Machine Learning

Unit I

# Solution to most common problems in ML

Canul Cal Jorge Carlos

Merida, Yucatan

September 15th, 2023

**Overfitting and Underfitting**

The underfitting problem is presented when a statistical model or machine learning algorithm is extremely simple to capture the data needed to perform analysis. With that being said, the model will be unable to learn the patterns presented in the dataset, so the performance is going to be poor both on the training and testing data (dewangNautiyal).

However, underfitting is due to a model not complex enough to be capable of represent and interpret the given data. Yet, the model would not be the problem itself but the input data by not having the adequate representations of underlying factors influencing the target variable. Other possible situation is that the training data set.

On the other hand, the overfitting situation is that when the model learns the data to the very minimum detail and noise thus showing negative impacts in the performance of the model on new data so the model will be unable to generalize but to perfectly fit what it has learned along the way (Brownlee, J.).

When it comes to overfitting problem there are several techniques when evaluating to reduce the situation such as a resampling technique to estimate model accuracy, hold back a validation dataset, etc.

**Characteristics of outliers**

Outliers, or outlying observations, are values in data which appear aberrant or unrepresentative taking place in almost every single dataset, so it is common to deal with them. Yet, if a test shows the outlier to be inconsistent with the model, it is characterized as discordant. A discordant outlier may be identified as of intrinsic interest, rejected as a nuisance item, or incorporated in the main data via a revised model (Lewis, T.). Outliers can be problematic when they exert a disproportionate influence on the analysis or modeling results such as intercepting the regression line, distorting the relationship between variables, or impact in the clustering algorithms by affecting the distance metrics and the formation of clusters (Firdose, T.).

**Common solutions for overfitting, underfitting and presence of outliers in dataset**

The common solutions when dealing with overfitting issues is having enough data that represents the diversity and complexity of the problem. For example, when the model fails to generalize to new data it is highly recommended to add more data in the training because exists the possibility that the current observations does not actually represent the data required for the analysis, so richer and more diverse data, according to what it is needed, should improve the performance of the model (Chemama, J.). Another possible solution is removing features from the data by interpreting which data could be irrelevant in the model training.

Meanwhile, underfitting occurs due to high bias and low variance, so the problem can be solved by increasing the number of features in the dataset (Biswal, A), increasing the model complexity so it can detect the characteristics of the observations, reducing the noise in the data or increasing the duration of training data (add more observations to make predictions).

In the field of outliers, one of the most common solutions are just erase the outliers if they are not part of the population that is going to be studied. For example, if you are studying the relationship between income and health in the American population, a musical artists would be an outlier that is clearly not part of the intended population (Lee, G.). Another possibility is to transform the data to see if it removes the outlier or significantly changes how much it skews the data.

**Dimensionality problem**

Normally known as 'Curse of Dimensionality', it is defined as: as the number of dimensions or features increases, the amount of data needed to generalize the machine learning model accurately increases exponentially. The increase in dimensions makes the data sparse, and it increases the difficulty of generalizing the model. More training data is needed to generalize that model better (Sriam).

**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 (Pramoditha, R.). The dimensionality reduction is a pre-process of the data before training

the model. However, when reducing the dimensionality of the dataset it will be lost some percentage of the variability in the original data. It can be performed by applying feature selection that, in simple words, consists in selecting a subset of the original features that are most relevant to the problem, so the ultimate goal is to preserve the critical features. Then, in the feature extraction process new features are created by combining or transforming the original ones, so the essence will be captured in a lower-dimensional space (GeeksforGeeks). For example, the methods are principal component analysis (PCA), linear discriminant analysis (LDA), and t-distributed stochastic neighbor embedding (t-SNE). PCA is a popular technique that projects the original features onto a lower-dimensional space while preserving as much of the variance as possible.

**Bias-variance trade-off**

The bias in machine learning refers to the difference between a model's prediction and the actual distribution of the value it tries to predict (Logunova, I. & Khaciyants, A.). Models with high bias oversimplify the data distribution rule/function, resulting in high errors in both the training outcomes and test data analysis results. In other words, the bias is a systematic error which takes place due to incorrect assumptions in the machine learning process, leading to the misinterpretation of data distribution, indicating that the model does not accurately represent the data that is trying to be predicted.

## References


Biswal, A. (2021, April 27). *The complete guide on overfitting and underfitting in machine learning*. Simplilearn. Retrieved from: https://www.simplilearn.com/tutorials/machine-learning-tutorial/overfitting-and-underfitting

Brownlee, J. (2016, March 21). *Overfitting and Underfitting With Machine Learning Algorithms*. Machine Learning Mastery. Retrieved from: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Chemama, J. (2020, March 10). *How to solve underfitting and overfitting data models*. AllCloud. Retrieved from: https://allcloud.io/blog/how-to-solve-underfitting-and-overfitting-data-models/

Firdose, T. (2023, June 1). *Understanding outliers: Impact, detection, and remedies.* Medium. Retrieved from: https://tahera-firdose.medium.com/understanding-outliers-impact-detection-and-remedies-ea2192174477

Follow, D. (2017, November 23). *ML | Underfitting and Overfitting.* GeeksforGeeks. Retrieved from: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

GeeksforGeeks (2017, June 1). *Introduction to dimensionality reduction*. GeeksforGeeks. Retrieved from: https://www.geeksforgeeks.org/dimensionality-reduction/

Lee, G. (2022, September 28). *4 easy ways to handle outliers in your data*. Medium. Retrieved from: https://medium.com/mlearning-ai/4-easy-ways-to-handle-outliers-in-your-data-47f125a3f779

Lewis, T. (2015). Statistical analysis, special problems of: Outliers. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 387–391). Elsevier.

Logunova, I., & Khaciyants, A. (2023, March 13). *Bias-Variance Tradeoff in Machine Learning*. Serokell Software Development Company. Retrieved from: https://serokell.io/blog/bias-variance-tradeoff

Pramoditha, R. (2021, April 14). 11 Dimensionality reduction techniques you should know in 2021. Medium. Retrieved from: https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b

Sriram. (2023, February 25*). Curse of dimensionality in machine learning: How to solve the curse?* upGrad. Retrieved September 15, 2023, from upGrad blog website: https://www.upgrad.com/blog/curse-of-dimensionality-in-machine-learning-how-to-solve-the-curse/