Universidad Politécnica de Yucatán

Computational Robotics

Victor Ortiz

Fernando Alberto Loera Te

Quarter 9°A

15/09/2023

Machine Learning

Solution to most common problems in ML - Portfolio evidence

In the realm of data science and machine learning, the challenges of overfitting and underfitting loom large, where models can either grasp noise or fail to capture underlying patterns, often aggravated by the presence of outliers in datasets. Addressing these issues involves employing strategies like cross-validation, regularization, and outlier management. Meanwhile, the dimensionality problem poses computational and generalization challenges as datasets become increasingly complex, calling for dimensionality reduction techniques to distill essential information. Understanding the bias-variance trade-off is paramount, as it navigates the delicate balance between model complexity and generalization. In this exploration, we delve into these critical aspects, offering insights into combating overfitting, underfitting, handling outliers, dimensionality reduction, and mastering the bias-variance trade-off.

Overfitting

It is refered to a model which has been constructed and trained for too long on sample data o it is too complex, what makes it to learn irrelevant information, within the dataset, in other words, the model is not capable to generalize well to new data, which compromises the predictions and their classification work. Low error rates and high variance indicates the presence of overfitting.

How to prevent overfitting:

- Early stopping
- Train withmore data
- Data aumentating
- Ensemble methods
- Regularization.

Underfitting

It is learning model which can not model the training data effectively, what ends in a poor performance when training and testing data. This kind of model mainly happens when using simplified models, what turns them on inaccurate models when working with new and unseen examples.

How to prevent underfitting:

- Increase model complexity
- Increase numbre of features
- REmoving noise from the data

Outliers

An outlier is a data point that is noticeably different from the rest, They are extreme values within the dataset. The easiest way to detect them, is when you find errors in measurements, bad data collection or in variables that are not considered when collecting data, which means that their presence can skew the results of statical analyses on the datasets.

They can also find their way into a dataset naturally through variability, or they can be the result of issues like human error, faulty equipment, or poor sampling.

Most common solutions for overfitting, underfitting and presence of outliers in datasets

1. Overfitting

- Cross-validation: Implement techniques such as k-fold cross-validation to evaluate model performance on different subsets of data. This helps detect overfitting and guides hyperparameter tuning.
- Regularization: Apply techniques such as L1 and L2 regularization to penalize large model coefficients, promoting simpler models and reducing overfitting.
- Simpler models: Choose simpler models with fewer parameters when complex models are prone to overfitting.
- Feature selection: Select the most relevant features and remove irrelevant or redundant ones to reduce model complexity.
- Ensemble methods: Use ensemble methods such as Random Forests and Gradient Boosting, which combine multiple models to reduce overfitting.
- Early stopping: Monitor model performance on a validation set during training and stop training when performance starts to degrade, preventing overfitting.

2. Underufitting

- Feature engineering: Create additional relevant features or transform existing ones to help the model capture underlying patterns.
- Increase model complexity: Use more complex models with a greater number of parameters when simpler models fail to capture the complexity of the data.
- Hyperparameter tuning: Tune hyperparameters, such as learning rates or model depth, to tune model performance.
- Collect more data: Collecting more data can help the model learn better if the problem is due to a small data set.

3. Presence of outliers

- Outlier detection: Identify and flag outliers using statistical methods such as Z-scores, IQR, or visual inspection.
- Data transformation: Apply data transformations such as log scaling or robust scaling to reduce the impact of outliers on the model.

- Outlier removal: Consider removing extreme outliers from the data set if they are erroneous or negatively impact model training.
- Robust algorithms: Use machine learning algorithms that are less sensitive to outliers.

Dimensionality problems

As the number of dimensions or features in a data set increases, it has a profound impact on the machine learning process. A key effect is that the amount of data needed to accurately generalize a machine learning model increases exponentially. This phenomenon is often known as the curse of dimensionality. The reason behind this is that with higher dimensions, the data becomes sparser, meaning there is more space between data points in the feature space. As a result, the model must consider a greater number of possible combinations of features, making it increasingly difficult to find meaningful patterns and relationships within the data.

In high-dimensional spaces, the concept of distance between data points changes significantly. The points become more equidistant from each other, meaning they are all approximately similar distances away. This equidistant separation can cause difficulties in data sampling because it decreases the randomness in the selection of data points. In simpler terms, when there are many features, it is more difficult to collect diverse and representative samples from the data set, which can lead to biased or non-representative subsets. The challenge of high dimensionality also affects clustering techniques that rely on measuring similarity between observations.

Dimensionality reduction process

Dimensionality reduction is a critical process in data analysis and machine learning that involves reducing the number of features or dimensions in a data set while striving to preserve as much relevant information as possible. The main goal of dimensionality reduction is to simplify the data set by removing redundant or less important features. This process offers several advantages:

- Reducing model complexity
- Improving algorithm performance
- Enhacing visualization
- Simplifying data interpretation

Bias-variance trade-off

The balance between bias and variance is fundamental in machine learning and refers to the delicate balance between two key sources of error in predictive models: bias and variance.

The trade-off between bias and variance arises because an algorithm cannot simultaneously be more complex (to reduce bias) and less complex (to reduce variance). There is a balance that needs to be achieved. The goal is to find the

appropriate level of model complexity that minimizes the total error, which is the sum of bias and variance. This level of complexity leads to the best generalization performance for unseen data. Achieving this balance is essential to building models that are accurate and robust.

References

[1] Jason Brownlee. (2019). “Overfitting and Underfitting With Machine Learning Algorithms”. Machine Learning Mastery. Available:

https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

[2] Avijeet Biswal. (2023). “The Complete Guide on Overfitting and Underfitting in Machine Learning”. Simple Learn. Available:

https://www.simplilearn.com/tutorials/machine-learning-tutorial/overfitting-and-underfitting

[3] Eric Kleppen. (2023). “How To Find Outliers in Data Using Python (and How To Handle Them)”. CFBlog. Available:

https://careerfoundry.com/en/blog/data-analytics/how-to-find-outliers/

[4] Sriram. (2023). “Curse of dimensionality in Machine Learning: How to Solve The Curse?” upGrad. Available:

https://www.upgrad.com/blog/curse-of-dimensionality-in-machine-learning-how-to-solve-the-curse/

[5] Jason Brownlee. (2020). “Introduction to Dimensionality Reduction for Machine Learning” Machine Learning Mastery. Available:

https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/

[6] Seema Singh. (2018). “Understanding the Bias-Variance Tradeoff”.Towards Data Science. Available:

https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229