MACHINE LEARNING WORKSHEET-1

Answers

1. a) Least Square Error
2. b) Linear Regression is sensitive to outliers
3. b) Negative
4. a) Correlation
5. c) Low bias and High Variance
6. b) Predictive Model
7. d) Regularizations
8. d)SMOTE
9. a) TPR and FPR
10. b) False
11. b) Apply PCA to project high dimensional data
12. a) We don't have to choose the learning rate
    b) It becomes slow when number of features is very large
13. Overfitting is the modelling error in statistics observed commonly in machine learning models, where the models perform well on training data sets but cannot generalize well on unseen(test) datasets. If the model is overfitting the data it may be due to the too many parameters increasing the model complexity, it may be also called high variance. Underfitting(high bias) is observed when the model fails to recognize patterns from the parameters given. Tuning the complexity of the model with Regularization helps in finding the good bias-variance tradeoff.
    Regularization helps in sorting overfitting problems by restricting degrees of freedom (weights of polynomial function). The concept behind regularization is to penalize the extreme parameter weights.

14. The different types of Regularizations.
    a. RIDGE(L2-Norm)
       The cost function is altered by adding a penalty term to the cost function, which multiples lambda with squared weights of individual features. The penalty term regularizes the coefficients decreasing the complexity of the model
       Uses of Ridge Regression:
       - The variable with multicollinearity problems, the general regression fail to solve the problems, we can use Ridge regression.
       - If we have more parameters than samples, Ridge regression can solve problem

       Limitations

       - It helps in reducing complexity, but the number of parameters remains the same, which fails to help in feature selection
       - The model interpretability, final model will contain all the parameters as the co-efficient are reduced to zero but never zero.

    b. LASSO (Least Absolute Shrinkage and Selection Operator)
       It is also called L1-Norm, In this technique, the penalty factor can force some of the coefficients to exact zero which means removal of coefficients completely for model evaluation.
       Limitations
       - If there are multiple collinear variables in the model, then LASSO selects one of variables randomly which is not good for model interpretation.

15. Linear Regression Equation:

    Y = a+bx+e
    Where,
    Y = the variable to be predicted
    a = intercept
    b = slope
    e = regression residual error.
    Error is the residual value which is difference between predicted value and actual value. In regression line, it is the vertical distance between dataset and regression line. Each data point will have one residual,
    - Positive, If datapoint is above regression line
    - Negative, if datapoint is below regression line
    - Zero, if datapoint lies on regression line

The error term can be formulated as,

Residual(error) = Observed value – predicted value

The Mean Absolute Error(MAE) represents average error. Mean Squared Error(MSE) is square of absolute error, which exaggerate noise and punishes the larger errors.