Machine Learning

Research 2

Oscar Andre Dorantes Victor

IRC 9B

9/15/2023

- Define the concepts of: Overfitting & Underfitting.

Overfitting:

A model overfits when it learns too much from its training data, including its noise and errors. While it may excel with the training data, it performs poorly on new, unseen data because it's too tailored to the original dataset.

Underfitting:

An underfitted model is too simple to understand the data's complexities, leading to weak performance on both training and unseen data. It hasn't effectively captured the core patterns in the training data.

- Define and distinguish the characteristics of outliers.

Outliers:

Outliers are distinct data entries that diverge notably from the majority of observations. These anomalies can emerge from mistakes in gathering or processing the data. The process of examining these unique data points is called outlier scrutiny or extraction.

- Discuss the most common solutions for overfitting, underfitting and presence of outliers in datasets.

Ways to Counter Underfitting:

- Opt for more intricate models.
- Introduce more relevant features.
- Clean the data to reduce noise.
- Train for longer durations or use more epochs.

Ways to Counter Overfitting:

- Use more diverse training data.
- Opt for simpler models.
- Cease training once validation errors increase.
- Apply regularizations like Ridge or Lasso.
- For neural networks, use dropout to prevent reliance on specific neurons

Solutions for handling outliers:

- Visualization: Use scatter plots or box plots to visually identify outliers.
- IQR Method: Define outliers as values outside the range of (Q1 - 1.5IQR) and (Q3 + 1.5IQR).
- Z-Score: Label data points as outliers if their Z-score is beyond a set threshold, such as 3 standard deviations from the mean.
- Winsorizing: Replace outliers with the nearest non-outlier value.

- Robust Methods: Employ statistical methods less sensitive to outliers, like using medians instead of means.

- Describe the dimensionality problem.

In machine learning, managing high-dimensional data is feasible but can lead to the "curse of dimensionality" as dimensions increase. This elevates computational needs and complicates understanding. While more features can boost model accuracy, there's a limit before it drops. High-dimensional data analysis is challenging due to human pattern recognition limits and the growing computational demands. As dimensions rise, data becomes sparse, affecting sampling and clustering.

- Describe the dimensionality reduction process.

Dimensionality reduction simplifies datasets by reducing feature count without losing significant information. It converts high-dimensional data to a more manageable lower-dimensional form.

In machine learning, as feature count increases, model efficiency can suffer, leading to overfitting. Two main strategies combat this:

- Feature Selection: Chooses essential original features. Techniques include filter, wrapper, and embedded methods.
- Feature Extraction: Creates new features by combining originals. Popular methods are PCA, LDA, and t-SNE.

- Explain the bias-variance trade-off.

In machine learning, the bias-variance tradeoff is a important concept. It revolves around managing the accuracy of predictions on training data versus the model's applicability to new, unseen data. Essentially, it’s a balancing act between a model’s flexibility and its generalization capability.

References

GeeksforGeeks. (2023). ML Underfitting and overfitting. *GeeksforGeeks*. https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

GeeksforGeeks. (2020). Machine Learning Outlier. *GeeksforGeeks*. https://www.geeksforgeeks.org/machine-learning-outlier/

Sriram. (s. f.). Top 12 Commerce Project Topics & Ideas in 2023 [For Freshers]. *upGrad blog*. https://www.upgrad.com/blog/curse-of-dimensionality-in-machine-learning-how-to-solve-the-curse/

GeeksforGeeks. (2023a). Introduction to dimensionality reduction. *GeeksforGeeks*. https://www.geeksforgeeks.org/dimensionality-reduction/

Khaciyants, I. L. A. (2023). Bias-Variance tradeoff in machine learning. *Serokell Software Development Company*. https://serokell.io/blog/bias-variance-tradeoff