Machine 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.

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?

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?

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?

C) both are performing equal

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

A) Ridge D) Lasso

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

B) Decision Tree C) Random Forest

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

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

10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

The adjusted R2 will penalize you for adding independent variables (K in the equation) that do not fit the model. Why? In regression analysis, it can be tempting to add more variables to the data as you think of them. Some of those variables will be significant, but you can't be sure that significance is just by chance. The adjusted R2 will compensate for this by that penalizing you for those extra variables.

While values are usually positive, they can be negative as well. This could happen if your R2 is zero; After the adjustment, the value can dip below zero. This usually indicates that your model is a poor fit for your data. Other problems with your model can also cause sub-zero values, such as not putting a constant term in your model.

11. Differentiate between Ridge and Lasso Regression.

Lasso tends to give sparse weights (most zeros), because the l1 regularization cares equally about driving down big weights to small weights, or driving small weights to zeros. If you have a lot of predictors (features), and you suspect that not all of them are that important, Lasso may be really good idea to start with.

Ridge tends to give small but well distributed weights, because the l2 regularization cares more about driving big weight to small weights, instead of driving small weights to zeros. If you only have a few predictors, and you are confident that all of them should be really relevant for predictions, try Ridge as a good regularized linear regression method.

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) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

VIFs are usually calculated by software, as part of regression analysis. You'll see a VIF column as part of the output. VIFs are calculated by taking a predictor, and regressing it against every other predictor in the model. This gives you the R-squared values, which can then be plugged into the VIF formula. "i" is the predictor you're looking at (e.g. x1 or x2)

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

Most research papers consider a VIF (Variance Inflation Factor) > 10 as an indicator of multicollinearity, but some choose a more conservative threshold of 5 or even 2.5.

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

The input variables are those that the network takes on the input or visible layer in order to make a prediction.

A good rule of thumb is that input variables should be small values, probably in the range of 0-1 or standardized with a zero mean and a standard deviation of one.

Whether input variables require scaling depends on the specifics of your problem and of each variable.

You may have a sequence of quantities as inputs, such as prices or temperatures.

If the distribution of the quantity is normal, then it should be standardized, otherwise the data should be normalized. This applies if the range of quantity values is large (10s, 100s, etc.) or small (0.01, 0.0001).

If the quantity values are small (near 0-1) and the distribution is limited (e.g. standard deviation near 1) then perhaps you can get away with no scaling of the data.

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

- 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.

Given TP = 1000

FP = 50

FN = 250

TN = 1200

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

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

= 0.88 = 88%

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

= 1000/1000+50

= 0.95 = 95%

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

= 1000/1000+250

= 0.8 = 80%

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

= 1200/1200+50

=0.96 = 96%