**Machin-Learning Assignment-6**

**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: 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.

**ANS: B) Decision trees are highly prone to overfitting.**

**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: C) Precision**

**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**

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

A) Ridge

B) R-squared

C) MSE

D) Lasso

**ANS: A) Ridge and 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**

**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: D) All of the above**

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

**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 accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not.

The R-squared, also called the coefficient of determination, is used to explain the degree to which input variables (predictor variables) explain the variation of output variables (predicted variables). It ranges from 0 to 1. For example, if the R-squared is 0.9, it indicates that 90% of the variation in the output variables are explained by the input variables

Models with tons of predictors tend to perform better in sample than when tested out of sample. The adjusted R 2 "penalizes" you for adding the extra predictor variables that don't improve the existing model. It can be helpful in model selection. Adjusted R 2 will equal R 2 for one predictor variable**.**

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

**ANS:**

Regression analysis is a way that can be used to determine the relationship between the predictor variable (x) and the target variable (y).

When people begin their Machine Learning journey, they often start with Linear Regression, one of the most simple algorithms out there. However, this model quickly shows its limitations, especially when working with datasets that lead models to overfit. The main solutions to this are called Ridge and Lasso regressions**.**

**Ridge regression:**

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. It has been used in many fields including econometrics, chemistry, and engineering. Also known as Tikhonov regularization, named for Andrey Tikhonov, it is a method of regularization of ill-posed problems.It is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters.

**Lasso Regression:**

This is a regularization technique used in feature selection using a Shrinkage method also referred to as the penalized regression method. Lasso is short for Least Absolute Shrinkage and Selection Operator, which is used both for regularization and model selection. If a model uses the L1 regularization technique, then it is called lasso regression.

The main difference between Ridge and LASSO Regression is that if ridge regression can shrink the coefficient close to 0 so that all predictor variables are retained. Whereas LASSO can shrink the coefficient to exactly 0 so that LASSO can select and discard the predictor variables that have the right coefficient of 0.

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

**ANS: Variance Inflation Factor (VIF):**

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**.** Variance inflation factor (VIF) is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis.VIF measures the number of inflated variances caused by multicollinearity.

Variance inflation factors allow a quick measure of how much a variable is contributing to the standard error in the regression. When significant multicollinearity issues exist, the variance inflation factor will be very large for the variables involved. After these variables are identified, several approaches can be used to eliminate or combine collinear variables, resolving the multicollinearity issue.

The Variance Inflation Factor (VIF) measures the impact of collinearity among the variables in a regression model. The Variance Inflation Factor (VIF) is 1/Tolerance, it is always greater than or equal to 1. There is no formal VIF value for determining presence of multicollinearity. Values of VIF that exceed 10 are often regarded as indicating multicollinearity, but in weaker models values above 2.5 may be a cause for concern. In many statistics programs, the results are shown both as an individual R2 value (distinct from the overall R2 of the model) and a Variance Inflation Factor (VIF). When those R2 and VIF values are high for any of the variables in your model, multicollinearity is probably an issue. When VIF is high there is high multicollinearity and instability of the b and beta coefficients. It is often difficult to sort this out.

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

**ANS:** Scaling of the data comes under the set of steps of data pre-processing when we are performing machine learning algorithms in the data set.

If scaling is required, then it should be done on both the train and test data sets.it depends on the data we are working with and the type of scaling to be done.

We do data scaling, when we are seeking for some relation between data point. In ANN and other data mining approaches we need to normalize the inputs, otherwise network will be ill-conditioned. We do the scaling to reach a linear, more robust relationship. Moreover, data scaling can also help you a lot to overcome outliers in the data. In short, data scaling is highly recommended in each type of machine learning algorithms. You can do normalization or standardization in order to scale your data. [Notice that do not confuse normalization with standardization (e.g. Z-score)] Hope that helps.Scaling is better to be done in general, because if all the features are on the same scale, the Gradient Descent Algorithm converges faster to the global or optimum local minimum.

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

**ANS**: linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

Evaluation metrics are a measure of how good a model performs and how well it approximates the relationship. Let us look at MSE, MAE, R-squared, Adjusted R-squared, and RMSE. .

**Mean Squared Error (MSE):**

The most common metric for regression tasks is MSE. It has a convex shape. It is the average of the squared difference between the predicted and actual value. Since it is differentiable and has a convex shape, it is easier to optimize.MSE penalizes large errors.

**Mean Absolute Error (MAE):**

This is simply the average of the absolute difference between the target value and the value predicted by the model. Not preferred in cases where outliers are prominent**.**MAE does not penalize large errors.

**R-squared:**

This metric represents the part of the variance of the dependent variable explained by the independent variables of the model. It measures the strength of the relationship between your model and the dependent variable.

**Root Mean Squared Error (RMSE):**

This is the square root of the average of the squared difference of the predicted and actual value.Basically, RMSE is just the root of the average of squared residuals. We know that residuals are a measure of how distant the points are from the regression line. Thus, RMSE measures the scatter of these residuals.