Worksheet-Set 6: Machine learning Solutions

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

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?**

   B) Decision Tree

   C) Random Forest

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

   A) Pruning

   C) Restricting the max depth of the tree

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

   A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points

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

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

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.

11. **Differentiate between Ridge and Lasso Regression.**

Ridge and Lasso regression uses two different penalty functions. Ridge uses l2 whereas lasso goes with l1. In ridge regression, the penalty is the sum of the squares of the coefficients and for the Lasso, it's the sum of the absolute values of the coefficients. It's a shrinkage towards zero using an absolute value (l1 penalty) rather than a sum of squares (l2 penalty).

As we know, ridge regression can't have zero coefficients. Here, you either select all the coefficients or none of them whereas LASSO does both parameter shrinkage and variable selection automatically because it zero out the coefficients of collinear variables. Here it helps to select the variable(s) out of given n variables while performing lasso regression.

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

Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity. Variance inflation factors (VIF) measure how much the variance of the estimated regression coefficients is inflated as compared to when the predictor variables are not linearly related.

It is obtained by regressing each independent variable, say X on the remaining independent variables (say Y and Z) and checking how much of it (of X) is explained by these variables.

$$VIF = \frac{1}{(1-R^2)}$$

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

In many machine learning algorithms, to bring all features in the same position, we need to do scaling so that one significant number doesn't impact the model just because of their large magnitude.

Feature scaling in machine learning is one of the most critical steps during the preprocessing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

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

The various metrics used to evaluate the results of the prediction are:
Mean Squared Error (MSE)
Root-Mean-Squared-Error (RMSE). Mean-Absolute-Error
(MAE).
R^2 or Coefficient of Determination.
Adjusted R^2

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

Sensitivity & Recall = 80.00%
Specificity = 96.00%
Accuracy = 88.00%
Precision - 95%