Universidad Politécnica de Yucatán

Solution to most common problems in ML

Computational Robotics

9°B

Machine learning

Unit I

Professor: MSc. Victor Alejandro Ortiz Santiago

Student: Diego Armando May Pech

Control Number: 2009092

I- Overfitting and Underfitting

Overfitting. It is quite common that, when embarking on the process of learning machine learning, we face the challenge of overfitting. What happens in this scenario is that our machine is limited to memorizing the specific examples it was given during training, losing the ability to generalize and recognize new input data. Our data sets often include unusual samples or samples with unexpected variations in some of their dimensions, as well as examples that may not be completely representative. When our model is overtrained and falls into overfitting, the algorithm considers valid only data that exactly matches that of the training set, including its imperfections, and cannot distinguish between quality data that deviates slightly from the predefined ranges.

Underfitting. Underfitting originates when excessive generalization of the input data is made to the model, which has a negative impact on the accuracy of the predictions. Underfitting occurs when a model cannot effectively capture the structure of the data set or achieve adequate generalization to new data. Furthermore, the model cannot establish a meaningful relationship between the input variables and the target variables**.**

II- characteristics of outliers

- Unusual Values: Outliers are values that differ significantly from the general pattern of the data set. Outliers are values that are in the extreme tails of the data distribution.

- Outliers: Outliers are noticeably far away from most other data points in the set. This is often measured in terms of distance or standard deviation.

- Potential to bias analysis: Outliers can distort statistical analysis and results, which can lead to erroneous interpretations or the creation of inappropriate models.

- Visual Identification: Outliers are often visible on scatter plots as points that lie well above or below the main group of points.

- Informational Potential: In some cases, outliers may contain valuable information about unusual or exceptional events in the data. Therefore, it is important to carefully consider whether to remove them or not, depending on the analysis objectives.

III- solutions for overfitting, underfitting and presence of outliers

To address overfitting:

- Use more training data.
- Apply cross-validation.
- Select relevant features.
- Apply regularization (L1, L2).
- Use simpler models.

To address underfitting:

- Increase the complexity of the model.
- Add relevant features.
- Use more advanced models.
- Adjust the hyperparameters of the model.

To address the presence of outliers:

- Identify and verify if outliers are errors or genuine data.
- Eliminate outliers if they are errors.
- Use data transformation techniques, such as standardization or normalization.
- Use learning algorithms robust to outliers, such as support vector machines (SVM) with robust loss function or random forests.
- Apply outlier detection techniques to specifically identify and treat them.

IV- Dimensionality problem.

The dimensionality problem, often referred to as the curse of dimensionality, is a challenge that arises when working with high-dimensional data in various fields, including machine learning, data analysis, and optimization. It is characterized by the negative impact of having a large number of features or dimensions in a dataset, and it can lead to several issues:

- Increased Computational Complexity
- Sparsity of Data
- Overfitting
- Curse of Sample Size
- Increased Model Complexity

V- Dimensionality reduction process.

Dimensionality reduction is a process used in data analysis and machine learning to reduce the number of features or dimensions in a dataset while preserving as much relevant information as possible. It is done to mitigate the curse of dimensionality, improve computational efficiency, and enhance the performance of models. The dimensionality reduction process typically involves the following steps:

- Data Preparation
- Understand the Data
- Choose a Dimensionality Reduction Technique
- Apply the Chosen Technique
- Determine the Number of Components
- Transform the Data
- Evaluate and Validate
- Interpretation (Optional)
- Visualization (Optional)
- Deploy in Machine Learning Models

VI- Bias-variance trade-off.

The balance between bias and variance, known as the bias-variance "trade-off", is a fundamental concept in machine learning.

- Bias (Underfitting): Represents errors due to overly simple models that cannot capture the complexity of the data.
- Variance (Overfitting): Represents errors due to overly complex models that fit noise in the data and do not generalize well.

The goal is to find a balance to create models that fit well on the training data and generalize well to new data. This involves choosing an appropriate complexity, using cross-validation to estimate performance, and regularizing models if necessary. In short, we are looking for models that are complex enough to capture patterns but not so complex that they are overfit.

References

- *Diferencias: underfitting vs overfitting*. (s.f.). KeepCoding Bootcamps. https://keepcoding.io/blog/underfitting-vs-overfitting/

- *¿Qué es un outlier o atípico?* (s.f.). Aprende con Eli. https://aprendeconeli.com/que-es-un-outlier-atipico/#:~:text=Un%20outlier%20es%20una%20observación,de%20los%20parámetros%20del%20mismo.

- *Qué es overfitting y underfitting y cómo solucionarlo*. (s.f.). Aprende Machine Learning. https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/

- *What is the curse of dimensionality and why does deep learning overcome it?* (s.f.). Impact with Artificial intelligence | Xomnia. https://www.xomnia.com/post/what-is-the-curse-of-dimensionality-and-why-does-deep-learning-overcome-it/

- (s.f.). https://www.themachinelearners.com/tradeoff-bias-variance/