MACHINE LEARNING

1) In which of the following you can say that the model is overfitting?

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?

Answer: B) Decision trees are highly prone to overfitting.

****
****

3) Which of the following is an ensemble technique?

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?

Answer: 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?

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??

Answer: A) Ridge and D) Lasso

****
****

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

Answer: B) Decision Tree and c) Random Forest

****
****

8) Which of the techniques are used for regularization of Decision Trees?

Answer: A) Pruning and C) Restricting the max depth of the tree

****
****

9) Which of the following statements is true regarding the Adaboost technique?

Answer: 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 R-squared formula penalizes the presence of unnecessary predictors in the model by subtracting a term from the R-squared. This term increases as the number of predictors increases

relative to the sample size. The adjusted R-squared value will only increase if the additional predictor significantly improves the model's predictive power beyond what would be expected by chance. In summary, the adjusted R-squared is a modified version of R-squared that takes into account the number of predictors in the model. It helps to avoid overfitting and provides a more accurate measure of the model's generalization performance.

****
****

11) Differentiate between Ridge and Lasso Regression.

Answer: 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).

Ridge regression: Adds a penalty term equal to the square of the coefficients. It shrinks the coefficients of all predictors towards zero, but does not eliminate any predictors. It is useful when all predictors are potentially useful, but some may have a small effect. It generally produces less sparse models.

Lasso regression: Adds a penalty term equal to the absolute value of the coefficients. It can eliminate some predictors by setting their coefficients to zero, effectively removing them from the model. It is useful when the number of predictors is high, and some are less important or irrelevant. It can produce very sparse models.

In short we can say, Ridge regression shrinks all the coefficients towards zero, while Lasso regression can eliminate some of the coefficients entirely. The choice of which technique to use depends on the problem at hand and the characteristics of the predictors.

****
****

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

Answer: VIF stands for Variance Inflation Factor, and it is a measure of the degree of multicollinearity (correlation) among the predictor variables in a linear regression model. Specifically, it measures the extent to which the variance of the estimated regression coefficient is increased due to multicollinearity among the predictors.

The VIF for a predictor variable is calculated as the ratio of the variance of its estimated coefficient in a model that includes all the predictor variables to the variance of its estimated coefficient in a model that excludes that particular predictor variable. The VIF for a predictor variable ranges from 1 to infinity, with a VIF of 1 indicating no multicollinearity and a VIF greater than 1 indicating some degree of multicollinearity.

The suitable value of VIF for a feature to be included in a regression model depends on the degree of multicollinearity that is acceptable for the problem at hand. As a general rule of thumb, a VIF value of 1 to 2 is considered low and indicates little to no multicollinearity, while a VIF value of 5 to 10 or higher indicates high multicollinearity and suggests that the variable may need to be removed from the model.

However, the suitable value of VIF also depends on the context of the problem and the complexity of the model. In some cases, a higher VIF may be acceptable if the variable is essential to the model and removing it would result in a loss of important information. On the other hand, in other cases, a lower VIF may be required if the model needs to be simpler and more interpretable. Ultimately, the decision of what VIF value to use should be based on a careful consideration of the problem at hand and the goals of the analysis.

****
****

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

Answer: Scaling the data before feeding it to a machine learning model can be important for several reasons:

Different scales: When the features of the dataset are measured in different units, or have different ranges of values, the scale of each feature can impact the way the model interprets the importance of each feature. For example, if one feature has values that range from 0 to 1 and another feature has values that range from 0 to 100, the model may incorrectly assign more importance to the feature with the larger range of values. Scaling the data puts all the features on a similar scale, which can improve the accuracy and stability of the model.

Gradient-based algorithms: Many machine learning algorithms rely on gradient-based optimization techniques to minimize a loss function and find the optimal set of model parameters. If the data is not scaled, the gradients can be different for each feature and this can cause the optimization process to take longer or get stuck in local optima. Scaling the data can speed up the optimization process and improve the overall performance of the model.

Regularization: Some regularization techniques, such as Ridge and Lasso regression, penalize large coefficients in the model. If the data is not scaled, features with larger scales will tend to have larger coefficients, even if they are not more important. Scaling the data ensures that all features are equally penalized for having large coefficients, which can help to prevent overfitting.

Overall, scaling the data can improve the performance, stability, and interpretability of machine learning models. Common methods of scaling data include standardization (subtracting the mean and dividing by the standard deviation), min-max scaling (scaling the values to a specific range, such as 0 to 1), and normalization (scaling the values to have unit norm).

****
****

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

Answer: In linear regression, there are several metrics that can be used to check the goodness of fit of the model. Some of the commonly used metrics are:

R-squared: R-squared is a measure of how well the linear regression model fits the data. It represents the proportion of the variance in the dependent variable that is explained by the independent variables. R-squared values range from 0 to 1, with higher values indicating a better fit.

Mean squared error (MSE): MSE measures the average squared difference between the predicted and actual values of the dependent variable. It gives an indication of the average magnitude of the errors in the model predictions.

Root mean squared error (RMSE): RMSE is the square root of the MSE and represents the average magnitude of the errors in the same units as the dependent variable.

Mean absolute error (MAE): MAE is the average absolute difference between the predicted and actual values of the dependent variable. It gives an indication of the average magnitude of the errors in the model predictions.

Residual plots: Residual plots can be used to visually check the goodness of fit of the model. A residual plot shows the difference between the predicted and actual values of the dependent variable for each observation, plotted against the independent variable. The plot should not show any clear patterns, which would indicate that the model is not capturing all of the relevant information in the data. Overall, these metrics can be used to evaluate the performance of a linear regression model and to compare the performance of different models. However, it's important to keep in mind that no single metric can provide a complete picture of the goodness of fit, and it's often useful to use a combination of metrics and visual inspection to evaluate the model.

****
****

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

Answer:

Sensitivity = TP / (TP + FN) = 1000 / (1000 + 250) = 0.8

Specificity = TN / (TN + FP) = 1200 / (1200 + 50) = 0.96

Precision = TP / (TP + FP) = 1000 / (1000 + 50) = 0.95

Recall = TP / (TP + FN) = 1000 / (1000 + 250) = 0.8

Accuracy = (TP + TN) / (TP + TN + FP + FN) = (1000 + 1200) / (1000 + 1200 + 50 + 250) = 0.88

****
****