ASSIGNMENT – 6

MACHINE LEARNING

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.

2. Which among the following is a disadvantage of decision trees?

B) Decision trees are highly prone to overfitting.

3. Which of the following is an ensemble technique?

A) SVM

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?

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?

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??

D) Lasso

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

D) Xgboost.

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

D) All of the above

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

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?

- 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.
- Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model.

11. Differentiate between Ridge and 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?

- A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.

- Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.
- A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

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

We scale down the images before feeding it into the network in order to reduce the number of parameters**.** When the number of parameters are high, we tend to increase the requirement of computation power. Scaling down images does decreases the detail and the scale size is purely dependent on the target of our model.

MACHINE LEARNING

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

Three statistics are used in Ordinary Least Squares (OLS) regression to evaluate model fit: R-squared, the overall F-test, and the Root Mean Square Error (RMSE). All three are based on two sums of squares: Sum of Squares Total (SST) and Sum of Squares Error (SSE).

R-squared

The difference between SST and SSE is the improvement in prediction from the regression model, compared to the mean model. Dividing that difference by SST gives R-squared. It is the proportional improvement in prediction from the regression model, compared to the mean model. It indicates the goodness of fit of the model.

R-squared has the useful property that its scale is intuitive. It ranges from zero to one. Zero indicates that the proposed model does not improve prediction over the mean model. One indicates perfect prediction. Improvement in the regression model results in proportional increases in R-squared.

One pitfall of R-squared is that it can only increase as predictors are added to the regression model. This increase is artificial when predictors are not actually improving the model's fit. To remedy this, a related statistic, Adjusted R-squared, incorporates the model's degrees of freedom.

RMSE

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance. It has the useful property of being in the same units as the response variable.

Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response. It's the most important criterion for fit if the main purpose of the model is prediction.

The best measure of model fit depends on the researcher's objectives, and more than one are often useful. The statistics discussed above are applicable to regression models that use OLS estimation.

Many types of regression models, however, such as mixed models, generalized linear models, and event history models, use maximum likelihood estimation. These statistics are not available for such models.

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

Actual/Predicted	**True**	**False**
True	**1000**	**50**
False	**250**	**1200**

Accuracy = 46.15%	**Precision Score = 97.56%**	**Recall Score = 83.33%**