Machine Learning

Solution to most common problems in ML

Author:

Christian Adriel Rodriguez

15/09/2023

Unit 1 Classwork 1

Solution to most common problems in Machine Learning.

Overfitting and Underfitting

- Overfitting
    1. 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 noise and inaccurate data entries in our data set.
       The model does not categorize the data correctly, because of too many details and noise. [1]

    2. Reasons for Overfitting
        - Data used for training is not cleaned and contains noise.
        - High variance (*Variance means the measure of the inconsistency of different predictions over varied datasets. Higher variance is an indication of overfitting in which the model loses the ability to generalize*)
        - The size of the dataset is not enough. [2]

    3. Techniques to avoid overfitting.
        - Train with more data: If we increase the training data, the important features are extracted and become prominent, so the model can recognize the relationship between the input attributes and the output variable.

        - Data argumentation: Makes a sample data look slightly different every time the model process it.

        - Feature selection: Every model has several parameters or features depending on the number of layers, number of neurons. he models can detect many redundant features or features determinable from other features leading to unnecessary complexity. [3]
- Underfitting
    1. The term “Underfitting” refers to a model that cannot capture the complexity of training data, therefore is does not adjust correctly to them. In other words, the model is too simple y is not able to identify the relationship between the input data and the output layers.

       When a model is underfitted is possible that has a high bias, that means that is too simple and cannot model appropriately.
    2. Prevent the underfitting.

        - Variety and balanced classes. In learning supervised we must classify a diverse of classes or categories. So, it is important to have a balanced class. For

example, we have apples, bananas, and lemons, we must have a lot of photos of each category.

- Parameter Tunning or Parameter Adjustment: we must experiment above all giving more/less "time/iterations" to the training and its learning until finding the balance.

- Excessive number of Dimensions (features), with many different variants, without enough samples. Sometimes it's wise to eliminate or reduce the number of features we'll use to train the model. A useful tool for doing so is PCA.

3. How detect a underfitting:
   If the model trained with the train set has a 90% success rate and with the test set has a very low percentage, this clearly indicates an overfitting problem.

   If in the Test set only one class type (for example "pears") is successful, or the only result obtained is always the same value will be that an underfitting problem occurred.

Characteristics of outliers

1. What are outliers in the data?
   An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.
   There are two activities that are essential for characterizing a set of data/
   - Examination of the overall shape of the graphed data for important features, including symmetry and departures from assumptions.
   - Examination of the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers. Two graphical techniques for identifying outliers, scatter plots and box plots, along with an analytic procedure for detecting outliers when the distribution is normal (Grubbs' Test), are also discussed in detail in the EDA chapter. [4]

Common solutions for overfitting and presence of outliers in dataset

1. **Hold-out (data)**

The recommended method is dividing a dataset into two separate sets—a training set and a testing set—typically utilizing an 80% training to 20% testing split ratio. The main goal is to train a model until both the training set and the testing set show that it performs proficiently. This demonstrates the model's strong generalizability to previously unobserved variables, which is essential for accurate prediction. It is underlined that, even after the split, having a sufficiently large dataset is crucial to ensuring the model can train efficiently and generalize in the right way.

2. **Data augmentation (data)**

A crucial method for preventing overfitting, a major issue in machine learning, is expanding a dataset's size. When collecting more data is not practical, the dataset can be artificially expanded by using data augmentation. Data augmentation is the process of adding different transformations

to an existing image dataset in the context of tasks like image classification, including flipping, rotating, rescaling, and shifting. By generating new data instances, these modifications effectively diversity the dataset and improve the model's ability to generalize. This method helps to enhance model performance, especially when the original dataset is small.

3. **L1 / L2 regularization (Learning algorithm)**
   Regularization is a technique to constrain our network from learning a model that is too complex. In L1 or L2 regularization, we can add a penalty term on the cost function to pusg the estimated coefficient towards zero. [5]

Dimensionality problem

Gathering a huge number of data may lead to the dimensionality problem where highly noisy dimensions with fewer pieces of information and without significant benefit can be obtained due to the large data. The exploding nature of spatial volume is at the forefront is the reason for the curse of dimensionality. [6]

Dimensionality reduction process

The process of dimensionality reduction is divided into two components. Feature selection and feature extraction. n feature selection, smaller subsets of features are chosen from a set of many dimensional data to represent the model by filtering, wrapping, or embedding. Feature extraction reduces the number of dimensions in a dataset to model variables and perform component analysis.

Methods of dimensionality reduction include:

- Factor analysis
- Low variance filter
- High Correlation Filer
- Principal Component Analysis
- UMAP

Bias variance trade-off

When a model makes simplifying assumptions to make the target function easier to approximate, this is referred to as bias. A model with a strong bias tends to oversimplify the underlying data patterns and may miss crucial relationships, which results in a poor fit on both the training and testing sets of data.

An underfit model is stated to be too inflexible or simplistic to adequately represent the complexity of the data and has a high bias. [7]

Bibliography

[1] Biswal. A (2023). *The Complete Guide on Overfitting and Underfitting in Machine Learning*. Simplilearn. Recovered by: https://www.simplilearn.com/tutorials/machine-learning-tutorial/overfitting-and-underfitting

[2] Baheti. P. (2021). *What is Overfitting in Deep Learning [+10 Ways to Avoid It]. V7labs. Recovered by: https://www.v7labs.com/blog/overfitting*

[3]

[4] Browniee. J. (2016). *Overfitting and Underfitting with Machine Learning Algorithms. Machinelearnigmastery.com Recovered by: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/*

[5] Chuan. D. (2020). *8 simple techniques to prevent overtfitting.* Towardsdatasetsicence.com Recovered by: https://towardsdatascience.com/8-simple-techniques-to-prevent-overfitting-4d443da2ef7d

[6] Choudhury. A. (2019). *Curse Of Dimensionality And What Beginners Should Do To Overcome It.* Analyticsindiamag.com Recovered by: https://analyticsindiamag.com/curse-of-dimensionality-and-what-beginners-should-do-to-overcome-it/#:~:text=Gathering%20a%20huge%20number%20of,for%20the%20curse%20of%20dimensionality.

[7] Bajwa. A (2020) *What is the tradeoff between Bias and Variance?* Educative.io Recovered by: https://www.educative.io/answers/what-is-the-tradeoff-between-bias-and-variance