Universidad Politécnica de Yucatán

Robotics Engineering

Ortiz Victor

Machine Learning

Solution to most common problems in ML

by

Christian Dzul Canul

School of Computing Science and Engineering

UPY

Mérida, Yucatán.

September 2023

Instructions

Solutions to most common problems in Machine Learning:

- Defines the concepts of overfitting.
- Define the concept of overgeneralization (Underfitting).
- Distinguish the characteristics of outliers.
- List the most common solutions for overfitting, over-generalization and outliers.
- Describe the process of Dimensionality Reduction.
- Dimensionality Reduction.
- Define the dimensionality problem.
- Explain the bias-variance trade-off.

Solutions to the most common problems in machine learning

Overfitting & Overgeneralization

Overfitting is a problem that 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 not representative of the real world.

Overgeneralization (Underfitting) is a problem that occurs when a machine learning model does not learn the training data well enough and is unable to make accurate predictions on new data. This can happen when the model is too simple or when the training data is not large enough.

Outliers

An outlier is a data point that lies significantly outside the overall distribution of data. Outliers can be caused by a variety of factors, such as measurement errors, data entry errors, or genuine anomalies in the data.

Characteristics of outliers

- Outliers are relatively rare in a dataset.
- Outliers are much further away from the center of the distribution than most of the other data points.
- Outliers do not form tight clusters with other data points.

Examples of outliers

- A customer who spends 100 times more money than any other customer in a retail store.
- A test score that is significantly lower than all of the other test scores in a class.
- A measurement of temperature that is much higher or lower than all of the other temperature measurements in a given location.

Solutions for overfitting

- **Reduce the complexity of the model:** This can be done by using a simpler model architecture, reducing the number of parameters in the model, or using regularization techniques.
- **Increase the size of the training dataset:** This gives the model more data to learn from and helps to prevent it from overfitting to the specific training data.
- **Use cross-validation:** Cross-validation is a technique for evaluating the performance of a model on unseen data. This can be used to identify and address overfitting.

Solutions for underfitting

- **Increase the complexity of the model:** This can be done by using a more complex model architecture, increasing the number of parameters in the model, or using fewer regularization techniques.
- **Increase the size of the training dataset:** This gives the model more data to learn from and helps to improve its performance.
- **Use data augmentation:** Data augmentation is a technique for creating new training data from existing training data. This can be used to increase the size and diversity of the training dataset.

Solutions for the presence of outliers

- **Remove outliers:** Outliers can be removed from the training dataset before training the model. This can help to improve the accuracy of the model.
- **Use a robust model:** Robust models are less sensitive to outliers than traditional models. Some examples of robust models include decision trees, random forests, and support vector machines.
- **Use outlier detection techniques:** Outlier detection techniques can be used to identify and flag outliers in the training dataset. This information can then be used to remove outliers or to handle them in other ways.

Dimensionality Problem

The dimensionality problem is a set of challenges that arise when working with high-dimensional data. High-dimensional data is data that has a large number of features. For example, a dataset of images might have features such as pixel intensity, color, and texture.

One of the challenges of high-dimensional data is that it can be very sparse. This means that there are many data points that are very far away from each other. This can make it difficult to find patterns and relationships in the data.

Another challenge of high-dimensional data is that it can be difficult to train machine learning models on. This is because machine learning models need to be able to learn the relationships between the features in the data. However, when there are many features, it can be difficult for the model to learn these relationships.

There are a number of techniques that can be used to address the dimensionality problem. Some common techniques include:

- **Feature selection:** Feature selection is the process of selecting a subset of features that are most important for the machine learning model. This can help to reduce the dimensionality of the data and improve the performance of the model.
- **Dimensionality reduction:** Dimensionality reduction is the process of transforming the data into a lower-dimensional space while preserving the important information. This can also help to improve the performance of machine learning models.

Dimensionality reduction

The dimensionality reduction process involves transforming high-dimensional data into a lower-dimensional space while preserving the important information. This can be done using a variety of techniques, such as:

- **Principal component analysis (PCA):** It identifies the principal components of the data, which are the directions of greatest variance. The data is then projected onto the principal components, which reduces the dimensionality of the data while preserving as much of the variance as possible.
- **Singular value decomposition (SVD):** It is a similar technique to PCA, but it can also be used to compress data. SVD decomposes the data matrix into three matrices: the left singular vectors, the right singular vectors, and the singular values. The singular values are then used to reduce the dimensionality of the data.
- **Linear discriminant analysis (LDA):** It is a technique that is specifically designed for dimensionality reduction for classification problems. LDA identifies the directions that best separate the different classes of data. The data is then projected onto these directions, which reduces the dimensionality of the data while preserving as much of the information about the class labels as possible.

The dimensionality reduction process can be summarized in the following steps:

- **Choose a dimensionality reduction technique:** There are several different dimensionality reduction techniques available, so it is important to choose the one that is most appropriate for the specific problem and dataset.
- **Preprocess the data:** This may involve scaling the data, removing outliers, or transforming the data into a different format.
- **Apply the dimensionality reduction technique to the data:** This will produce a lower-dimensional representation of the data.
- **Evaluate the results:** It is important to evaluate the results of the dimensionality reduction technique to ensure that the important information has been preserved.

The bias-variance trade-off

The bias-variance trade-off is a trade-off between the bias and variance of a machine learning model. Bias is the error that occurs when the model is not a good fit for the data. Variance is the error that occurs when the model is sensitive to small changes in the data. A model with high bias is likely to underfit the data, while a model with high variance is likely to overfit the data. The goal is to find a model with a low bias and a low variance.

References

[1]. Dmytro Nikolaiev (Dimid. (2021, November 2). Overfitting and Underfitting Principles | by Dimid | Towards Data Science. Medium; Towards Data Science. https://towardsdatascience.com/overfitting-and-underfitting-principles-ea8964d9c45c

[2]. Dionysia Lemonaki. (2021, August 24). What is an Outlier? Definition and How to Find Outliers in Statistics. FreeCodeCamp.org; freeCodeCamp.org. https://www.freecodecamp.org/news/what-is-an-outlier-definition-and-how-to-find-outliers-in-statistics/

[3]. ML Underfitting and Overfitting. (2017, November 23). GeeksforGeeks; GeeksforGeeks. https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

[4]. Overfitting vs Underfitting in Machine Learning [Differences]. (2023). V7labs.com. https://www.v7labs.com/blog/overfitting-vs-underfitting

[5]. Introduction to Dimensionality Reduction Technique - Javatpoint. (2021). Www.javatpoint.com. https://www.javatpoint.com/dimensionality-reduction-technique#:~:text=The%20number%20of%20input%20features,predictive%20modeling%20task%20more%20complicated.

[6]. Singh, S. (2018, May 21). Understanding the Bias-Variance Tradeoff - Towards Data Science. Medium; Towards Data Science. https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229