Machine learning Worksheet-1
Answers

Q.1.Ans:- D) Both A and B

Q.2.Ans:- A) Linear regression is sensitive to outliers

Q.3.Ans:- B) Negative

Q.4.Ans:- B) Correlation

Q.5.Ans:- C) Low bias and high variance

Q.6.Ans:- B) Predictive model

Q.7.Ans:- D) Regularization

Q.8.Ans:- D) SMOTE

Q.9.Ans:- A) TPR and FPR

Q.10.Ans:- B) False

Q.11.Ans:- A) Construction bag of words from a email.

In Q12, more than one options are correct

Q.12.Ans:- A) We don't have to choose the learning rate.
B) It becomes slow when number of features is very large.

Q13 and Q15 are subjective answer

Q.13.Ans:- Regularization is a technique used in machine learning and statistics to prevent overfitting and improve the generalization performance of models. The core idea behind regularization is to add a penalty term to the objective function being optimized during the model training process. This penalty term discourages the model from fitting the training data too closely, thus promoting simpler models that are less likely to overfit.

There are various types of regularization techniques, but two of the most common ones are:

1. L1 Regularization (Lasso):

- **In L1 regularization, a penalty term proportional to the absolute value of the coefficients is added to the objective function.**
- **This regularization technique encourages sparsity in the model, meaning it tends to drive some coefficients to exactly zero, effectively performing feature selection.**
- **L1 regularization is particularly useful when dealing with high-dimensional data or when there is a suspicion that many features are irrelevant.**

2. L2 Regularization (Ridge):

- **In L2 regularization, a penalty term proportional to the square of the coefficients is added to the objective function.**
- **This regularization technique penalizes large coefficients more heavily than small ones, effectively shrinking the coefficients towards zero.**

- L2 regularization helps to reduce the variance of the model by smoothing out the parameter space and preventing extreme parameter values.
- It is particularly useful when multicollinearity is present among the features or when the number of features is not excessively large.

Regularization helps to find a balance between fitting the training data well and maintaining model simplicity, thereby improving the model's ability to generalize to unseen data. It is a crucial tool in machine learning for building robust and reliable models, especially when dealing with complex datasets or limited training data.

Q.14.Ans:- Regularization techniques can be applied to a variety of machine learning algorithms to prevent overfitting. Some of the common algorithms that can incorporate regularization include:

1. Linear Regression:
   - Ridge Regression (L2 regularization)
   - Lasso Regression (L1 regularization)
   - Elastic Net Regression (Combination of L1 and L2 regularization)

2. Logistic Regression:
   - Ridge Logistic Regression (L2 regularization)
   - Lasso Logistic Regression (L1 regularization)
   - Elastic Net Logistic Regression (Combination of L1 and L2 regularization)

3. Support Vector Machines (SVM):
   - SVM with L2 regularization (often referred to as C parameter)

- SVM with L1 regularization (using the "penalty" hyperparameter)

4. Neural Networks:
   - Weight decay (L2 regularization)
   - L1 regularization
   - Dropout regularization
   - Batch normalization (indirect regularization effect)

5. Decision Trees and Ensemble Methods:
   - Pruning techniques to prevent overfitting (indirect regularization)
   - Regularized versions of ensemble methods like Regularized Random Forest, Gradient Boosting with regularization (e.g., XGBoost, LightGBM, CatBoost)

6. Generalized Linear Models (GLMs):
   - Regularized GLMs such as regularized logistic regression, regularized Poisson regression, etc.

Regularization techniques can be applied to many other algorithms as well, either through built-in parameters or by manually adding penalty terms to the objective function during training. The choice of which regularization technique to use depends on the specific characteristics of the dataset and the problem at hand.

Q.15.Ans:- In the context of linear regression, the term "error" refers to the discreoancy between the observed values of the dependent variable and the values predicted by the linear regression model.

The linear regression equation typically takes the form:

y = $\beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \ldots + \beta_n\, x_n + \varepsilon$

Where:

- **y is the dependent variable**
- **x_1, x_2, …, x_n are the independent variables**
- **β_0 , β_1, β_2, …, β_n are the coefficients or parameters estimated by the regression model,**
- **ε represents the error term.**

The error term captures the difference between the actual observed values of the dependent variable and the values predicted by the linear regression model. It represents the inherent variability in the data that cannot be explained by the linear relationship between the independent and dependent variables. In other words, it accounts for factors other than the predictors included in the model that influence the outcome.

The goal of linear regression is minimize the error term ε by estimating the coefficients β_0 , β_1, …, β_n that best fit the observed data. This is typically done by choosing coefficients that minimize the sum of squared differences between the observed and predicted values, a method known as ordinary least squares regression.