Name of the Assignment: Machine Learning (Worksheet-6)

Submitted by : Shalini Joshi

Designation : Data Science Intern

Date of Submission : 26th Feb,2023

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

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

Ans: A) Decision trees are prone to outliers

3. Which of the following is an ensemble technique?

A) SVM B) Logistic Regression C) Random Forest D) Decision tree

Ans: 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.

Ans: 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?

A) Model A B) Model B C) both are performing equal D) Data Insufficient

Ans: 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

Ans: 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.

Ans: 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

Ans: 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

Ans: A) & C)

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 is a modified version of R-squared that adjusts for the number of predictors in a regression model. It is calculated as: Adjusted R2 = 1 – [(1-R2)*(n-1)/(n-k-1)]

Where:

- R2: The R2 of the model
- n: The number of observations
- k: The number of predictor variables

Because R-squared always increases as you add more predictors to a model, the adjusted R-squared can tell you how useful a model is, adjusted for the number of predictors in a model.The advantage of Adjusted R-squared:

- Adjusted R-squared tells us how well a set of predictor variables is able to explain the variation in the response variable, adjusted for the number of predictors in a model.

- Because of the way it’s calculated, adjusted R-squared can be used to compare the fit of regression models with different numbers of predictor variables.

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.

Whereas 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: A **variance inflation factor (VIF)** is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

VIF = 1 is suitable value for a feature to be included in a regression modelling

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

Ans: Given the use of small weights in the model and the use of error between predictions and expected values, the scale of inputs and outputs used to train the model are an important factor. Unscaled input variables can result in a slow or unstable learning process, whereas unscaled target variables on regression problems can result in exploding gradients causing the learning process to fail.

Data preparation involves using techniques such as the normalization and standardization to rescale input and output variables prior to training a neural network model.

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

Ans: The different metrics which are used to check the goodness of fit in linear regression are as follows:

- R square/Adjusted R square
- Mean Absolute Error (MAE)
- Mean Square Error (MSE)
- Root Mean Square 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

Ans:

Here, TP = 1000, FN = 50, FP = 250 and TN = 1200

- Specificity= TN/(TN+FP)
  = 1200/(1200+250)
  = **0.827586**

- Precision = TP/(TP+FP)
  =1000/(1000+250)
  = **0.8**

- Recall/Sensitivity=TP/(TP+FN)
  =1000/(1000+50)
  = **0.952381**
- Accuracy = (TP+TN)/(TP+TN+FP+FN)
  =(1000+1200)/(1000+1200+250+50)
  =**0.88**

***************** **The End** ************************

******************************* **The End***