

MACHINE LEARNING WORKSHEET-1

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

Ans:

a) Least Square Error

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

Ans:

A) Linear regression is sensitive to outliers

3. A line falls from left to right if a slope is _____?

Ans:

a) Negative

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

Ans:

b) Correlation

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

Ans:

c) Low bias and high variance

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

Ans:

b) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?

Ans:

d) Regularization

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

Ans:

d) SMOTE

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

Ans:

a) TPR and FPR

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

Ans:

b) False

11. Pick the feature extraction from below:

Ans:

c) Apply PCA to project high dimensional data

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

Ans:

a) We don't have to choose the learning rate. b) It becomes slow when number of features is very large. c) We need to iterate.

13. Explain the term regularization?

Ans:

Regularization is a technique used to calibrate machine learning models in order to minimize the adjusted loss function by fitting the function appropriately on the given training set and avoid overfitting or underfitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

Sometimes the machine learning model is able to deliver accurate results on training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data (i.e. different type of data which is not same with training data) and hence the model is called overfitted. Regularization helps to sort this overfitting problem by restrict the degrees of freedom of the given equation. This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables or corresponding weights. Hence, it maintains accuracy as well as a generalization of the model. Actually, regularization technique help the machine to learn more than just memorize.

14. Which particular algorithms are used for regularization?

Ans:

There are three main regularization techniques, namely:

- i. LASSO / L1 form
- ii. Ridge Regression/ L2 form
- iii. Elastic-Net regression (Less popular then previous two)

LASSO means Least Absolute Shrinkage and Selection Operator. LASSO penalizes the model based on the sum of magnitude of the coefficient. The regularization term is given by:

Regularization = $\lambda * |\beta_j|$, Where λ = Shrinkage Factor

As loss function only considers absolute weights, optimization algorithms penalize higher weight values. It eliminates less important features and sets respective weight values to zero. Thus, LASSO also performs feature selection along with regularization.

Ridge Regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions and reduce the complexity of the model. The regularization term is given by:

Regularization = $\lambda * |\beta_j|^2$

We can calculate it by multiplying with the λ to the squared weight of each individual feature. It is mostly used to reduce the overfitting in the model, and it includes all the features present in the model. It reduces the complexity of the model by shrinking the coefficients of the less important features which contribute very less in the model but have huge weights, very close to zero. But it never make them zero.

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

Ans:

The error term is the sum of the deviations of each actual observation from a model regression line. It is the term in a regression equation that tallies up and accounts for the difference between the actually observed values of the independent variable and the results predicted by the model. If there were no error, all the data points would be located on the regression line. Hence, the error term is a measure of how accurately the regression model reflects the actual relationship between the independent and dependent variable or variables. It can indicate either that the model can be improved, such as by adding in another independent variable that explains some or all of the difference, or by randomness, meaning that the dependent and independent variable or variables are not correlated to any greater degree.

There are actually two types of error terms commonly used in regression analysis: absolute error and relative error. Absolute error is the difference between the actually observed values of the independent variable and the results predicted by the model. In other hand relative error is defined as the absolute error divided by the exact value predicted by the model.