Universidad Politécnica de Yucatán

Machine Learning

Teacher:

Victor Ortiz

Student:

Esau Alberto May Ceh

Activity

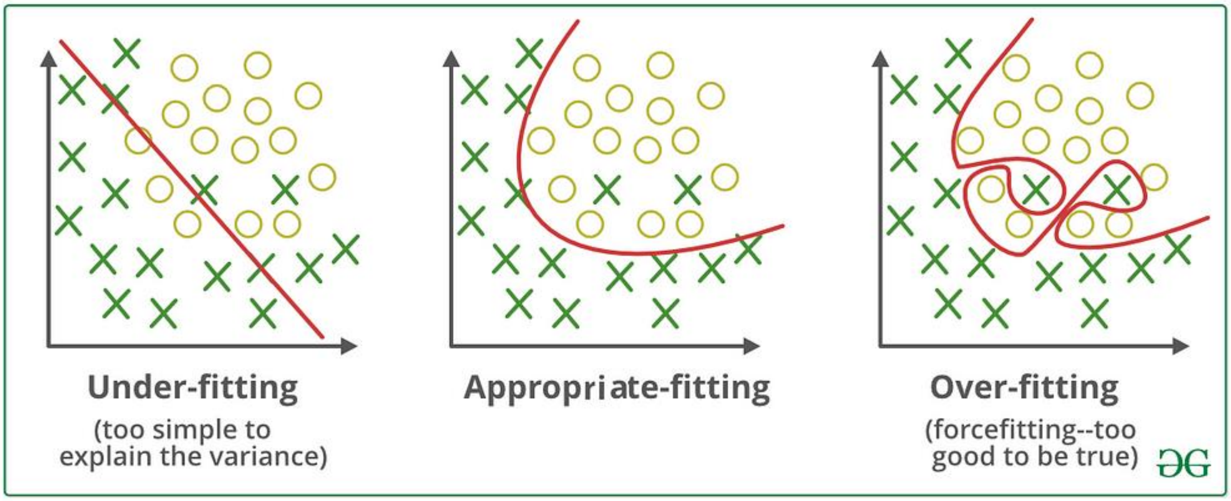
Solution to most common problems in Machine Learning

09/15/2023

Overfitting and Underfitting

Overfitting frequently occurs in machine learning when a model attempts to excessively scrutinize and adapt to the training data, leading it to memorize all the data patterns. This behavior can cause a predictive algorithm to produce inaccurate results with low accuracy, especially when making predictions with high variability.

Underfitting arises from overly simplistic representation of the input data in the model, leading to a decline in data accuracy. In cases of underfitting, a module is unable to adequately represent the dataset or generalize to new datasets. Similarly, the model cannot establish a meaningful relationship between the input variables and the target variables.


Under-fitting
(too simple to explain the variance)
Appropriate-fitting
Over-fitting
(forcefitting--too good to be true)

Outliers

An outlier is a data point that deviates significantly from its nearest neighbors and the surrounding values in a data graph or dataset you are analyzing. Outliers are data points that exhibit extreme values, markedly distinct from the general trend within a dataset or graph.

Common solutions for overfitting, underfitting and presence of outliers in datasets

Overfitting: Divide our data into training, validation, and testing, obtain a larger amount of data, adjust the parameters of our models, use simpler models, the data come from different distributions, lower the number of iterations in iterative algorithms.

Underfitting: Treat data correctly, eliminating outliers and unnecessary variables, use more complex models, adjust the parameters of our models, increase iterations in iterative algorithms.

Outliers: Trimming or remove outliers, replace the outliers with the median value, reduce the weight of outliers, changing the values, use robust estimation techniques.

Dimensionality Problem

Dimensionality refers to the challenges encountered when dealing with high-dimensional data, where dimension signifies the number of attributes or features in a dataset. High-dimensional data often exceed a hundred or more features. These challenges arise when attempting to analyze or visualize data to detect patterns, as well as during the training of machine learning models. The issues associated with training machine learning models on high-dimensional data are commonly termed the 'Curse of Dimensionality'.

Dimensionality reduction involves the act of reducing the quantity of attributes within a dataset while retaining the maximum amount of original dataset variability. This constitutes a data preprocessing step, implying that we conduct dimensionality reduction prior to model training."

Bias-variance trade-off

It means that if our model is overly simplistic with few parameters, it can result in high bias and low variance. Conversely, a model with numerous parameters tends to exhibit high variance and low bias. Therefore, striking the correct balance is crucial to avoid both overfitting and underfitting.

This trade-off in complexity illustrates the inherent balance between bias and variance. An algorithm cannot simultaneously be more complex and less complex.

References

Bonthu, H. (2023, June 26). *Detecting and treating outliers: Treating the odd one out!*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/#2e43

Great Learning Team. (2022, December 13). *Understanding curse of dimensionality*. Great Learning Blog: Free Resources what Matters to shape your Career! https://www.mygreatlearning.com/blog/understanding-curse-of-dimensionality/

KeepCoding. (2023, April 27). *Diferencias: Underfitting vs Overfitting*. KeepCoding Bootcamps. https://keepcoding.io/blog/underfitting-vs-overfitting/

Lemonaki, D. (2021, August 24). *What is an outlier? definition and how to find outliers in statistics*. freeCodeCamp.org. https://www.freecodecamp.org/news/what-is-an-outlier-definition-and-how-to-find-outliers-in-statistics/

Outlier detection and Treatment - World Bank. (n.d.). https://thedocs.worldbank.org/en/doc/20f02031de132cc3d76b91b5ed8737d0-0050012017/related/lecture-12.pdf

Pramoditha, R. (2023, August 24). *11 dimensionality reduction techniques you should know in 2021*. Medium. https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b

Rubiales, A. (2020, June 26). *¿Qué es underfitting y overfitting?*. Medium. https://rubialesalberto.medium.com/qu%C3%A9-es-underfitting-y-overfitting-c73d51ffd3f9

Singh, S. (2018, October 9). *Understanding the bias-variance tradeoff*. Medium. https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229