RESEARCH: SOLUTION TO MOST COMMON PROBLEMS IN MACHINE LEARNING

Ricardo Hernández Ramírez

15 DE SEPTIEMBRE DE 2023
UNIVERSIDAD POLITÉCNICA DE YUCATÁN
Robótica 9A

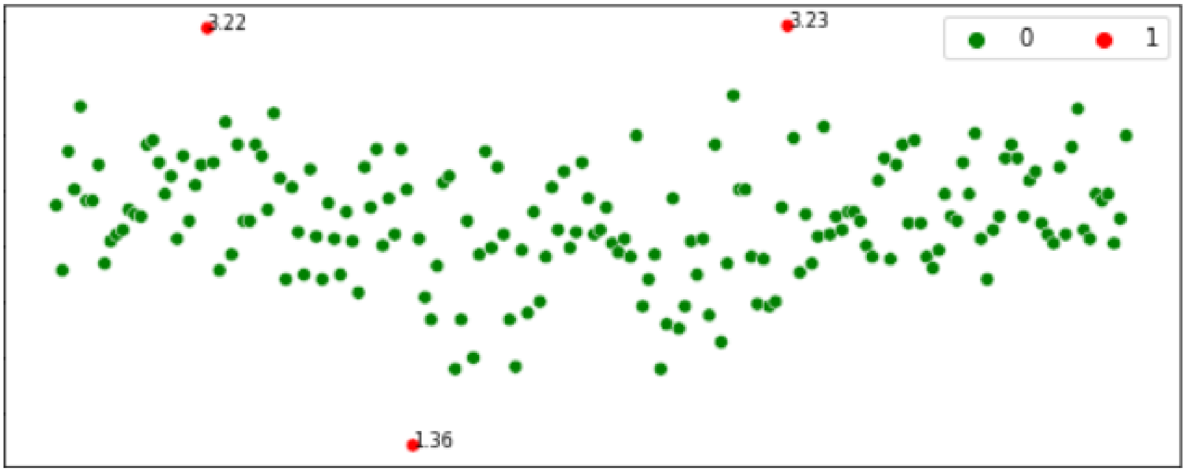
Overfitting and Underfitting:

A statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

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 result in poor performance both on the training and testing data. In simple terms, an underfit model's are inaccurate, especially when applied to new, unseen examples. It mainly happens when we uses very simple model with overly simplified assumptions. To address underfitting problem of the model, we need to use more complex models, with enhanced feature representation, and less regularization.

Outliers, definition and characteristics:

Outliers in data are values that look significantly different from other values in a dataset, for example:



Outliers can occur by errors during data entry, such as a faulty sensor, or naturally, such as the difference between employees of opposite hierarchy.

Outliers can cause issues during model fitting or might inflate error metrics that give higher weights to large errors.

Common solutions for overfitting, underfitting and outliers:

- Overfitting:
  - Increase training data.
  - Reduce model complexity.
  - Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
  - Ridge Regularization and Lasso Regularization.
  - Use dropout for neural networks to tackle overfitting.

- Underfitting
  - Increase model complexity.
  - Increase the number of features, performing feature engineering.
  - Remove noise from the data.
  - Increase the number of epochs or increase the duration of training to get better results.

- Outliers:
  - Increase model complexity.
  - Increase the number of features, performing feature engineering.
  - Remove noise from the data.
  - Increase the number of epochs or increase the duration of training to get better results.

What is the dimensionality problem?

The curse of dimensionality refers to a set of problems that arise when working with high-dimensional data. The dimension of a data set corresponds to the number of attributes/characteristics that exist in a data set. A data set with a large number of attributes, usually on the order of one hundred or more, is called high-dimensional data. Some of the difficulties that arise with high-dimensional data manifest themselves during analysis or visualization of the data to identify patterns, and some manifest themselves when training machine learning models. Difficulties related to training machine learning models due to high-dimensional data are called the “Curse of Dimensionality.” The popular aspects of the curse of dimensionality; “Data scarcity” and “distance concentration” are discussed in the following sections.

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. 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.

- 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.

The various methods used for dimensionality reduction include:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Generalized Discriminant Analysis (GDA)

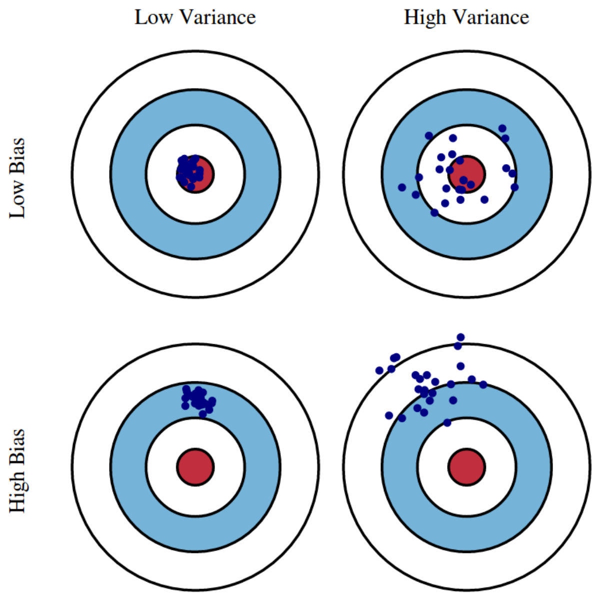
The Bias-Variance trade-off:

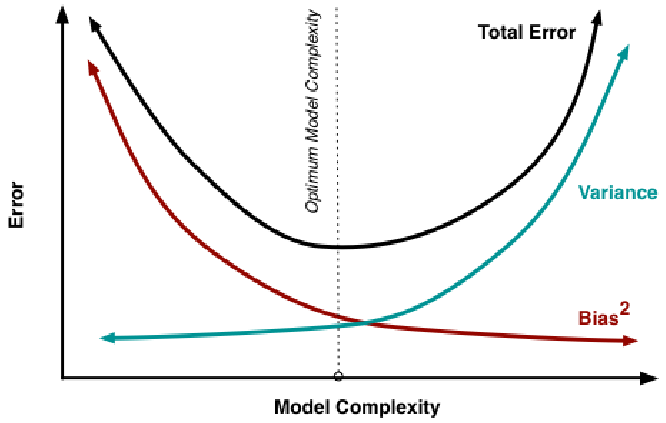
The bias–variance tradeoff describes the relationship between a model's complexity, the accuracy of its predictions, and how well it can make predictions on previously unseen data that were not used to train the model. In general, as we increase the number of tunable parameters in a model, it becomes more flexible, and can better fit a training data set. It is said to have lower error, or bias. However, for more flexible models, there will tend to be greater variance to the model fit each time we take a set of samples to create a new training data set. It is said that there is greater variance in the model's estimated parameters.

The bias–variance dilemma or bias–variance problem is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set:

- The bias error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

- The variance is an error from sensitivity to small fluctuations in the training set. High variance may result from an algorithm modeling the random noise in the training data (overfitting).





References:

[1] S. Singh, “Understanding the bias-variance tradeoff,” Medium, https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229 (accessed Sep. 15, 2023).

[2] “Introduction to dimensionality reduction,” Introduction to Dimensionality Reduction, https://www.geeksforgeeks.org/dimensionality-reduction/ (accessed Sep. 15, 2023).

[3] “Understanding curse of dimensionality,” Understanding Curse of Dimensionality, https://www.mygreatlearning.com/blog/understanding-curse-of-dimensionality/ (accessed Sep. 15, 2023).

[4] Muralidhar, “Outlier detection methods in machine learning,” Medium, https://towardsdatascience.com/outlier-detection-methods-in-machine-learning-1c8b7cca6cb8 (accessed Sep. 15, 2023).

[5] “ML: Underfitting and overfitting,” GeeksforGeeks, https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/ (accessed Sep. 15, 2023).

[6] A. Suresh, “How to remove outliers for machine learning?,” Medium, https://medium.com/analytics-vidhya/how-to-remove-outliers-for-machine-learning-24620c4657e8 (accessed Sep. 15, 2023).