gives the output.

**2. Mean Squared Error or MSE-**MSE is calculated by taking the average of the square of the difference between the original and predicted values of the data.



Hence, MSE =

Here N is the total number of observations/rows in the dataset. The sigma symbol denotes that the difference between actual and predicted values taken on every i value ranging from 1 to n.

**3. Root Mean Squared Error or RMSE-**RMSE is the standard deviation of the errors which occur when a prediction is made on a dataset. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model.

**4. R Squared-**It is also known as the coefficient of determination. This metric gives an indication of how good a model fits a given dataset. It indicates how close the regression line (i.e the predicted values plotted) is to the actual data values. The R squared value lies between 0 and 1 where 0 indicates that this model doesn't fit the given data and 1 indicates that the model fits perfectly to the dataset provided.