Parth Makwana

MACHINE LEARNING

1. In which of the following you can say that the model is overfitting?
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?
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 followingmetrics you would focus on?
A) **Accuracy**

5. The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and ofmodel B is 0.85. Which of these two models is doing better job in classification?
B) **Model B**

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?
B) **L2 regularization**

9. Which of the following statements is true regarding the Adaboost technique?
C) **It is example of bagging technique**

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

- An adjusted R-squared is a measure of how well a regression model fits a given dataset. It indicates the percentage of variance in the response variable that is
- explained by the model, taking into account the number of predictors used. It is an adjusted version of the R-squared statistic that penalizes models for using too many predictors. A higher adjusted R-squared value indicates a better fit for the model.

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

LASSO Model

- The LASSO method aims to produce a model that has high accuracy and only uses a subset of the original features. The way it does this is by putting in a constraint wherethe sum of the absolute values of the coefficients is less than a fixed value. To that endit lowers the size of the coefficients and leads to some features having a coefficient of0, essentially dropping it from the model. In this way, it is also a form of filtering your features and you end up with a model that is simpler and more interpretable.

Ridge Regression

- The Ridge Regression method was one of the most popular methods before the LASSO method came about. The idea is similar, but the process is a little different. The Ridge Regression also aims to lower the sizes of the coefficients to avoid over-fitting, but it does not drop any of the coefficients to zero. The constraint it uses is to have the sum of the squares of the coefficients below a fixed value. The Ridge Regression improves the efficiency, but the model is less interpretable due to the potentially high number of features.

- It performs better in cases where there may be high multi-colinearity, or high
- correlation between certain features. This is because it reduces variance in exchange for bias. You also need to make sure that the number of features is less than the
- number of observations before using Ridge Regression because it does not dropfeatures and in that case may lead to bad predictions.

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

- The Variance Inflation Factor (VIF) measures the severity of multicollinearity
- in regression analysis. It is a statistical concept that indicates the increase in thevariance 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.
- 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

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

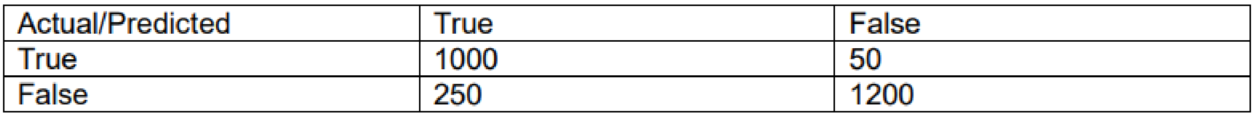
- In general, scaling is not an absolute requirement, its a recommendation,
- primarily for similarity based algorithms. For many algorithms, you may need to consider data transformation prior to normalization.There's also various
- normalization techniques you can try out, and there's no one size fits best for all problems. The main reason for normalization for error based algorithms such as linear, logistic regression, neural networks is faster convergence to the global
- minimum due to the better initialization of weights.Information based algorithms (Decision Trees, Random Forests) and probability based algorithms (Naive Bayes, Bayesian Networks) don't require normalization either.

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

- **Mean Absolute Error(MAE)**
- This is the simplest of all the metrics. It is measured by taking the average of the absolute difference between actual values and the predictions.
- **Root Mean Square Error(RMSE)**
- The Root Mean Square Error is measured by taking the square root of the average of the squared difference between the prediction and the actual value. It represents the sample standard deviation of the differences between predicted values and observed values(also called residuals).
- **Adjusted R-squared**
- There is a drawback of R^2 that it improves every time when we add new variables in the model.
- Think about it, whenever you add a new variable there can be two circumstances, either the new variable improves your model or not. When the new variable improves your model then it is ok. But what if it does not improve your model? Then the problem occurs. The value of R^2 keeps on increasing with the addition of more independent variables even though they may not have a significant impact on the prediction
- **Coefficient of Determination or R^2**
- It measures how well the actual outcomes are replicated by the regression line. It helps you to understand how well the independent variable adjusted with the variance in your model. That means how good is your model for a dataset.

Parth Makwana

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



Actual/Predicted	True	False
True	1000	50
False	250	1200

		Predicted	
		True	False
Actual	True	1000	50
	False	250	1200

To calculate the various metrics:

Sensitivity = True Positive / (True Positive + False Negative)

- True Positive = 1000
- False Negative = 250
- Sensitivity = 1000 / (1000 + 250) = 0.8

Specificity = True Negative / (True Negative + False Positive)

- True Negative = 1200
- False Positive = 50
- Specificity = 1200 / (1200 + 50) = 0.96

Precision = True Positive / (True Positive + False Positive)

- True Positive = 1000
- False Positive = 50
- Precision = 1000 / (1000 + 50) = 0.9524

Recall (same as sensitivity)

Accuracy = (True Positive + True Negative) / (True Positive + False Positive + True Negative + False Negative)

- True Positive = 1000
- False Positive = 50
- True Negative = 1200
- False Negative = 250
- Accuracy = (1000 + 1200) / (1000 + 50 + 1200 + 250) = 0.88

Therefore, the sensitivity is 0.8, specificity is 0.96, precision is 0.9524, recall is 0.8 and accuracy is 0.88.