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?

C) High R-squared value for train-set and Low R-squared value for test-set.

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

B) Decision trees are highly prone to overfitting.

3. Which of the following is an ensemble technique?

C) Random Forest

4. 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?

B) Sensitivity

5. 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?

B) Model B

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

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

A) Ridge

D) Lasso

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

C) Random Forest

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

A) Pruning

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

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.

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

ANS: The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

11. Differentiate between Ridge and Lasso Regression.

ANS: **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. During training, the objective function become:

$$\frac{1}{2m}\sum_{i=1}^{m}(y-Xw)^2+alpha\sum_{j=1}^{p}\left|w_j\right|$$

As you see, Lasso introduced a new hyperparameter, *alpha*, the coefficient to penalize weights.

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. The objective function becomes:

$$\sum_{i=1}^{n}(y - Xw)^2 + alpha\sum_{j=1}^{p} w_j{}^2$$

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

modelling?

ANS: The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity. Generally, a VIF above 4 or tolerance below 0.25 indicates that multicollinearity might exist, and further investigation is required. When VIF is higher than 10 or tolerance is lower than 0.1, there is significant multicollinearity that needs to be corrected.

13. 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.

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

ANS: There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are: Mean Squared Error (MSE). Root Mean Squared Error (RMSE). Mean Absolute Error (MAE)

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

Actual/Predicted True False

True 1000 50

False 250 1200

ANS:

- Sensitivity= 0.8000
- Specificity= 0.9600
- Precision = 0.9524
- Recall = 0.8000
- Accuracy = 0.8800