MOST COMMON MACHINE LEARNING PROBLEMS

15 DE SEPTIEMBRE DE 2023

AMAURY DAVID CASTELLANOS PALOMO

MACHINE LEARNING

Overfitting & Underfitting

Overfitting occurs when a machine learning model learns the training data too well and is unable to generalize to new data. This can happen when the model is too complex or when the training data is too small or noisy.

Underfitting occurs when the model is too simple and cannot learn the training data well enough. This can lead to poor performance on both the training and test data.

Outliers

Outliers are data points that are significantly different from the rest of the data. They can be caused by errors in data collection or measurement, or they may be genuine but rare data points.

Outliers can have a negative impact on machine learning models, as they can cause the model to overfit the training data.

Solutions for Overfitting, Underfitting, and Outliers

Here are some common solutions for overfitting, underfitting, and outliers in datasets:

Overfitting

- Reduce the complexity of the model. This can be done by using a simpler model architecture, removing unnecessary features, or using regularization techniques.
- Increase the size of the training data. This gives the model more data to learn from and makes it less likely to overfit the training data.
- Use cross-validation to evaluate the model on unseen data. This can help to identify overfitting early and make necessary adjustments to the model.

Underfitting

- Increase the complexity of the model. This can be done by using a more complex model architecture, adding more features, or using fewer regularization techniques.
- Engineer new features. This can help the model to better understand the data and make more accurate predictions.
- Use a larger training dataset. This gives the model more data to learn from and makes it less likely to underfit the training data.

Outliers

- Identify and remove outliers from the dataset. This can be done manually or using statistical methods.
- Use robust machine learning algorithms. These algorithms are less sensitive to outliers than traditional machine learning algorithms.
- Use regularization techniques. Regularization techniques can help to reduce the impact of outliers on the model.

Dimensionality Problem

The dimensionality problem refers to the challenges that arise when working with high-dimensional data. High-dimensional data is data that has a large number of features.

One of the main challenges of working with high-dimensional data is that it can be difficult to find meaningful patterns in the data. This is because the data is spread out over a large number of dimensions.

Another challenge is that high-dimensional data can lead to overfitting. This is because complex machine learning models are more likely to overfit high-dimensional data.

Dimensionality Reduction

Dimensionality reduction is the process of transforming high-dimensional data into a lower-dimensional space. This can be done using a variety of techniques, such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE).

Dimensionality reduction has a number of benefits. It can help to improve the performance of machine learning models by making it easier for the models to find meaningful patterns in the data. It can also help to reduce overfitting by making the data less complex.

Bias-Variance Tradeoff

Bias refers to the error introduced by approximating a real-world problem that may be complex by a simplified model. High bias can cause the model to underfit the data, this is when the model is too simplistic to represent the true relationship between the features and the target variable.

Variance, on the other hand, occurs because the model's sensitivity to small fluctuations or noise in the training data. High variance can cause the model to overfit the data, meaning it fits the training data very closely but fails to generalize well to unseen data. In this case, the model has essentially memorized the training data rather than learning the true underlying patterns.

It is impossible to completely eliminate both bias and variance from a machine learning model. The goal in machine learning is to find a model that strikes the right balance between bias and variance. It should be complex enough to capture the underlying patterns in the data (low bias) but not so complex that it fits the noise in the data (low variance). This tradeoff is essential for building models that generalize well to new, unseen data.

References

Machine Learning. (2023). How can you scale ML models to handle high traffic and large datasets? www.linkedin.com. https://www.linkedin.com/advice/1/how-can-you-scale-ml-models-handle-high-traffic

Kumar, N. (2023). Overfitting and underfitting in ML. *Spark By {Examples}*. https://sparkbyexamples.com/machine-learning/overfitting-and-underfitting-in-ml/

¿Qué es el sobreajuste? - Explicación del sobreajuste en Machine Learning - AWS. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/overfitting/

Awan, A. A. (2023, 13 septiembre). *The Curse of dimensionality in Machine Learning: challenges, impacts, and solutions*. https://www.datacamp.com/blog/curse-of-dimensionality-machine-learning

Lambert, N. (2021, 13 diciembre). Bias-Variance tradeoff — Fundamentals of Machine Learning. *Medium*. https://towardsdatascience.com/bias-variance-tradeoff-fundamentals-of-machine-learning-64467896ae67