UNIVERSIDAD
POLITÉCNICA
DE YUCATÁN

BIS
Universities

UNIVERSIDAD POLITÉCNICA DE YUCATÁN

COMPUTATIONAL ROBOTICS ENGINEERING

GROUP: 9° “B”

PROFESSOR: ORTIZ VICTOR

SUBJECT: MACHINE LEARNING

STUDENT: JOSEPH JESUS AGUILAR RODRIGUEZ

Table of Contents

Define the concepts of: Overfitting & Underfitting. 3

What is Overfitting? 3

What is underfitting? 3

Underfitting vs. overfitting 3

Define and distinguish the characteristics of outliers. 4

Discuss the most common solutions for overfitting, underfitting and presence of outliers in datasets. 4

What is the solution for underfitting and overfitting? 4

How do you deal with outliers in a dataset? 4

Describe the dimensionality problem. 5

What is dimensionality problem? 5

Describe the dimensionality reduction process. 5

Explain the bias-variance trade-off. 5

Why is Bias Variance Tradeoff? 5

Works Cited 6

Define the concepts of: Overfitting & Underfitting.

What is Overfitting?

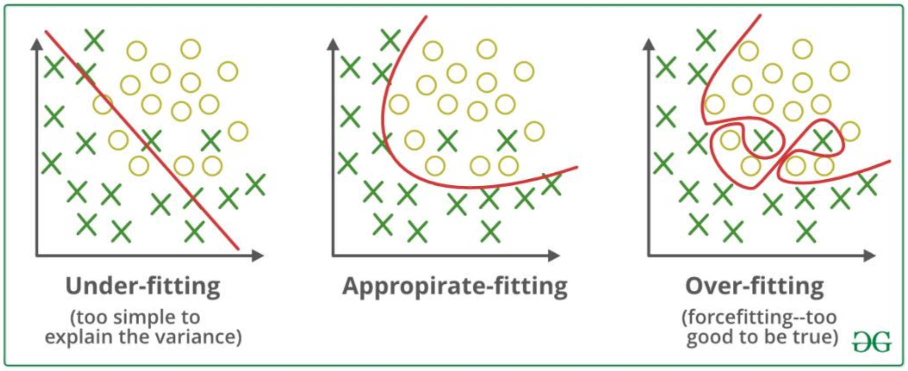
Overfitting is an undesirable machine learning behavior that occurs when the machine learning model gives accurate predictions for training data but not for new data. When data scientists use machine learning models for making predictions, they first train the model on a known data set. Then, based on this information, the model tries to predict outcomes for new data sets. An overfit model can give inaccurate predictions and cannot perform well for all types of new data.

What is underfitting?

Underfitting is another type of error that occurs when the model cannot determine a meaningful relationship between the input and output data. You get underfit models if they have not trained for the appropriate length of time on a large number of data points.

Underfitting vs. overfitting

Underfit models experience high bias—they give inaccurate results for both the training data and test set. On the other hand, overfit models experience high variance—they give accurate results for the training set but not for the test set. More model training results in less bias but variance can increase. Data scientists aim to find the sweet spot between underfitting and overfitting when fitting a model. A well-fitted model can quickly establish the dominant trend for seen and unseen data sets. (aws amazon, 2022)

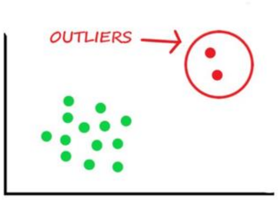

Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)
GG

Define and distinguish the characteristics of outliers.

Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

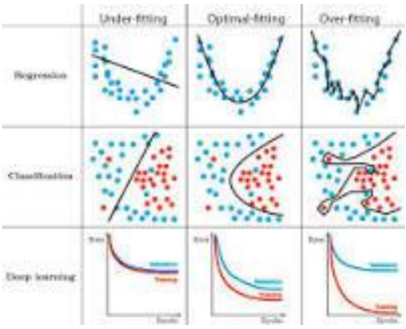
To ensure that the trained model generalizes well to the valid range of test inputs, it's important to detect and remove outliers.

In this guide, we'll explore some statistical techniques that are widely used for outlier detection and removal. (C, 2022)



Discuss the most common solutions for overfitting, underfitting and presence of outliers in datasets.

What is the solution for underfitting and overfitting?



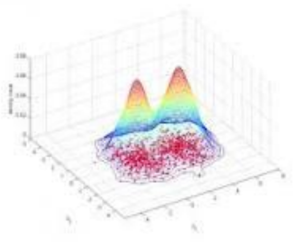
In this situation, the best strategy is to increase the model complexity by either increasing the number of parameters of your deep learning model or the order of your model. Underfitting is due to the model being simpler than needed. It fails to capture the patterns in the data. (Minhas, 2021)

How do you deal with outliers in a dataset?

There are different approaches such as replacing the outlier with the mean value, or median value or in some cases dropping the observation with the suspected outlier so as to avoid any bias in them. We tend to delete the outlier if they are due to data entry errors caused due to human error, data processing errors. (Snehal_bm, 2021)

Describe the dimensionality problem.

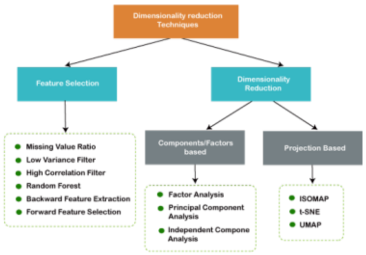
What is dimensionality problem?



Gathering a huge number of data may lead to the dimensionality problem where highly noisy dimensions with fewer pieces of information and without significant benefit can be obtained due to the large data. The exploding nature of spatial volume is at the forefront is the reason for the curse of dimensionality. (Choudhury, 2019)

Describe the dimensionality reduction process.

Dimensionality reduction simply refers to the process of reducing the number of attributes in a dataset while keeping as much of the variation in the original dataset as possible . It is a data preprocessing step meaning that we perform dimensionality reduction before training the model. (Pramoditha, 2021)



Explain the bias-variance trade-off.

Why is Bias Variance Tradeoff?

If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time. (Singh, 2018)

Works Cited

(s.f.). Obtenido de https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b#:~:text=Dimensionality%20reduction%20simply%20refers%20to,reduction%20before%20training%20the%20model.

aws amazon. (1 de january de 2022). *aws*. Obtenido de aws: What is Overfitting? - Overfitting in Machine Learning Explained - AWS. (n.d.-b). Amazon Web Services, Inc. https://aws.amazon.com/what-is/overfitting/#:~:text=Underfitting%20vs.,not%20for%20the%20test%20set.

C, B. P. (5 de July de 2022). *freecodecamp*. Obtenido de freecodecamp: https://www.freecodecamp.org/news/how-to-detect-outliers-in-machine-learning/#:~:text=Outliers%20are%20those%20data%20points,data%20entry%2C%20or%20erroneous%20observations.

Choudhury, A. (22 de May de 2019). *analyticsindiamag*. Obtenido de Mapping the Zeitgeist: https://analyticsindiamag.com/curse-of-dimensionality-and-what-beginners-should-do-to-overcome-it/#:~:text=Gathering%20a%20huge%20number%20of,for%20the%20curse%20of%20dimensionality.

Minhas, M. S. (5 de June de 2021). *towardsdatascience*. Obtenido de towardsdatascience: https://towardsdatascience.com/techniques-for-handling-underfitting-and-overfitting-in-machine-learning-348daa2380b9

Pramoditha, R. (13 de April de 2021). *towardsdatascience*. Obtenido de towards data science: https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b#:~:text=Dimensionality%20reduction%20simply%20refers%20to,reduction%20before%20training%20the%20model.

Singh, S. (20 de May de 2018). *towardsdatascience*. Obtenido de towardsdatascience: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Snehal_bm. (8 de July de 2021). *analyticsvidhya*. Obtenido de analyticsvidhya: https://www.analyticsvidhya.com/blog/2021/07/how-to-treat-outliers-in-a-data-set/#:~:text=There%20are%20different%20approaches%20such,human%20error%2C%20data%20processing%20errors.