Portfolio Evidence - Solution to most common problems in Machine Learning

Marco Alejandro González Barbudo

Universidad Politécnica de Yucatán

Computational Robotics Engineering

9th Quarter, Group B

Professor Víctor Alejandro Ortiz Santiago

September 15th, 2023

Portfolio Evidence - Solution to most common problems in Machine Learning

Overfitting & Underfitting:

Overfitting and underfitting are two common issues in machine learning and statistical modeling that relate to the performance of a model when applied to new, unseen data. These issues arise when a model doesn't generalize well from the training data to the testing or validation data.

Overfitting occurs when a machine learning model is excessively complex or flexible, and it learns not only the underlying patterns in the training data but also captures noise, random fluctuations, or outliers in the data.

Characteristics:

- The model performs exceptionally well on the training data.
- However, when tested on new, unseen data (validation or test data), the performance significantly degrades.
- Overfit models tend to have high variance, meaning they are sensitive to small changes in the training data, leading to poor generalization.

Causes:

- Using a very complex model with too many parameters.
- Having insufficient training data, making it easier for the model to memorize the training examples.
- Not applying regularization techniques or early stopping to prevent excessive learning.

Mitigation: To mitigate overfitting, is possible to:

- Use simpler models.
- Increase the amount of training data.
- Apply regularization techniques (e.g., L1, L2 regularization).
- Use cross-validation to tune hyperparameters.
- Employ early stopping during training.

Underfitting occurs when a machine learning model is too simple or inflexible to capture the underlying patterns in the training data.

Characteristics:

The model performs poorly on both the training data and new, unseen data.

It has high bias, meaning it oversimplifies the relationships in the data and fails to capture important patterns.

Causes:

Using an overly simple model that cannot represent the underlying data distribution.

Insufficient model complexity, such as using linear models for highly nonlinear data.

Mitigation: To mitigate underfitting, it is possible to:

- Use more complex models, such as deep neural networks or ensemble methods.
- Engineer better features that capture the data's complexity.
- Ensure you have enough training data.
- Tune hyperparameters appropriately.

The goal in machine learning is to find a model that strikes the right balance between fitting the training data well (reducing bias) and generalizing to new, unseen data (reducing variance). This balance is crucial for building models that perform well in real-world applications.

Characteristics of outliers:

In the realm of data analysis, outliers are data points that significantly deviate from the normal pattern of the dataset. They are the exceptional values that stand out amidst the majority of data points, often due to their unusual characteristics. Detecting outliers is a crucial step in data analysis as they can distort statistical measures, affect machine learning models, and lead to erroneous conclusions.

Types of Outliers:

Outliers come in various forms, each with its distinct characteristics:

1. **General or Normal Outliers:** These outliers do not have any specific patterns or associations with other data features. They are merely data points that deviate significantly from the overall dataset, like unusually high or low values.
2. **Private or Individual Outliers:** In this type, data points become outliers under specific conditions or contexts. For instance, a high temperature reading during winter or a low stock price in a bull market can be considered private outliers.
3. **Collective Outliers:** Collective outliers involve a group of data points that deviate together. These can signify issues or events affecting a group of related data points, such as delayed deliveries during a transportation strike.
4. **High-Dimensional Data Outliers:** These outliers are specific to datasets with numerous dimensions or features. They manifest as deviations across multiple dimensions, making them challenging to detect.

Detecting Outliers:

Detecting outliers is a critical task in data analysis, and several methods and techniques can be employed for this purpose. Here are some common approaches:

1. **Visualization Techniques:** Graphical tools like histograms, box plots, and scatter plots can help visualize the data distribution. Outliers often appear as data points far from the central clusters in these visualizations.

2. **Statistical Methods:** Statistical methods, such as the Z-score and the Interquartile Range (IQR), are effective in identifying outliers. The Z-score measures how many standard deviations a data point is away from the mean, while the IQR defines outliers as values lying outside a specific range around the median.

3. **Distance-Based Algorithms:** These algorithms analyze the proximity of data points to their neighbors. If a data point has significantly fewer neighbors than others, it may be considered an outlier. Examples of distance-based methods include DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and OPTICS (Ordering Points To Identify the Clustering Structure).

4. **Clustering Techniques:** Clustering algorithms, such as K-means, can be used to group similar data points together. Outliers are often those data points that do not fit well into any cluster or belong to small, sparse clusters.

5. **Machine Learning Approaches:** Machine learning models, particularly anomaly detection algorithms like Isolation Forests or One-Class SVM (Support Vector Machine), can automatically identify outliers in datasets. These models learn the normal patterns in the data and flag deviations as outliers.

Dimensionality problem:

The dimensionality problem is a common challenge encountered in the field of machine learning and data analysis. It arises when datasets contain a large number of features or variables, often referred to as dimensions. In high-dimensional spaces, the complexity of data increases significantly, leading to several issues and challenges.

One of the primary problems associated with high dimensionality is the increased risk of overfitting. Overfitting occurs when a machine learning model becomes too complex and captures noise or random variations in the data rather than the underlying patterns. This can result in poor generalization, where the model performs well on the training data but poorly on unseen or new data.

Another issue is the computational burden. With a large number of dimensions, the computational resources required for training and evaluating machine learning models can become impractical. Training times increase, and the models may become slow and inefficient, hindering their real-world applicability.

Moreover, high dimensionality can lead to the curse of dimensionality. This phenomenon refers to the increased sparsity of data points in high-dimensional spaces, making it challenging to find meaningful patterns or similarities between data points. It can affect various tasks such as clustering, classification, and regression.

Dimensionality reduction process:

Dimensionality reduction is a critical technique used to address the challenges posed by high-dimensional data. It involves the transformation of a dataset from a high-dimensional space to a lower-dimensional space while preserving essential information and patterns. This process aims to mitigate issues such as overfitting, computational complexity, and the curse of dimensionality.

Several dimensionality reduction techniques are available, and they can be broadly categorized into two types: feature selection and feature extraction.

1. **Feature Selection:** In feature selection, a subset of the most relevant features is chosen from the original dataset. Irrelevant or redundant features are discarded, simplifying the data representation. Common techniques for feature selection include methods based on statistical tests, correlation analysis, and domain knowledge.
2. **Feature Extraction:** Feature extraction methods create new features that are linear or nonlinear combinations of the original features. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are popular linear techniques for feature extraction. Nonlinear methods such as t-Distributed Stochastic Neighbor Embedding (t-SNE) can capture complex relationships in the data.

The choice of dimensionality reduction technique depends on the specific problem, dataset characteristics, and the goals of the analysis. After reducing the dimensionality of the data, machine learning models can be trained more efficiently, reducing the risk of overfitting and improving generalization performance. Additionally, dimensionality reduction can enhance data visualization and interpretation, making it a valuable tool in exploratory data analysis and model development.

Bias-variance trade-off:

The bias-variance trade-off is a critical concept in the field of machine learning and statistics, and it plays a fundamental role in the development of accurate predictive models. At its core, this trade-off involves finding the right balance between two types of errors that can occur when we build models based on data: bias and variance.

Bias, in this context, represents the error that arises from simplifying a real-world problem. It occurs when a model is too basic or makes overly simplistic assumptions, leading to predictions that consistently deviate from the actual values we are trying to predict. In simpler terms, a biased model is one that consistently gets things wrong in a systematic way.

On the other side of the trade-off is variance, which refers to the error introduced when a model is too sensitive to the specific training data it's exposed to. High variance models tend to overfit the training data, capturing not just the underlying patterns but also the noise or random fluctuations in that data. As a result, they struggle to generalize to new, unseen data because they've essentially memorized the training set rather than truly understanding the underlying patterns.

The bias-variance trade-off suggests that there is an optimal point of complexity for a model. If we aim to reduce bias, we typically increase the complexity of the model. However, this increase in complexity also leads to higher variance. Conversely, reducing model complexity to minimize variance may introduce more bias. The challenge is to strike the right balance where both bias and variance are minimized, resulting in a model that generalizes well to new, unseen data.

To navigate this trade-off effectively, machine learning practitioners use techniques like cross-validation. Cross-validation involves splitting the data into multiple subsets and repeatedly training and testing the model on different combinations of these subsets. It helps us assess how well a model will perform on data it hasn't seen before and aids in selecting the appropriate model complexity, regularization techniques, or hyperparameters.

Bibliography:

- Allamy, Haider. (2014). METHODS TO AVOID OVER-FITTING AND UNDER-FITTING IN SUPERVISED MACHINE LEARNING (COMPARATIVE STUDY).
- Falahi, Thaer & Nasserddine, Ghalia & Younis, Joumana. (2023). Detecting Data Outliers with Machine Learning. Al-Salam Journal for Engineering and Technology. 2. 10.55145/ajest.2023.02.02.018.
- Abdul Salam, Mustafa & Taher, Ahmad & Elgendy, Mustafa & Mohamed, Khaled. (2021). The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem. International Journal of Advanced Computer Science and Applications. 12. 10.14569/IJACSA.2021.0120480.