MACHINE LEARNING

ASSIGNMENT - 6

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: **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: **C) both are performing equal**

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, 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: **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) 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?

Ans: The adjusted R-squared is a modified version of R-squared that adjusts for predictors that are not significant in a regression model.

Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model.

Adjusted R-Squared

Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Typically, the adjusted R-squared is positive, not negative. It is always lower than the R-squared.

Adding more independent variables or predictors to a regression model tends to increase the R-squared value, which tempts makers of the model to add even more variables. This is called overfitting and can

return an unwarranted high R-squared value. Adjusted R-squared is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables.2

In a portfolio model that has more independent variables, adjusted R-squared will help determine how much of the correlation with the index is due to the addition of those variables. The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

Key Differences

The most obvious difference between adjusted R-squared and R-squared is simply that adjusted R-squared considers and tests different independent variables against the stock index and R-squared does not. Because of this, many investment professionals prefer using adjusted R-squared because it has the potential to be more accurate. Furthermore, investors can gain additional information about what is affecting a stock by testing various independent variables using the adjusted R-squared model.

R-squared, on the other hand, does have its limitations. One of the most essential limits to using this model is that R-squared cannot be used to determine whether or not the coefficient estimates and predictions are biased. Furthermore, in multiple linear regression, the R-squared can not tell us which regression variable is more important than the other.

Adjusted R-Squared vs. Predicted R-Squared

The predicted R-squared, unlike the adjusted R-squared, is used to indicate how well a regression model predicts responses for new observations. So where the adjusted R-squared can provide an accurate model that fits the current data, the predicted R-squared determines how likely it is that this model will be accurate for future data.

R-Squared vs. Adjusted R-Squared Examples

When you are analyzing a situation in which there is a guarantee of little to no bias, using R-squared to calculate the relationship between two variables is perfectly useful. However, when investigating the relationship between say, the performance of a single stock and the rest of the S&P500, it is important to use adjusted R-squared to determine any inconsistencies in the correlation.

If an investor is looking for an index fund that closely tracks the S&P500, they will want to test different independent variables against the stock index such as the industry, the assets under management, how long the stock has been available on the market, and so on to ensure they have the most accurate figure of the correlation.

Conclusion:Using adjusted R-squared over R-squared may be favored because of its ability to make a more accurate view of the correlation between one variable and another. Adjusted R-squared does this by taking into account how many independent variables are added to a particular model against which the stock index is measured.

11. Differentiate between Ridge and Lasso Regression.

Ans: Ridge and lasso regression are **two common machine learning approaches for constraining model parameters**.

Both methods try to get the coefficient estimates as close to zero as possible because minimizing (or shrinking) coefficients can reduce variance dramatically (i.e., overfitting).

- Similar to the lasso regression, ridge regression puts a similar constraint on the coefficients by introducing a penalty factor.

However, **while lasso regression takes the magnitude of the coefficients, ridge regression takes the square**.

Ridge regression is also referred to as L2 Regularization.

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.

The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

Variance inflation factor (VIF) is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis.

Multicollinearity inflates the variance and type II error. It makes the coefficient of a variable consistent but unreliable.

VIF measures the number of inflated variances caused by multicollinearity.

VIF can be calculated by the formula below:

Variance Inflation Factor - Formula

$$VIF_i = \frac{1}{1 - R_i^2} = \frac{1}{\text{Tolerance}}$$

Where Ri2 represents the unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones. The reciprocal of VIF is known as tolerance. Either VIF or tolerance can be used to detect multicollinearity, depending on personal preference.

If Ri2 is equal to 0, the variance of the remaining independent variables cannot be predicted from the ith independent variable. Therefore, when VIF or tolerance is equal to 1, the ith independent variable is not correlated to the remaining ones, which means multicollinearity does not exist in this regression model. In this case, the variance of the ith regression coefficient is not inflated.

Generally, a VIF above 4 or tolerance below 0.25 indicates that multicollinearity might exist, and further investigation is required. When VIF is higher than 10 or tolerance is lower than 0.1, there is significant multicollinearity that needs to be corrected.

However, there are also situations where high VFIs can be safely ignored without suffering from multicollinearity. The following are three such situations:

1. High VIFs only exist in control variables but not in variables of interest. In this case, the variables of interest are not collinear to each other or the control variables. The regression coefficients are not impacted.

2. When high VIFs are caused as a result of the inclusion of the products or powers of other variables, multicollinearity does not cause negative impacts. For example, a regression model includes both x and x2 as its independent variables.

3. When a dummy variable that represents more than two categories has a high VIF, multicollinearity does not necessarily exist. The variables will always have high VIFs if there is a small portion of cases in the category, regardless of whether the categorical variables are correlated to other variables

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

Ans: we need to scale the data before feeding it to the train the model

- To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features,
- we scale the data before feeding it to the model.
- Scaling the target value is a good idea in regression modelling; scaling of the data makes it easy for a model to learn and understand the problem. 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.
- Feature scaling is essential for machine learning algorithms that calculate distances between data. If not scale, the feature with a higher value range starts dominating when calculating distances, as explained intuitively in the “why?” section.
- The ML algorithm is sensitive to the “relative scales of features,” which usually happens when it uses the numeric values of the features rather than say their rank.

- In many algorithms, when we desire faster convergence, scaling is a MUST like in Neural Network.
- Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions do not work correctly without normalization. For example, the majority of classifiers calculate the distance between two points by the distance. If one of the features has a broad range of values, the distance governs this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.
- Even when the conditions, as mentioned above, are not satisfied, you may still need to rescale your features if the ML algorithm expects some scale or a saturation phenomenon can happen. Again, a neural network with saturating activation functions (e.g., sigmoid) is a good example.
- Rule of thumb we may follow here is an algorithm that computes distance or assumes normality, scales your features.
- Some examples of algorithms where feature scaling matters are:
- K-nearest neighbors (KNN) with a Euclidean distance measure is sensitive to magnitudes and hence should be scaled for all features to weigh in equally.
- K-Means uses the Euclidean distance measure here feature scaling matters.
- Scaling is critical while performing Principal Component Analysis(PCA). PCA tries to get the features with maximum variance, and the variance is high for high magnitude features and skews the PCA towards high magnitude features.
- We can speed up gradient descent by scaling because θ descends quickly on small ranges and slowly on large ranges, and oscillates inefficiently down to the optimum when the variables are very uneven.
- Algorithms that do not require normalization/scaling are the ones that rely on rules. They would not be affected by any monotonic transformations of the variables. Scaling is a monotonic transformation. Examples of algorithms in this category are all the tree-based algorithms — CART, Random Forests, Gradient Boosted Decision Trees. These algorithms utilize rules (series of inequalities) and do not require normalization.
- Algorithms like Linear Discriminant Analysis(LDA), Naive Bayes is by design equipped to handle this and give weights to the features accordingly. Performing features scaling in these algorithms may not have much effect.

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

Ans: • Mean Squared Error (MSE)

• Mean Absolute Error (MAE)

• R-squared or Coefficient of Determination

• Adjustable R-squared

• 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

1. Ans: Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN.
2. Misclassification (all incorrect / all) = FP + FN / TP + TN + FP + FN.
3. Precision (true positives / predicted positives) = TP / TP + FP.
4. Sensitivity aka Recall (true positives / all actual positives) = TP / TP + FN.
5. Specificity (true negatives / all actual negatives) =TN / TN + FP

Solution:

Given TP = 1000

FP = 50

FN = 250

TN = 1200

1. Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN

= 1000+1200/1000+1200+50+250

= 0.88 = 88% accuracy.

2. Misclassification (all incorrect / all) = FP + FN / TP + TN + FP + FN

   (50 + 250) / 1000+1200+50+250

   = 300 / 2500 = 0.12 or 12% Misclassification

   You can also just do 1 — Accuracy, so:

   1–0.88 = 0.12 or 12% Misclassification

3. Precision (true positives / predicted positives) = TP / TP + FP

= 1000/1000+50

= 0.95 = 95% precision.

4. Sensitivity aka Recall (true positives / all actual positives) = TP / TP + FN

= 1000/1000+250

= 0.8 = 80% sensitivity.

5. Specificity (true negatives / all actual negatives) =TN / TN + FP

= 1200/1200+50

=0.96 = 96%specificity.
