FUNDAMENTAL CONCEPTS OF MACHINE LEARNING
PART 2

YULIANA ALEJANDRA MOLINA CORTES

UNIT 1

MACHINE LEARNING

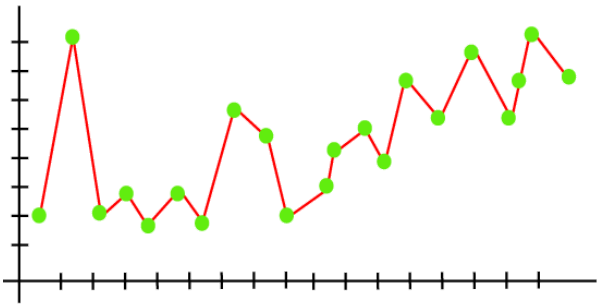
COMPUTATIONAL ROBOTICS ENGINEERING

GROUP 9B

Solution to most common problems in machine learning

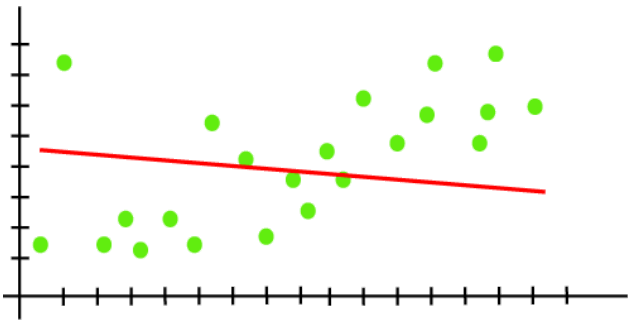
- **Define the concept of Overfitting:**

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. It means the more we train our model, the more chances of occurring the overfitted model. [1]



- **Define the concept of Overgeneralization(Underfiting):**

Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data. To avoid 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. Thus, the model is unable to capture the data points present in the plot. [1]



- **Distinguish the characteristics of outliers:**
  an outlier contains a value that is inconsistent or doesn't comply with the general behavior. A measurement error or an input error can lead to the existence of outlier values.
  Outliers introduce noise into a dataset and don't prove helpful. [2]

- **List the most common solutions for overfitting, overgeneralization and outliers**
  - Some ways by which we can reduce the occurrence of overfitting in our model, are Cross-Validation, Training with more data, Removing features, Early stopping the training, Regularization and Ensembling.
  - Ways to avoid underfitting are: By increasing the training time of the model and by increasing the number of features.
  - Solutions for outliers are: Z-Score, Local Outlier Factor (LOF), Isolation Forest, DBSCAN is a density-based clustering technique and Coresets.

- **Define the dimensionality problem**

The dimensionality problem in machine learning refers to the challenges and complications that arise when dealing with high-dimensional data, where the number of features or variables greatly exceeds the number of observations or samples. High dimensionality can lead to several issues:

1. Increased computational complexity: High-dimensional data requires more computational resources and time to train machine learning models, making it less efficient.

2. Curse of dimensionality: As the number of features grows, the volume of the feature space increases exponentially, causing data to become sparse. This sparsity can hinder the ability of models to generalize from the available data.

3. Overfitting: Models trained on high-dimensional data are more prone to overfitting, as they may capture noise or irrelevant patterns in the data, making them less effective on unseen data.

4. Interpretability and visualization: Understanding and visualizing high-dimensional data become challenging, making it harder to gain insights from the model or the data itself.

To address the dimensionality problem, techniques like feature selection, dimensionality reduction, and regularization are often employed to reduce the number of features or manage their impact on model performance.

- **Describe the process of Dimensionality Reduction:**

The number of input variables or features for a dataset is referred to as its dimensionality. Dimensionality reduction refers to techniques that reduce the number of input variables in a dataset. More input features often make a predictive modeling task more challenging to model, more generally referred to as the curse of dimensionality. [3]

High-dimensionality statistics and dimensionality reduction techniques are often used for data visualization. Nevertheless, these techniques can be used in applied machine learning to simplify a classification or regression dataset in order to better fit a predictive model.

Dimensionality reduction is the process of reducing the number of features (or dimensions) in a dataset while retaining as much information as possible. Reduce the complexity of a model, to improve the performance of a learning algorithm, or to make it easier to visualize the data. 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. This is a common problem in machine learning, where the performance of the model deteriorates as the number of features increases. [3]

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

Feature Selection: 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. Filter methods rank the features based on their relevance to the target variable, wrapper methods use the model performance as the criteria for selecting features, and embedded methods combine feature selection with the model training process.

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). PCA is a popular technique that projects the original features onto a lower-dimensional space while preserving as much of the variance as possible.

- **Explain the bias-variance trade-odd**

The bias-variance trade-off is a fundamental concept in machine learning that describes the balance between two types of errors a model can make: bias and variance. [4]

1. **Bias**: Bias refers to the error introduced by approximating a real-world problem, which may be complex, by a simplified model. A model with high bias makes strong assumptions about the underlying data distribution and is likely to underfit the data. In other words, it doesn't capture the underlying patterns in the data and has poor predictive performance.

2. **Variance**: Variance, on the other hand, refers to the model's sensitivity to the noise or fluctuations in the training data. A model with high variance is very flexible and fits the training data closely, but it may not generalize well to unseen data. Such a model is prone to overfitting, where it captures noise in the training data rather than the true underlying patterns. [4]

The trade-off arises because increasing the complexity of a model (e.g., adding more features or making it more flexible) typically reduces bias but increases variance, and vice versa. The goal in machine learning is to find the right balance between bias and variance to create a model that generalizes well to unseen data. This balance can lead to better model performance and predictive accuracy.

Techniques like cross-validation, regularization, and ensemble methods are commonly used to help strike an appropriate bias-variance balance when building machine learning models.

[1] "Overfitting and Underfitting in Machine Learning - Javatpoint". www.javatpoint.com. Accedido el 12 de septiembre de 2023. [En línea]. Disponible:
https://www.javatpoint.com/overfitting-and-underfitting-in-machine-learning

***[2] Top 5 Outlier Detection Methods Every Data Enthusiast Must Know*. (s.f.). DataHeroes.**
https://dataheroes.ai/blog/outlier-detection-methods-every-data-enthusiast-must-know/

***[3]Introduction to Dimensionality Reduction*. (s.f.). geeksforgeeks.**
https://www.geeksforgeeks.org/dimensionality-reduction/

[4] *Bias and Variance in Machine Learning - A Fantastic Guide for Beginners!* (s.f.). Analytics Vidhya.
https://www.analyticsvidhya.com/blog/2020/08/bias-and-variance-tradeoff-machine-learning/