Solution to most common problems in ML

UNIVERSIDAD POLITÉCNICA DE YUCATÁN

BIS Universities

Machine Learning.

Professor: Victor Alejandro Ortiz Santiago.

Porfolio Evidence.

Erubiel Tun Moo (2009137)

Date: 15/09/2023

Solution to most common problems in ML

I. Overfitting & Underfitting concepts

Overfitting and Underfitting are the most common problems within Machine Learning and the main causes of obtaining bad results in the data. When training the model, an attempt is made to fit the input data with each other and with the output, and it is in this process where These problems arise as overfitting or underfitting can occur. [1]

A. *Overfitting.*

It is called Overfitting when a model fits too much to the training data and this prevents it from generalizing well to new data; In other words, overfitting means that our training has focused so much on the particular training set that it has missed the point entirely. In this way, the model is not able to adapt to new data, since it is too focused on the training set. [2]

B. *Underfitting.*

On the other hand, we have Underfitting, which occurs when a model is too simple and cannot capture the complexity of the training data in such a way that it produces a poor generalization. In simpler words, we can say in underfitting what happens is that the model He does not know what to do with the task we have given him and therefore provides an answer that is far from correct. [2]

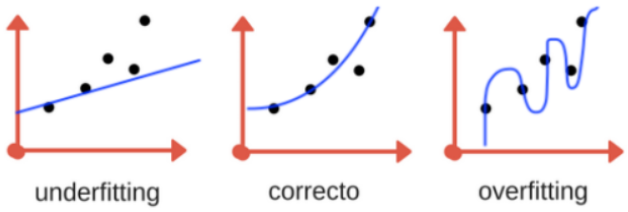

Fig. 1. Overfitting and Underfitting.

II. Define and distinguish the characteristics of outliers.

Outliers are data points in the data set where there are abnormal observations among the normal observations and can lead to strange precision scores that can bias the measurements as the results do not present the actual results. [3]

These can be caused by a variety of factors, such as measurement errors, data entry errors, or simply unusual patterns in the data. If outliers are caused by measurement or data entry errors, they can distort the results of the analysis and lead to to erroneous conclusions. On the other hand, if outliers are caused by interesting patterns in the data, they can provide valuable and relevant information for the analysis [4]; To detect outliers, there are several techniques that can be used in exploratory data analysis. Some of these techniques include: Data visualization, statistical analysis and also some Machine Learning techniques. Outliers can be useful as they can mean several things:

- Error: This refers to when some of our data is incorrect and in this case the outlier is quite useful as it helps us detect that error.
- Limits: In other cases, we may have values that fall outside the "middle group", but we want to keep the modified data so that it does not harm the learning of the ML model.
- Points of interest: This is something very useful since sometimes what we want is to observe or detect anomalous data, in these cases the outliers are of great help to achieve our objective. [5]

On the other hand, these can distort descriptive statistics, such as the mean and standard deviation. The mean tends to be affected by outliers, while the standard deviation can increase significantly due to the variability introduced by outliers. Something that can help us identify outliers is the visualization of the data and graphs such as the box plot. The box plot and scatter plot can reveal the presence of outliers in a data set.

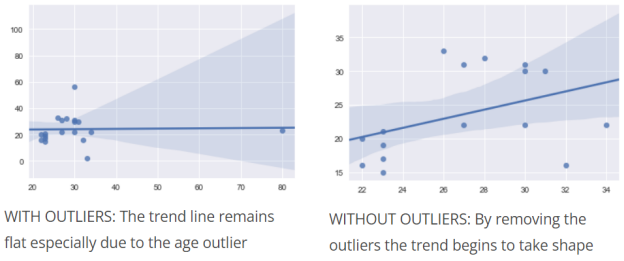

Fig. 2. Outliers.

III. Solutions for Overfitting, Underfitting and presence of outliers in datasets.

A. *For Overfitting*

- Model simplification: The first step is to reduce the complexity of the model, for example in a neural network you can reduce or reduce the number of layers or neurons, as well as use regularization techniques such as dropout or early stop.
- Eliminate noise from the training set: Because overfitting can be caused by poor data cleaning, it is very important to perform data cleaning when we receive the data in order to eliminate outliers, standardize the data and delete information that may cause noise to our modeling.
- Get more observations: Getting more data could help solve the problem. However, it is also possible that this

is not the case and another methodology will have to be used. [6]

B. For Underfitting

- Decrease regularization: Regularization is typically used to reduce the variance of a model by applying a penalty to the input parameters with the largest coefficients. There are several different methods such as L1 regularization, Lasso regularization, dropout, etc. that help reduce noise and outliers within a model. [7]
- Increase the duration of training: Some of the problems that cause underfitting are training time, when you stop training too soon this problem occurs, therefore it is necessary to extend the duration of training to avoid underfitting. However, It is important to be aware of overtraining and subsequently overfitting. Finding the balance between both scenarios will be key.
- More complex models: Use more complex models with more parameters, such as increasing the depth of a decision tree or using larger neural networks.

C. Ouliers.

- Trimming: It excludes the outlier values from our analysis. By applying this technique, our data becomes thin when more outliers are present in the dataset. Its main advantage is its fastest nature.
- Capping: In this technique, wecap our outliers data and make the limit i.e, above a particular value or less than that value, all the values will be considered as outliers, and the number of outliers in the dataset gives that capping number.
- Discretization: In this technique, by making the groups, we include the outliers in a particular group and force them to behave in the same manner as those of other points in that group. This technique is also known as Binning. [8]

IV. Describe the dimensionality problem.

This problem is known as the curse of dimensionality which occurs in machine learning when data sets with many features are handled and as more features are added, the complexity of the model increases and the accuracy decreases. This is because the data becomes sparse in a high-dimensional space, making it difficult to identify patterns and relationships between features. Higher dimensions lead to equidistant separation between points. The larger the dimensions, the more difficult it is to take samples because sampling loses its randomness.

It becomes more difficult to collect observations if there are many features. These dimensions make all observations in the data set equidistant from all other observations. Clustering uses Euclidean distance to measure the similarity between observations. Meaningful groups cannot be formed if the distances are equidistant. [9]

V. Describe the dimensionality reduction process.

Dimensionality reduction is the process of reducing the number of entities in a data set while preserving as much information as possible. It is used to overcome the curse of dimensionality, which refers to the fact that high-dimensional data is often sparse and difficult to analyze. There are two main approaches to dimensionality reduction: feature selection and feature extraction. Feature selection involves selecting a subset of the original features, while feature extraction involves transforming the original features into a new set of features. A popular technique for feature extraction is Principal Component Analysis (PCA), which involves finding the linear combinations of the original features that capture the most variation in the data. These linear combinations are called principal components and can be used to represent the data in a lower dimensional space. Another popular technique for feature extraction is t-SNE, which is a nonlinear method that preserves local structure in the data. It is often used to visualize high-dimensional data in two or three dimensions. [10]

Other techniques could be; Factor Analysis, Independent Component Analysis, ISOMAP, CMAP.

VI. Explain the bias-variance trade-off.

The bias-variance tradeoff is an important aspect of machine/statistical learning.

All learning algorithms use a mathematical/statistical approach that contains an "error" term which can be further split into two components: reducible and irreducible error. As the name suggests, the Irreducible error is an inherent uncertainty associated with the model and is associated with a natural variability in a system. This cannot be reduced and nothing can be done about it. On the other hand, Reducible error, as the name suggests, can be and should be minimized further to maximize accuracy.

In supervised learning algorithms, this reducible error can be further decomposed into "error due to squared bias" and "error due to variance" The goal of the learning algorithm is to simultaneously reduce bias and variance in order to obtain an model that is the most feasible. However, achieving that is not so easy and in real life, there is a tradeoff to be made when selecting models of different flexibility or complexity and to minimize these sources of error!; Ideally a tilt towards either of them is not desired but while modelling real world problems, it is impossible to get rid of both of them at the same time. This is where the term "tradeoff" comes in. [11]

References

[1] Na, "Qué es overfitting y underfitting y cómo solucionarlo," Aprende Machine Learning, 12-Dec-2017. [Online]. Available: https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/. [Accessed: 15-Sep-2023].

[2] "Overfitting vs. Underfitting: What is the difference?," 365 Data Science, 27-Aug-2021. [Online]. Available: https://365datascience.com/tutorials/machine-learning-tutorials/overfitting-underfitting/. [Accessed: 15-Sep-2023].

[3] P. Nichani, "OutLiers in machine learning," Analytics Vidhya, 22-Apr-2020. [Online]. Available: https://medium.com/analytics-vidhya/outliers-in-machine-learning-e830b2bd8660. [Accessed: 15-Sep-2023].

[4] O. A. T. Briseño, “Análisis exploratorio de datos en Python: técnicas y herramientas esenciales para comprender tus datos,” Linkedin.com, 1678286523000. [Online]. Available: https://www.linkedin.com/pulse/an%C3%A1lisis-exploratorio-de-datos-en-python-t%C3%A9cnicas-y-tello-brise%C3%B1o/?originalSubdomain=es. [Accessed: 15-Sep-2023].
[5] Na, “Detección de outliers en Python,” Aprende Machine Learning, 02-Jun-2020. [Online]. Available: https://www.aprendemachinelearning.com/deteccion-de-outliers-en-python-anomalia/. [Accessed: 15-Sep-2023].
[6] R. Canadas, “Qué es el overfitting,” abdatum, 17-Aug-2021. .
[7] “What is underfitting?,” Ibm.com. [Online]. Available: https://www.ibm.com/topics/underfitting. [Accessed: 15-Sep-2023].
[8] C. Goyal, “Outlier detection removal,” Analytics Vidhya, 19-May-2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/. [Accessed: 15-Sep-2023].
[9] Sriram, “Curse of dimensionality in machine learning: How to solve the curse?,” upGrad blog, 25-Feb-2023. .
[10] “Introduction to dimensionality reduction,” GeeksforGeeks, 01-Jun-2017. [Online]. Available: https://www.geeksforgeeks.org/dimensionality-reduction/. [Accessed: 15-Sep-2023].
[11] R. Shankar, “Bias-Variance Tradeoff: What is it and why is it important?,” Linkedin.com, 1483476131000. [Online]. Available: https://www.linkedin.com/pulse/bias-variance-tradeoff-what-why-important-ravi-shankar/. [Accessed: 15-Sep-2023].