ML Worsksheet 6

Q.1) A

Q.2) A

Q.3) C

Q.4) A

Q.5) B

Q.6) A , D

Q.7) A, C and D

Q.8) D

Q.9) A, B

Q. 10) Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model ?

Answer : 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.
If a model has too many predictors and higher order polynomials, it begins to model the random noise in the data. This condition is known as overfitting the model and it produces misleadingly high R-squared values and a lessens ability to make predictions.

Q. 11) Differentiate between Ridge and Lasso Regression ?

Answer : Lasso is a modification of linear regression, where the model is penalized for the sum of absolute values of the weights. Thus, the absolute values of weight will be (in general) reduced, and many will tend to be zeros while Ridge takes a step further and penalizes the model for the sum of squared value of the weights. Thus, the weights not only tend to have smaller absolute values, but also really tend to penalize the extremes of the weights, resulting in a group of weights that are more evenly distributed.

Q. 12) What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling ?

Answer : A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.
Most research papers consider a VIF (Variance Inflation Factor) > 10 as an indicator of multicollinearity, but some choose a more conservative threshold of 5 or even 2.5. When VIF is higher than 10 or tolerance is lower than 0.1, there is significant multicollinearity that needs to be corrected.

Q. 13) Why do we need to scale the data before feeding it to the train the model?

Answer : Scaling the target value is a good idea in regression modelling ; scaling of the data makes it easy for a model to learn and understand the problem. In the case of neural networks, an independent variable with a spread of values may result in a large loss in training and testing and cause the learning process to be unstable.

Q. 14) What are the different metrics which are used to check the goodness of fit in linear regression?

Answer :
1) Mean Absolute Error (MAE) : MAE is a very simple metric which calculates the absolute difference between actual and predicted values.
2) Mean Squared Error(MSE) : MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.
3) Root Mean Squared Error(RMSE) : As RMSE is clear by the name itself, that it is a simple square root of mean squared error.
4) Root Mean Squared Log Error(RMSLE) : Taking the log of the RMSE metric slows down the scale of error. The metric is very helpful when you are developing a model without calling the inputs. In that case, the output will vary on a large scale.
5) R Squared (R2) : R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.
6) Adjusted R Squared : The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because It assumes that while adding more data variance of data increases.

Q. 15) From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.
Answer :
1) Recall / Sensitivity = TP / (TP + FN)
= 1000/(1000+50)
= 1000/1050
= 0.95

2) Precision = TP / (TP + FP)
= 1000/(1000+250)
= 1000/1250

= 0.80

3) Specificity = TN / (TN + FP)
= 1200/(1200+250)
= 1200/1450
= 0.82

4) Accuracy = (TP + TN)/ (TP + TN+FP+FN)
= (1000+1200)/(1000+1200+250+50)
= (2200)/(2500)
= 0.88