MACHINE LEARNING WORKSHEET

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?
A) Least Square Error   B) Maximum Likelihood
C) Logarithmic Loss   D) Both A and B

Ans : A)Least Square Error

2. Which of the following statements is true about outliers in linear regression?
A) Linear regression is sensitive to outliers
B) linear regression is not sensitive to outliers
C) Can't say
D) none of these

Ans : A)Linear regression is sensitive to outliers

3. A line falls from left to right if a slope is _____?
A) Positive
B) Negative
C) Zero
D) Undefined

Ans : B)Negative

4. Which of the following will have symmetric relation between dependent variable and independent variable?
A) Regression
B) Correlation
C) Both of them
D) None of these

Ans : C)both of them

5. Which of the following is the reason for over-fitting?
A) High bias and high variance
B) Low bias and low variance
C) Low bias and high variance
D) none of these

Ans : C)Low bias and high variance

6. If output involves label then that model is called as:
A) Descriptive model
B) Predictive model
C) Reinforcement learning
D) All of the above

Ans : B)Predictive model

7. Lasso and Ridge regression techniques belong to _____?
A) Cross validation
B) Removing outliers
C) SMOTE
D) Regularization

Ans : D)Regularization

8. To overcome an imbalance dataset, which technique can be used?
A) Cross validation
B) Regularization
C) Kernel
D) SMOTE

Ans : SMOTE

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?
A) TPR and FPR
B) Sensitivity and precision
C) Sensitivity and Specificity
D) Recall and precision

Ans : TPR and FPR

10. The AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.
A) True
B) False

Ans : b)False

11. Pick the feature extraction from below:
A) Construction bag of words from a email
B) Apply PCA to project high dimensional data
C) Removing stop words
D) Forward selection

Ans : B)Apply PCA to project high dimensional data

In Q12, more than one options are correct, choose all the correct options:

12. Which of the following is true about the Normal Equation used to compute the coefficient of the Linear Regression?
A) We don't have to choose the learning rate.
B) It becomes slow when the number of features is very large.
C) We need to iterate.
D) It does not make use of dependent variables.

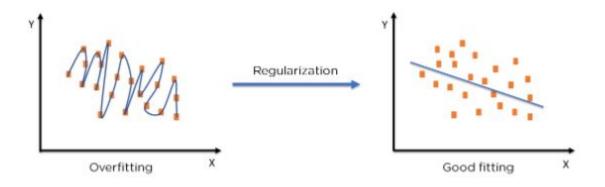**Ans : A) We don't have to choose the learning rate.**
     B) It becomes slow when the number of features is very large.

Q13 and Q15 are subjective answer type questions, Answer them briefly.

13. Explain the term regularization?

1)Regularization refers to techniques that are used to calibrate machine learning models in order to minimise the adjusted loss function and prevent overfitting or underfitting.

2)Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

14.Which particular algorithms are used for regularization?

There are two main types of regularization techniques:
1)Ridge Regularization
2)Lasso Regularization
1)Ridge Regularization

Also known as Ridge Regression, it modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients.This means that the mathematical function representing our machine learning model is minimised and coefficients are calculated. The magnitude of coefficients is squared and added. Ridge Regression performs regularization by shrinking the coefficients present.

2)Lasso Regularization

It modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients. Lasso regression also performs coefficient minimization, but instead of squaring the magnitudes of the coefficients, it takes the true values of coefficients. This means that the coefficient sum can also be 0, because of the presence of negative coefficients.

15. Explain the term error present in linear regression equation?

The standard error of regression (S), also known as the standard error of the estimate, represents the average distance that the observed values fall from the regression line. Conveniently, it tells you how wrong the regression model is on average using the units of the response variable. Smaller values are better because it indicates that the observations are closer to the fitted line.