**ASSIGNMENT 3**

**MACHINE LEARNING**

**In Q1 to Q11, only one option is correct, choose the correct option:**

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?

**Answer: -** A) Least Square Method

1. Which of the following statement is true about outliers in linear regression?

**Answer: -** A) Linear regression is sensitive to outliers

1. A line falls from left to right if a slope is \_\_\_\_\_\_?

**Answer: -** B) Negative

1. Which of the following will have symmetric relation between dependent variable and independent variable?

**Answer: -** C) Both of them

1. Which of the following is the reason for over fitting condition?

**Answer: -** C) Low bias and high variance

1. If output involves label, then that model is called as:

**Answer: -** B) Predictive model

1. Lasso and Ridge regression techniques belong to \_\_\_\_\_\_\_\_\_?

**Answer: -** D) Regularization

1. To overcome with imbalance dataset which technique can be used?

**Answer: -** D) SMOTE

1. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses \_\_\_\_\_ to make graph?

**Answer: -** C) Sensitivity and Specificity

1. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

**Answer: -** B) False

1. Pick the feature extraction from below:

**Answer: -**

A) Construction bag of words from an email

B) Apply PCA to project high dimensional data

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

1. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

**Answer: -** A) We don’t have to choose the learning rate.

C) We need to iterate.

D) It does not make use of dependent variable.

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

1. Explain the term regularization?

**Answer: -** When we use regression models to train some data, there is a good chance that the model will overfit the given training dataset. Regularization helps sort this overfitting problem by restricting the degrees of freedom of a given equation. i.e., Simply reducing the number of degrees of a polynomial function by reducing their corresponding weights. In a Linear equation, we do not want huge weights/ coefficients as a small change in weight can make a large difference for the dependent variable (y), So, regularization constraints the weights of such features to avoid overfitting. To regularise the model, a shrinkage penalty is added to the cost function.

**Overfitted Model: -**

* The Data will work on the training data but won't work on the test data
* The training data has the same questions which we will learn and the test data has different questions.
* So, the success of training data is more but it follows the old model hence, we should avoid overfitted model.
* Because It works on the train data but it won't work on test data.
* Because we don't know which data is going under training phase which is going under testing phase.

Let's see different types of Regularization in regression: -

#### **LASSO** (Least Absolute Shrinkage and Selection Operator) **Regression** (L1 form)

#### **Ridge Regression** (L2 form)

#### **ELASTICNET** (Less Popular) (Combination of LASSO and Ridge)

Overfitting is rare and in case if overfitting happens we have these regularisation types......

1. Which particular algorithms are used for regularization?

**Answer: -** The particular algorithm that are used for regularization are

#### **LASSO** (Least Absolute Shrinkage and Selection Operator) **Regression** (L1 form)

#### **Ridge Regression** (L2 form)

1. **ELASTICNET** (Less Popular) (Combination of LASSO and Ridge)

The Lasso and Ridge Regression are the main algorithms used for regularization

#### **LASSO (Least Absolute Shrinkage and Selection Operator) Regression (L1 form)**

LASSO Regression penalizes the model based on the sum of Magnitude of the Coefficients. LASSO gives Zero importance to the unwanted features in the dataset.

#### **Ridge Regression (L2 form)**

Ridge Regression penalises the model based on the Sum of Squares of Magnitude of the coefficients. Ridge Regression gives very less (1/2%) importance to the features in the dataset.

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

**Answer: -** An error term in statistics is a value which represents how observed data differs from actual population data.

It can also be a variable which represents how a given statistical model differs from reality. Although the terms error and residual are often interchanged, there is an important formal difference. While an error term represents the way observed data differs from the actual population, a residual represents the way observed data differs from sample population data. This means that a residual is often much easier to quantify. Although an error is generally unobservable, a residual is observable.