Universidad Politécnica de Yucatán

Computational Robotics Engineering

Machine Learning

Portfolio evidence

Solution to most common problems in ML

Teacher Victor Ortiz Santiago

Chi Centeno Mariana Guadalupe
2009038

September 15, 2023
9B

Solution to most common problems in ML

- **Define the concepts of: Overfitting & Underfitting.**
    - Overfitting

Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the model's ability to generalize.
Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning algorithms also include parameters or techniques to limit and constrain how much detail the model learns.
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.
The chances of occurrence of overfitting increase as much we provide training to our model. It means the more we train our model, the more chances of occurring the overfitted model.

Overfitting is the main problem that occurs in supervised learning.

- Reasons for Overfitting:
    - High variance and low bias.
    - The model is too complex.
    - The size of the training data.

    - 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. In simple terms, an underfit model's are inaccurate, especially when applied to new, unseen examples. It mainly happens when we uses very simple model with overly simplified assumptions. To address underfitting problem of the model, we need to use more complex models, with enhanced feature representation, and less regularization.
Underfitting refers to a model that can neither model the training data nor generalize to new data.

An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.
Underfitting is often not discussed as it is easy to detect given a good performance metric. The remedy is to move on and try alternate machine learning algorithms. Nevertheless, it does provide a good contrast to the problem of overfitting.
Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data. To avoid the 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.

- Reasons for Underfitting
  - The model is too simple, so it may be not capable to represent the complexities in the data.
  - The input features which is used to train the model is not the adequate representations of underlying factors influencing the target variable.
  - The size of the training dataset used is not enough.
  - Excessive regularizations are used to prevent the overfitting, which constraint the model to capture the data well.
  - Features are not scaled.

- **Define and distinguish the characteristics of 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.
Outliers are values at the extreme ends of a dataset.
Some outliers represent true values from natural variation in the population. Other outliers may result from incorrect data entry, equipment malfunctions, or other measurement errors.
An outlier isn’t always a form of dirty or incorrect data, so you have to be careful with them in data cleansing. What you should do with an outlier depends on its most likely cause.

- True outliers

True outliers should always be retained in your dataset because these just represent natural variations in your sample.
True outliers are also present in variables with skewed distributions where many data points are spread far from the mean in one direction. It’s important to select

appropriate statistical tests or measures when you have a skewed distribution or many outliers.

- Other outliers

Outliers that don't represent true values can come from many possible sources:

- Measurement errors
- Data entry or processing errors
- Unrepresentative sampling

This type of outlier is problematic because it's inaccurate and can distort your research results.
In practice, it can be difficult to tell different types of outliers apart. While you can use calculations and statistical methods to detect outliers, classifying them as true or false is usually a subjective process.

- **Ways of calculating outliers**

You can choose from several methods to detect outliers depending on your time and resources.

- Sorting method

You can sort quantitative variables from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.
This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.

- Using visualizations

You can use software to visualize your data with a box plot, or a box-and-whisker plot, so you can see the data distribution at a glance. This type of chart highlights minimum and maximum values (the range), the median, and the interquartile range for your data.
Many computer programs highlight an outlier on a chart with an asterisk, and these will lie outside the bounds of the graph.

- Statistical outlier detection

Statistical outlier detection involves applying statistical tests or procedures to identify extreme values.
You can convert extreme data points into z scores that tell you how many standard deviations away they are from the mean.
If a value has a high enough or low enough z score, it can be considered an outlier. As a rule of thumb, values with a z score greater than 3 or less than –3 are often determined to be outliers.

- Using the interquartile range

The interquartile range (IQR) tells you the range of the middle half of your dataset. You can use the IQR to create “fences” around your data and then define outliers as any values that fall outside those fences.
This method is helpful if you have a few values on the extreme ends of your dataset, but you aren’t sure whether any of them might count as outliers.

- Interquartile range method

1. Sort your data from low to high
2. Identify the first quartile (Q1), the median, and the third quartile (Q3).
3. Calculate your IQR = Q3 – Q1
4. Calculate your upper fence = Q3 + (1.5 * IQR)
5. Calculate your lower fence = Q1 – (1.5 * IQR)
6. Use your fences to highlight any outliers, all values that fall outside your fences.

Your outliers are any values greater than your upper fence or less than your lower fence.

- Find the outliers using tables.

The simplest way to find outliers in your data is to look directly at the data table or worksheet – the dataset, as data scientists call it.

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

Overfitting:

- Cross-Validation
- Training with more data
- Removing features
- Early stopping the training (have an eye over the loss over the training period as soon as loss begins to increase stop training).
- Regularization
- Describe the dimensionality problem.
- Reduce model complexity.
- Ridge Regularization
- Use dropout for neural networks to tackle overfitting.

Underfitting:

- Increasing the training time of the model.
- Increasing the number of features.
- Increase model complexity, performing feature engineering.
- Remove noise from the data.

- Increase the number of epochs or increase the duration of training to get better results.

- **Describe the 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.

As the dimensionality increases, the number of data points required for good performance of any machine learning algorithm increases exponentially.
Curse of Dimensionality refers to a set of problems that arise when working with high-dimensional data. The dimension of a dataset corresponds to the number of attributes/features that exist in a dataset. A dataset with a large number of attributes, generally of the order of a hundred or more, is referred to as high dimensional data. Some of the difficulties that come with high dimensional data manifest during analyzing or visualizing the data to identify patterns, and some manifest while training machine learning models. The difficulties related to training machine learning models due to high dimensional data are referred to as the 'Curse of Dimensionality'.

- **Describe the dimensionality reduction process.**

Dimensionality reduction is the process of reducing the number of features in a dataset while retaining as much information as possible.
This can be done to reduce the complexity of a model, improve the performance of a learning algorithm, or make it easier to visualize the data.

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.
In machine learning, high-dimensional data refers to data with a large number of features or variables. The curse of dimensionality is a common problem in machine learning, where the performance of the model deteriorates as the number of features increases. This is because the complexity of the model increases with the number of features, and it becomes more difficult to find a good solution. In addition, high-

dimensional data can also lead to overfitting, where the model fits the training data too closely and does not generalize well to new data.
Dimensionality reduction can help to mitigate these problems by reducing the complexity of the model and improving its generalization performance. There are two main approaches to dimensionality reduction: feature selection and feature extraction.

- **Explain the bias-variance trade-off.**

If the algorithm is too simple (hypothesis with linear equation) then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex (hypothesis with high degree equation) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as a Trade-off or Bias Variance Trade-off. 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

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.

Bias-variance decomposition and bias-variance tradeoff are closely related concepts.
Bias-variance decomposition is a mathematical technique that divides the generalization error in a predictive model into two components: bias and variance. In machine learning, as you try to minimize one component of the error (e.g., bias), the other component (e.g., variance) tends to increase, and vice versa. Finding the right balance of bias and variance is key to creating an effective and accurate model. This is called the bias-variance tradeoff.

References

- Bias-Variance Trade Off - Machine Learning - GeeksforGeeks. (n.d.). GeeksforGeeks. https://www.geeksforgeeks.org/ml-bias-variance-trade-off/
- Brownlee, J. (2019, August 12). Overfitting and Underfitting With Machine Learning Algorithms - MachineLearningMastery.com. MachineLearningMastery.com. https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/
- Curse of dimensionality in Machine Learning: How to Solve The Curse? | upGrad blog. (n.d.). upGrad blog. https://www.upgrad.com/blog/curse-of-dimensionality-in-machine-learning-how-to-solve-the-curse/
- Great Learning Team. (2022, December 13). Understanding Curse of Dimensionality. Great Learning Blog: Free Resources what Matters to shape your Career! https://www.mygreatlearning.com/blog/understanding-curse-of-dimensionality/
- How to Find Outliers | 4 Ways with Examples & Explanation. (n.d.-a). Scribbr. https://www.scribbr.com/statistics/outliers/
- Introduction to Dimensionality Reduction - GeeksforGeeks. (n.d.-b). GeeksforGeeks. https://www.geeksforgeeks.org/dimensionality-reduction/
- Karanam, S. (2021, August 10). Curse of Dimensionality — A “Curse” to Machine Learning. Medium. https://towardsdatascience.com/curse-of-dimensionality-a-curse-to-machine-learning-c122ee33bfeb
- Khaciyants, I. L. A. (2023, March 13). What Is the Bias-Variance Tradeoff in Machine Learning? Serokell Software Development Company. https://serokell.io/blog/bias-variance-tradeoff
- Lemonaki, D. (2021, August 24). What is an Outlier? Definition and How to Find Outliers in Statistics. freeCodeCamp.org. https://www.freecodecamp.org/news/what-is-an-outlier-definition-and-how-to-find-outliers-in-statistics/
- ML | Underfitting and Overfitting - GeeksforGeeks. (n.d.). GeeksforGeeks. https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/
- Overfitting and Underfitting in Machine Learning - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/overfitting-and-underfitting-in-machine-learning
- Singh, S. (2018, May 20). Understanding the Bias-Variance Tradeoff. Medium. https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229
- What are outliers and how to treat them in Data Analytics? - Aquarela. (n.d.). Aquarela. https://www.aquare.la/en/what-are-outliers-and-how-to-treat-them-in-data-analytics/