UNIVERSIDAD POLITECNICA DE YUCATAN

UNIT

MACHINE LEARNING

TEACHER:

Victor Alejandro Ortiz

by:

Mónica Hernández Alamilla

DUE DATE:

September 15, 2023

Overfitting & Underfitting:

Overfitting

Overfitting occurs when a machine learning model learns the training data too well to the extent that it captures noise or random fluctuations in the data, rather than just the underlying patterns. This results in a model that performs well on the training data but poorly on unseen or new data.

Underfitting

Underfitting is the opposite problem. It occurs when a model is too simple to capture the underlying patterns in the data. An underfit model performs poorly both on the training data and on new data because it fails to capture the relationships present in the data.

Outliers

Outliers are data points that significantly deviate from the rest of the data in a dataset. They can be caused by measurement errors, data corruption, or rare events. Outliers can distort statistical analyses and machine learning models because they don't conform to the expected patterns of most of the data.

Solutions for Overfitting, Underfitting, and Outliers

Overfitting Solutions

Techniques to combat overfitting include using simpler models, increasing the amount of training data, applying regularization methods (e.g., L1 or L2 regularization), and using techniques like cross-validation.

Underfitting Solutions

To address underfitting, you can use more complex models, increase the model's capacity, or engineer better features. Collecting more relevant data can also help.

Outlier Solutions

Dealing with outliers involves identifying and removing or transforming them. Techniques such as z-score or IQR-based outlier detection, robust statistical methods, or using anomaly detection algorithms can be effective.

Dimensionality Problem

The dimensionality problem arises when datasets have a large number of features or dimensions. High-dimensional data can be challenging to work with because it can lead to increased computational complexity, a risk of overfitting, and difficulty in visualizing and interpreting the data.

Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of features in a dataset while preserving as much relevant information as possible. This can be achieved through techniques such as Principal Component Analysis (PCA) or feature selection methods.

Bias-Variance Trade-off

The bias-variance trade-off is a fundamental concept in machine learning. It refers to the balance between two types of errors that a model can make:

- Bias (Underfitting): High bias occurs when a model is too simple and cannot capture the underlying patterns in the data. This leads to systematic errors.
- Variance (Overfitting): High variance occurs when a model is too complex and fits the noise in the data. This leads to errors due to sensitivity to small fluctuations.

The trade-off involves finding the right level of model complexity to minimize the total error, which includes both bias and variance. This is often achieved through techniques like regularization and cross-validation.

REFERENCES

- *Alpaydin, E. (2014). Introduction to Machine Learning. The MIT Press.*
- *Witte, R. S., & Witte, J. S. (2019). Statistics. Wiley.*
- *Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.*