UNIVERSIDAD POLITÉCNICA DE YUCATÁN

MACHINE LEARNING

UNIT 1

SOLUTION TO MOST COMMON PROBLEMS IN ML

TEACHER:

VICTOR ALEJANDRO

STUDENT:

ARI ISAÍAS QUINTAL VARGUEZ

07/09/2023

Overfitting

When a machine learning model predicts outcomes accurately for training data but not for new data, this undesired tendency is known as overfitting.

To solve the problem of overfitting, we can use different techniques.

Train with more data, feature selection, simplify data, regularization and early stopping are some of the most common techniques. These are used to by choosing the best features or removing useful features. Others work by just stopping the training when the number of epochs is high.

Underfitting

A data model that is underfitted has a high error rate on both the training set and unobserved data because it is unable to effectively represent the relationship between the input and output variables.

Some techniques to solve underfitting are increasing the model complexity, increasing the features of the model, and removing the noise from the data.

Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

Data points that stand out from the rest of the dataset are known as outliers.

The data distribution is frequently skewed by these anomalous observations, which are frequently the result of inaccurate observations or incorrect data entry.

Some solutions for outliers are setting up a filter in the testing tool, removing the outliers during post analysis or changing the value of outliers.

Dimensionality problem

Massive data collection may result in the dimensionality problem, where excessively noisy dimensions with few pieces of information and little discernible advantage can be achieved.

The primary cause of the dimensionality curse is the increasing nature of spatial volume.

Dimensionality reduction process

This technique is used to reduce the number of features in a dataset while preserving as much crucial data as feasible. In other words, it is a method for converting high-dimensional data into a lower-dimensional space while keeping the original data's essential characteristics.

Bias-variance trade-off.

The bias is known as the difference between the prediction of the values by the Machine Learning model and the correct value. The variability of model prediction for a given data point which tells us the spread of our data is called the variance of the model. The balance between an estimator's bias and variance is known as the "bias-variance tradeoff," and it is a key idea in machine learning and statistics.