Universidad Politécnica de Yucatán

Raymundo Dariel Baas Cabañas

9°B Robótica Computacional

Solution to most common problems in ML

Machine Learning

Overfitting and Underfitting

A statistical model or a machine learning algorithm is said to have underfitting when a model is too simple to capture data complexities. It represents the inability of the model to learn the training data effectively result in poor performance both on the training and testing data.

A statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise.

Outliers

In simple terms, an outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data graph or dataset you're working with.

Outliers are extreme values that stand out greatly from the overall pattern of values in a dataset or graph.

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

Overfitting

1. Increase training data.
2. Reduce model complexity.
3. Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
4. Ridge Regularization and Lasso Regularization.
5. Use dropout for neural networks to tackle overfitting.

Underfitting

1. Increase model complexity.
2. Increase the number of features, performing feature engineering.
3. Remove noise from the data.
4. Increase the number of epochs or increase the duration of training to get better results.

Outliers

1. Visualization techniques like box and scatter plots.
2. Interquartile range (IQR) Data points outside the range of Q1 - 1.5 * IQR to Q3 + 1.5 * IQR are considered outliers.
3. Machine learning models to identify anomalies or outliers in the data like isolation forests.
4. Visualization and dimensionality reduction using techniques like t-SNE or PCA.

Dimensionality problem

The curse of dimensionality in machine learning is defined as follows,

As the number of dimensions or features increases, the amount of data needed to generalize the machine learning model accurately increases exponentially. The increase in dimensions makes the data sparse, and it increases the difficulty of generalizing the model. More training data is needed to generalize that model better.

The higher dimensions lead to equidistant separation between points. The higher the dimensions, the more difficult it will be to sample from because the sampling loses its randomness.

Dimensionality Reduction

Dimensionality reduction is a technique used to reduce the number of features in a dataset while retaining as much of the important information as possible.

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.

There are two components of dimensionality reduction:

- Feature selection: In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:
    1. Filter
    2. Wrapper
    3. Embedded
- Feature extraction: This reduces the data in a high dimensional space to a lower dimension space, a space with lesser number of dimensions.

The various methods used for dimensionality reduction include:

1. Principal Component Analysis (PCA)
2. Linear Discriminant Analysis (LDA)
3. Generalized Discriminant Analysis (GDA)

Bias-Variance Tradeoff

To understand how the Bias-Variance Tradeoff works we need to understand first what Bias and Variance is.

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data.

If our model is too simple and has very few parameters, then it may have high bias and low variance. On the other hand, if our model has large number of

parameters then it's going to have high variance and low bias. So, we need to find the right/good balance without overfitting and underfitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time.

An optimal balance of bias and variance would never overfit or underfit the model.

References

C, B. P. (2022, July 2). *FreeCodeCamp*. Retrieved from How to Detect Outliers in Machine Learning – 4 Methods for Outlier Detection: https://www.freecodecamp.org/news/how-to-detect-outliers-in-machine-learning/#:~:text=Outliers%20are%20those%20data%20points,data%20entry%2C%20or%20erroneous%20observations.

Lemonaki, D. (2021, August 24). *FreeCodeCamp*. Retrieved from What is an Outlier? Definition and How to Find Outliers in Statistics: https://www.freecodecamp.org/news/what-is-an-outlier-definition-and-how-to-find-outliers-in-statistics/#:~:text=In%20simple%20terms%2C%20an%20outlier,in%20a%20dataset%20or%20graph.

Overfitting, M. |. (2023, August 31). *GeekforGeeks*. Retrieved from ML | Underfitting and Overfitting: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Singh, S. (2018, May 20). *Towards Data Science*. Retrieved from Understanding the Bias-Variance Tradeoff: https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Sriram. (2023, February 24). *Upgrad*. Retrieved from Curse of dimensionality in Machine Learning: How to Solve The Curse?: https://www.upgrad.com/blog/curse-of-dimensionality-in-machine-learning-how-to-solve-the-curse/

Uberoi, A. (2023, May 6). *GeekforGeeks*. Retrieved from Introduction to Dimensionality Reduction: https://www.geeksforgeeks.org/dimensionality-reduction/