Portfolio evidence

# Solution to Common Problems in Machine Learning

Rafael Marí Reyna

Robotics 9B

Due Date: September 15th, 2023

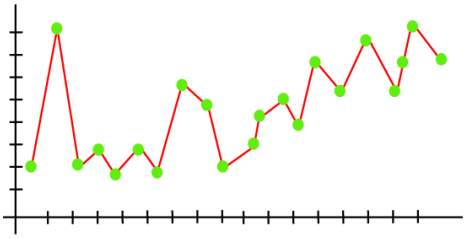
Teacher Victor Alejandro Ortiz

Machine Learning

**Overfitting**

Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

Overfitting occurs when our machine learning model tries to cover all the data points or more than the required data points present in the given dataset. Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model. The overfitted model has low bias and high variance. The chances of occurrence of overfitting increase as much we provide training to our model. It means the more we train our model, the more chances of occurring the overfitted model.



As we can see, the model tries to cover all the data points present. It may look efficient, but it is not. Because the goal of the regression model is to find the best fit line, but here we have not got any best fit, so, it will generate the prediction errors.

How to solve overfitting?

- Adding more data

It can be observed in a model when overfitting occurs, indicating a failure to generalize to new data. This suggests that the training data isn't truly representative of the data encountered in production. Training the algorithm on a larger, more varied, and diverse dataset might enhance its performance.

- Data augmentation

This encompasses techniques aimed at artificially expanding the size of a dataset by transforming existing data. For instance, with images, they can be flipped either horizontally or vertically, cropped, or rotated. They can also be converted to grayscale or have their color

saturation adjusted. From the algorithm's perspective, this is new data. Yet, not all transformations are universally applicable.

- Regularization

Regularization covers a broad spectrum of techniques, though not all are detailed here. The key takeaway is that these methods impose a "complexity penalty" on a model. To avoid this penalty, the model should emphasize the most conspicuous patterns, which are more likely to generalize effectively.

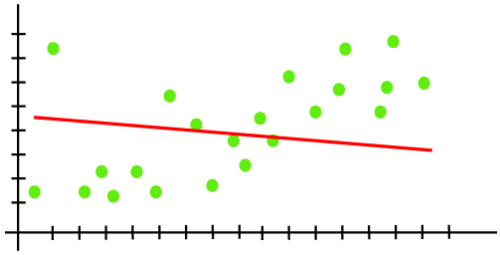
- Removing features from data

At times, a model might not generalize adequately because the training data is overly intricate, causing the model to overlook essential patterns. Eliminating some features to simplify the data might assist in curtailing overfitting.

**Underfitting**

Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data. To avoid the overfitting in the model, the fed of training data can be stopped at an early stage, due to which the model may not learn enough from the training data. As a result, it may fail to find the best fit of the dominant trend in the data. In the case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable predictions. An underfitted model has high bias and low variance.

A statistical model or a machine learning algorithm is said to have underfitting when a model is too simple to capture data complexities. It represents the inability of the model to learn the training data effectively resulting in poor performance both on the training and testing data.



As we can see from the diagram, the model is unable to capture the data points present in the plot.

How to solve overfitting?

- Increasing the model complexity

It can be observed in a model when underfitting happens, potentially because the model isn't sufficiently intricate to discern patterns in the data. Opting for a more complex model, like transitioning from a linear to a non-linear model or incorporating additional hidden layers in a neural network, can frequently address underfitting issues.

- Reducing regularization

The algorithms employed typically come with inbuilt regularization parameters designed to counteract overfitting. On occasions, these parameters might impede the algorithm's learning capabilities. Generally, lowering their values proves beneficial.

- Adding features to training data

Contrary to overfitting scenarios, underfitting in a model might arise due to overly simplistic training data. The data might not encompass the necessary features that would allow the model to discern pertinent patterns for accurate predictions. Enhancing the training data with additional features and complexity can aid in rectifying underfitting.

**Outliers**

Outlier is an observation that appears far away and diverges from an overall pattern in a sample. Outliers in input data can skew and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results. Outliers are datapoints in dataset in which are abnormal observations amongst the normal observations and can lead to weird accuracy scores which can skew measurements as the results do not present the actual results.

Detecting outliers:

- Data Visualization:

Visualization methods such as Distribution Curve, Box-plot, Histogram and Scatter Plot can be used to detect Outliers.

- Z-Score or Extreme Value Analysis:

The z-score or standard score of an observation is a metric that indicates how many standard deviations a data point is from the sample's mean, assuming a gaussian distribution.

$$z = \frac{x - \mu}{\sigma}$$

When computing the z-score for each sample on the data set a threshold must be specified.

- Clustering Methods:

In the K-Means clustering technique, each cluster has a mean value. Objects belong to the cluster whose mean value is closest to it. In order to identify the Outlier, firstly we need to initialize the threshold value such that any distance of any data point greater than it from its nearest cluster identifies it as an outlier for our purpose. Then we need to find the distance of the test data to each cluster mean. Now, if the distance between the test data and the closest cluster to it is greater than the threshold value then we will classify the test data as an outlier.

**Dimensionality Problem**

Coined by mathematician Richard E. Bellman, the curse of dimensionality references increasing data dimensions and its explosive tendencies. This phenomenon typically results in an increase in computational efforts required for its processing and analysis.

Regarding the curse of dimensionality, also known as the Hughes Phenomenon or Dimensionality Problem, there are two things to consider. On one hand, ML excels at analyzing data with many dimensions. Humans are not good at finding patterns that may be spread out across so many dimensions, especially if those dimensions are interrelated in counter-intuitive ways. On the other hand, as we add more dimensions we also increase the processing power we need to analyze the data, and we also increase the amount of training data required to make meaningful data models.

**Dimensionality Reduction**

Dimensionality reduction is a technique used to reduce the number of features in a dataset while retaining as much of the important information as possible. In other words, it is a process of transforming high-dimensional data into a lower-dimensional space that still

preserves the essence of the original data. Dimensionality reduction can help to mitigate these problems by reducing the complexity of the model and improving its generalization performance. There are two main approaches to dimensionality reduction: feature selection and feature extraction.

- Feature Selection:

Feature selection involves selecting a subset of the original features that are most relevant to the problem at hand. The goal is to reduce the dimensionality of the dataset while retaining the most important features. There are several methods for feature selection, including filter methods, wrapper methods, and embedded methods.

- Feature Extraction:

Feature extraction involves creating new features by combining or transforming the original features. The goal is to create a set of features that captures the essence of the original data in a lower-dimensional space. There are several methods for feature extraction, including principal component analysis (PCA), linear discriminant analysis (LDA), and t-distributed stochastic neighbor embedding (t-SNE).

**Bias-variance trade-off**

In machine learning, the bias-variance trade-off is a fundamental concept affecting the performance of any predictive model. It refers to the delicate balance between bias error and variance error of a model, as it is impossible to simultaneously minimize both. Striking the right balance is crucial for achieving optimal model performance.

Bias Variance Tradeoff is a design consideration when training the machine learning model. Certain algorithms inherently have a high bias and low variance and vice-versa. In this one, the concept of bias-variance tradeoff is clearly explained so you make an informed decision when training your ML models.

**References**


"Overfitting and Underfitting in Machine Learning - Javatpoint," www.javatpoint.com. https://www.javatpoint.com/overfitting-and-underfitting-in-machine-learning

"Underfitting and Overfitting in Machine Learning," GeeksforGeeks, Nov. 23, 2017. https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

"How to solve underfitting and overfitting?," AllCloud, Mar. 10, 2020. https://allcloud.io/blog/how-to-solve-underfitting-and-overfitting-data-models/

P. Nichani, "OutLiers in Machine Learning," Analytics Vidhya, Apr. 22, 2020. https://medium.com/analytics-vidhya/outliers-in-machine-learning-e830b2bd8660

"Machine Learning | Outlier," GeeksforGeeks, Jan. 12, 2019. https://www.geeksforgeeks.org/machine-learning-outlier/

"Curse of Dimensionality," Built In. https://builtin.com/data-science/curse-dimensionality

G. L. Team, "What is Curse of Dimensionality in Machine Learning?," GreatLearning Blog: Free Resources what Matters to shape your Career!, Oct. 01, 2020. https://www.mygreatlearning.com/blog/understanding-curse-of-dimensionality/

"Introduction to Dimensionality Reduction - GeeksforGeeks," GeeksforGeeks, Jun. 2017. https://www.geeksforgeeks.org/dimensionality-reduction/

"Bias Variance Tradeoff - Clearly Explained," ML+, Oct. 22, 2020. https://www.machinelearningplus.com/machine-learning/bias-variance-tradeoff/

"The Bias-Variance Trade-off in Machine Learning," Stack Abuse, May 23, 2023. https://stackabuse.com/the-bias-variance-trade-off-in-machine-learning/