Universidad Politécnica de Yucatán

Student’s name:

Fernando Rodríguez Zapata

Teacher’s name: Victor Alejandro Ortiz Santiago

Subject: Machine Learning

Group: Computational Robotics 9° A

Date: September 15, 2023

Overfitting and underfitting

There are two common problems we can face in the training of our model, those are overfitting and underfitting, first, overfitting according to IBM is when a statistical model fits exactly against its training data. In other words, it is when the model instead of learning the general features of a dataset, it learns or memorize the noise, particular features, and characteristics of the data, so it can seem it has a good performance in the training stage, but when it is presented new data, it has a poor performance. It could not generalize well to predict but learned the dataset.

In the other hand, underfitting when a model is unable to capture the relationship between the input and output variables accurately, according to IBM. This means that is not able to perform well or predict with seen data nor unseen data, this is related to the simpleness of the model. So, the model is not ready to do regression or classification tasks.

Solutions for overfitting

Early stopping: A possible reason why the model is learning the dataset instead of generalizing, it is that the training is taking too long, letting the model to learn the noise, to avoid this, we need to reduce the duration of the training, not leading it to memorize the dataset particular features but generalizing the patterns.

Using a larger training dataset: Enhancing the dataset by adding more data points can improve the model's accuracy as it offers a broader perspective on the relationship between input and output variables. However, it is important to ensure that the additional data is of good quality and relevant.

Feature Selection: Feature selection is about retaining the most critical features from the training data while discarding those that are redundant or irrelevant.

Regularization: Regularization introduces a penalty on the magnitude of model parameters. This "penalty" usually targets the coefficients of the input parameters, especially when they become too large, as large coefficients can lead to sensitivity to slight variations in input data.

Solutions for underfitting

Increasing the duration of training: Unlike overfitting, underfitting could be a problem presented when the duration of training is short. To avoid this, we could give the model more time to learn and adapt to the dataset.

Feature selection: In overfitting the problem with the features leads to redundance and irrelevant features, with underfitting we need to add features with greater importance.

Reducing regularization: If the features in the data are overly uniform due to excessive regularization, the model might struggle to discern the primary patterns, resulting in underfitting. By lessening the regularization, we introduce more complexity and diversity into the model, facilitating a more effective training process.

Outliers

An outlier is an observation that appears far away and diverges from an overall pattern in a sample (Nichani, P.). Also, we could define outliers as data points in a dataset that stand out from typical observations, being some distance away from the mean of the samples. In the training process the presence of outliers can lead to misleading accuracy scores, affecting the true performance of a model or extending training durations.

They have three main characteristics, deviation, rarity and impact. Deviation refers to the abnormal divergence from the typical distribution of the data. While, rarity it is about their commonness, which is low, not representing the general pattern of the data; finally, impact means the influence of the outliers in the statistical analyses of the dataset, leading to skewed results or inaccurate conclusions.

Deal with Outliers

Remove the observations: It is possible to exclude outlying data points to prevent them from adversely affecting model training. When dealing with a small dataset, however, eliminating the observations is not recommendable.

Imputation: To impute the outliers, we can use a variety of imputation values, ensuring that no data is lost. As impute values, we can choose between the mean, median, mode, and boundary values.

Dimensionality problem and dimensional reduction

Also called "Curse of dimensionality", it is defined by Siriram as follows: "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." Consequently, high-dimensional datasets can lead to longer training times, increased computational demands, and a heightened risk of overfitting.

Dimensionality reduction refers to the method of condensing the number of features in a dataset while maintaining its crucial information. Essentially, it takes data from a high-dimensional space and represents it in a lower-dimensional format, ensuring the core characteristics of the original data remain intact.

There are two main approaches to dimensionality reduction: feature selection and feature extraction.

Feature Selection: It aims to choose a subset of the initial features deemed most pertinent to the task. The objective is to decrease the dataset's dimensionality by preserving only the most significant features. There exists a variety of techniques for this purpose, such as filter methods, wrapper methods, and embedded methods.

Feature Extraction: It aims to generate new features by merging or altering the initial features. This process intends to represent the original data's core characteristics within a more concise dimensional space. Various techniques facilitate feature extraction, such as principal component analysis (PCA), linear discriminant analysis (LDA), and t-distributed stochastic neighbor embedding (t-SNE).

Bias-variance tradeoff

It is important what bias and variance are in order to understand the tradeoff related to them.

Bias: Bias refers to the systematic error in predictions when a model's assumptions about the data are incorrect. It manifests as the difference between the expected or actual values and the values predicted by the model. This discrepancy, often termed as bias error or error due to bias, stems from inappropriate assumptions made during the machine learning process.

Variance: Variance quantifies the dispersion of data points from their average value. In the context of machine learning, variance describes how the performance of a predictive model fluctuates when trained on various subsets of the training data. Essentially, it indicates the model's sensitivity or variability in response to different portions of the training dataset.

While the tradeoff refers to the relation they have when we choose to lower bias which typically increases variance, or lower variance which typically increases bias. When a model's complexity rises, the total error typically decreases, but only to a specific threshold. Beyond that, the variance starts to grow, leading to a rise in the total error. Practically, our primary focus is on reducing the model's total error rather than specifically minimizing variance or bias. Achieving the lowest total error involves finding an optimal balance between variance and bias.

References

AL-somiri, B. (2022, November 18). outliers in data set: definition, Characteristics, formulas, examples. Retrieved September 16, 2023, from StatisticsLingo website: https://statisticslingo.com/outliers-in-data-set/

Follow, K. (2020, February 6). Bias and variance in machine learning. Retrieved September 16, 2023, from GeeksforGeeks website: https://www.geeksforgeeks.org/bias-vs-variance-in-machine-learning/

Introduction to dimensionality reduction. (2017, June 1). Retrieved September 16, 2023, from GeeksforGeeks website: https://www.geeksforgeeks.org/dimensionality-reduction/

Nichani, P. (2020, April 22). OutLiers in machine learning. Retrieved September 16, 2023, from Analytics Vidhya website: https://medium.com/analytics-vidhya/outliers-in-machine-learning-e830b2bd8660

Sahu, A. (2021, April 3). How to handle outliers in machine learning. Retrieved September 16, 2023, from Analytics Vidhya website: https://medium.com/analytics-vidhya/how-to-handle-outliers-in-machine-learning-5d8105c708e5

Sriram. (2023, February 25). Curse of dimensionality in machine learning: How to solve the curse? Retrieved September 16, 2023, from upGrad blog website: https://www.upgrad.com/blog/curse-of-dimensionality-in-machine-learning-how-to-solve-the-curse/

What is overfitting? (n.d.). Retrieved September 16, 2023, from Ibm.com website: https://www.ibm.com/topics/overfitting

What is underfitting? (n.d.). Retrieved September 16, 2023, from Ibm.com website: https://www.ibm.com/topics/underfitting

Zach. (2020, October 25). What is the Bias-Variance Tradeoff in Machine Learning? Retrieved September 16, 2023, from Statology website: https://www.statology.org/bias-variance-tradeoff/