MACHINE LEARNING ASSIGNMENT – 6

In Q1 to Q5, only one option is correct, Choose the correct option:

1. In which of the following you can say that the model is overfitting?

Ans=B) Low R-squared value for train-set and High R-squared value for test-set.

1. Which among the following is a disadvantage of decision trees?

Ans=A) Decision trees are prone to outliers.

1. Which of the following is an ensemble technique?

Ans=B) Logistic Regression

1. Suppose you are building a classification model for detection of a fatal disease where detection of the disease is most important. In this case which of the following metrics you would focus on?

Ans=A) Accuracy

1. The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and of model B is 0.85. Which of these two models is doing better job in classification?

Ans=B) Model B

In Q6 to Q9, more than one options are correct, Choose all the correct options:

1. Which of the following are the regularization technique in Linear Regression??

Ans=A) Ridge D) Lasso

1. Which of the following is not an example of boosting technique?

Ans=B) Decision Tree C) Random Forest

1. Which of the techniques are used for regularization of Decision Trees?

Ans=B) L2 regularization C) Restricting the max depth of the tree

1. Which of the following statements is true regarding the Adaboost technique?

Ans=B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

C) It is example of bagging technique

Q10 to Q15 are subjective answer type questions, Answer them briefly.

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

Ans=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.

1. Differentiate between Ridge and Lasso Regression.

Ans=Lasso regression stands for Least Absolute Shrinkage and Selection Operator. It adds penalty term to the cost function.The difference between ridge and lasso regression is that it tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.

There are three popular regularization techniques, each of them aiming at decreasing the size of the coefficients: Ridge Regression, which penalizes sum of squared coefficients (L2 penalty). Lasso Regression, which penalizes the sum of absolute values of the coefficients (L1 penalty).

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

Ans=In general, a VIF above 10 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above. Sometimes a high VIF is no cause for concern at all. For example, you can get a high VIF by including products or powers from other variables in your regression, like x and x2.

1. Why do we need to scale the data before feeding it to the train the model?

Ans=To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

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

Ans=Five metrics give us some hints about the goodness-of-fit of our model. The first two metrics, the Mean Absolute Error and the Root Mean Squared Error (also called Standard Error of the Regression), have the same unit as the original data.

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit.

1. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.

Ans=Precision = TP / (TP+FP)=1000/(1000+250)=0.8

Recall = TP / (TP+FN)=1000/(1000+50)=0.95

Sensitivity=TP(TP+FN)=1000(1000+50)=1050000

Specificity=TN(TN+FP)=1200(1200+250)=1262400