WORKSHEET- 6 MACHINE LEARNING

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?

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

Answer : 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?

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.

Answer : c) Decision trees are not easy to interpret

3. Which of the following is an ensemble technique?

a) SVM

b) Logistic Regression

c) Random Forest

d) Decision tree

Answer : 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?

a) Accuracy
b) Sensitivity
c) Precision
d) None of the above

Answer : a) Accuracy

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?

a) Model A
b) Model B
c) both are performing equal
d) Data Insufficient

Answer : 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
b) R-squared
c) MSE
d) Lasso

Answer : a) Ridge

d) Lasso

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

a) Adaboost
b) Decision Tree
c) Random Forest
d) Xgboost

Answer : b) Decision Tree

c) Random Forest

8. 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

Answer : 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
c) It is example of bagging technique
d) None of the above

Answer : 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

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 : The adjusted R2 will penalize you for adding independent variables (K in the equation) that do not fit the model. Why? In regression analysis, it can be tempting to add more variables to the data as you think of them. Some of those variables will be significant, but you can't be sure that significance is just by chance. The adjusted R2 will compensate for this by that penalizing you for those extra variables. While values are usually positive, they can be negative as well. This could happen if your R2 is zero; After the adjustment, the value can dip below zero. This usually indicates that your model is a poor fit for your data. Other problems with your model can also cause sub-zero values, such as not putting a constant term in your model.

11. Differentiate between Ridge and Lasso Regression.

Answer : 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 that ridge regression can't 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 co-efficient 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?

Answer : A variance inflation factor(VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model. Variance inflation factors range from 1 upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor:

1 = not correlated.

Between 1 and 5 = moderately correlated.

Greater than 5 = highly correlated.

Exactly how large a VIF has to be before it causes issues is a subject of debate. What is known is that the more your VIF increases, the less reliable your regression results are going to be. 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.

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

Answer : When you're working with a learning model, it is important to scale the features to a range which is centered around zero. This is done so that the variance of the features are in the same range. If a feature's variance is orders of magnitude more

than the variance of other features, that particular feature might dominate other features in the dataset, which is not something we want happening in our model.

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

Answer : The Curve Fitting Toolbox supports these goodness of fit statistics for parametric models:

- The sum of squares due to error (SSE)
- R-square
- Adjusted R-square
- Root mean squared error (RMSE)

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

= 0.95

Specificity :

TN/TN+FP = 1200/1200+250

= 0.83

Precision :

TP/TP+FP = 1000/1000+250

= 0.80

Recall :

TP/TP+FN = 1000/1000+1200

= 0.45

Accuracy :

TP+TN/TP+TN+FP+FN = 1000+1200/1000+250+50+1200

= 0.88