**MACHINE LEARNING WORKSHEET-6**

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

Q1.In which of the following you can say that the model is overfitting?

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

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

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

D) None of the above

A1. ‘A’

Q2. Which among the following is a disadvantage of decision trees?

A) Decision trees are prone to outliers.

B) Decision trees are highly prone to overfitting.

C) Decision trees are not easy to interpret

D) None of the above.

A2.’B’

Q3. Which of the following is an ensemble technique?

A) SVM B) Logistic Regression

C) Random Forest D) Decision tree

A3. ‘C’

Q4. 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?

A) Accuracy B) Sensitivity

C) Precision D) None of the above.

A4: ‘A’

Q5. 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?

A) Model A B) Model B

C) both are performing equal D) Data Insufficient

A5. ‘A’

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

Q6. Which of the following are the regularization technique in Linear Regression??

A) Ridge B) R-squared

C) MSE D) Lasso

A6: A,B, & D

Q7. Which of the following is not an example of boosting technique?

A) Adaboost B) Decision Tree

C) Random Forest D) Xgboost.

A7: B &C

Q8. Which of the techniques are used for regularization of Decision Trees?

A) Pruning B) L2 regularization

C) Restricting the max depth of the tree D) All of the above

A8: A&C

Q9. 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

C) It is example of bagging technique

D) None of the above

A9: A &B

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?

Answer: R-squared called the coefficient of determination, it used to explain the degree to which input variables

(predictor variables) the variation of output variables (predicted variables). It ranges from 0 to 1.

It measures the proportion of variation explained by only those independent variables that really help

in explaining the dependent variable. It penalizes you for adding independent

variable that do not help in predicting the dependent variable.

Adjusted R-Squared can be calculated mathematically in terms of sum of squares.

Q11. Differentiate between Ridge and Lasso Regression.

A11. Ridge regression is an extension for linear regression. It’s basically a regularized linear regression model. The λ parameter is a scalar that should be learned as well, using a method called cross validation that will be discussed in another post.

A cruicial fact about ridge regression is that it enforces the β coefficients to be lower, but it does not enforce them to be zero. That is, it will not eliminate of irrelevant features but rather , only minimize their impact on the trained model.

The only difference between this and Ridge regression is that the regularisation term is expressed in absolute value. However, this distinction has a significant effect on the previously mentioned trade-off. The Lasso approach overcomes the drawback of Ridge regression by not only penalising high coefficient values but also setting them to zero if they are irrelevant. As a result, you could end up with fewer features in the model than you began with, which is a huge benefit.

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

Answer:

The variance inflation factor (VIF) quantifies the degree of multicollinearity in a set of multiple regression variables. The VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model with only that single independent variable. For each independent variable, this ratio is computed. A high VIF means that the independent variable associated with it is strongly collinear with the other variables in the model.

If the VIF value is more than 1 , it is sais that one can proceed with the regression procedure

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

Answer: Scaling is the process of placing values in the same range or scale such that no variable is controlled by the other

It is a technique used to standardize the independent features present in data in a fixed range. It is used during data pre-processing to deal with highly varying magnitudes, values, or units. If feature scaling is not performed, a machine learning algorithm would prefer to weight greater values as higher and consider smaller values as lower, regardless of the unit of measurement.

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

Answer: For check the goodness of fit in linear model we are use three different matrics which is below:

A. R Square/Adjusted R Square

B. Mean Squared Error (MSE)

C. Mean Absolute Error

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

Actual/Predicted True False

True 1000 50

False 250 1200

Answer:

Sensitivity= TP/TP+FN =1000/1000+50 = .95

Accuracy= TP+TN/TP+TN+FP+FN = 1000+1200/1000+250+50+1200 = .88

Precision= TP/TP+FP = 1000/1000+250 = .8

Specificity= TN/TN+FP = 1200/1200+250 = .82

F Score= 2\*Recall\*Precision/Recall+Precision 2\*(.95)\*(.8)/(.95)+(.8) = .86