MACHINE LEARNING PROBLEMS

Overfitting, Underfitting, Outliers, Dimensionality

Universidad Politécnica de Yucatán
Michelle Cámara González, 8°A Robotics

Machine Learning Fundamentals

Underfitting and Overfitting in Machine Learning

The first one is when 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 models are inaccurate, especially when applied to new, unseen examples. It mainly happens when we use 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.

Some reasons for this are the following:

1. The model is too simple. So, it may be not capable of represent the complexities in the data.
2. The input features which are used to train the model is not the adequate representation of underlying factors influencing the target variable.
3. The size of training dataset used is not enough.
4. Excessive regularizations are used to prevent the overfitting, which constraints the model to capture the data well.
5. Features are not scaled.

But we have techniques to reduce it.

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.

Meanwhile, the overfitting is when a statistical model s 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. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

In a nutshell, Overfitting is a problem where the evaluation of machine learning algorithms on training data is different from unseen data.

Some reasons for overfitting are the following.

1. High variance and low bias.
2. The model is too complex.
3. The size of the training data.

Like underfitting we have here some techniques to reduce the 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.

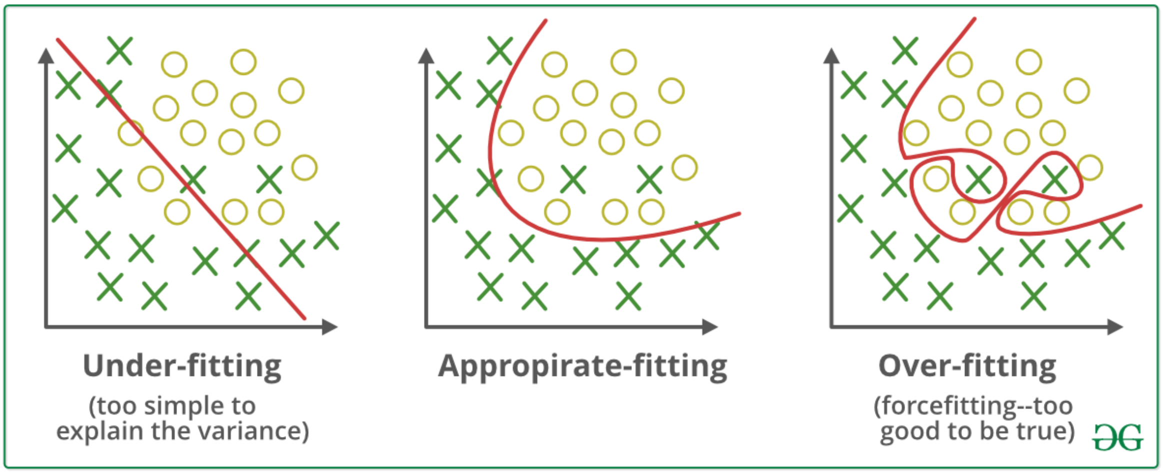
Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

Figure 1. Underfitting and Overfitting

Outliers in Data

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Outliers should be investigated carefully. Often, they contain valuable information about the process under investigation or the data gathering and recording process. Before considering the possible elimination of these points from the data, one should try to understand why they appeared and whether it is likely similar values will continue to appear. Of course, outliers are often bad data points.

Characteristics of outliers include:

- They are significantly different from most of the data points.
- They can be either higher or lower than most of the data.
- They can occur in one or multiple dimensions of the data.
- They can be identified using statistical methods or domain knowledge.

Also, we have solutions too for outliers, which are the following.

- Regularization: Regularization techniques, such as L1 and L2 regularization, can be used to add a penalty term to the model's loss function. This helps to reduce the complexity of the model and prevent overfitting.
- Cross-validation: Cross-validation techniques, such as k-fold cross-validation, can be used to evaluate the model's performance on multiple subsets of the data. This helps to assess the model's generalization ability and identify overfitting or underfitting.
- Feature selection: Feature selection techniques can be used to identify and remove irrelevant or redundant features from the dataset. This helps to reduce the complexity of the model and improve its performance.
- Outlier detection and handling: Outliers can be detected using statistical methods or domain knowledge. They can be handled by either removing them from the dataset or treating them separately during the modelling process.

Dimensionality problem and dimensionality Reduction

The dimensionality problem refers to the challenges that arise when working with high-dimensional data. As the number of dimensions increases, the complexity of the data increases, making it difficult to analyze and model effectively. This can lead to overfitting, increased computational requirements, and the need for more training data.

Dimensionality reduction techniques are used to address the dimensionality problem by reducing the number of features or dimensions in the data while preserving the important information. These techniques include:

- Principal Component Analysis (PCA): PCA transforms the original features into a new set of uncorrelated features called principal components. It aims to capture the maximum amount of variance in the data with a smaller number of components.
- Feature selection: Feature selection techniques, such as backward elimination or forward selection, can be used to select a subset of the most informative features.
- Feature extraction: Feature extraction techniques, such as linear discriminant analysis (LDA) or t-distributed Stochastic Neighbor Embedding (t-SNE), can be used to create new features that capture the underlying structure of the data.

Bias-Variance Trade-off

The bias-variance trade-off is a fundamental concept in machine learning that deals with the trade-off between a model's ability to fit the training data (low bias) and its ability to generalize to new, unseen data (low variance).

Bias refers to the error introduced by approximating a real-world problem with a simplified model. High bias models are too simple and may underfit the data.

Variance refers to the error introduced by the model's sensitivity to fluctuations in the training data. High variance models are too complex and may overfit the data.

The goal is to find the right balance between bias and variance to achieve good model performance. This can be achieved by selecting an appropriate model complexity, using regularization techniques, and applying cross-validation to assess the model's generalization ability.

References

- 7.1.6. What are outliers in the data? (n.d.). Nist.gov. Retrieved 16 September 2023, from https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm
- Follow, D. (2017, November 23). ML. GeeksforGeeks. https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/
- Sriram. (2023, February 25). Curse of dimensionality in machine learning: How to solve the curse? UpGrad Blog. https://www.upgrad.com/blog/curse-of-dimensionality-in-machine-learning-how-to-solve-the-curse/
- (N.d.-a). Statisticsbyjim.com. Retrieved 16 September 2023, from https://statisticsbyjim.com/basics/outliers/
- (N.d.-b). Javatpoint.com. Retrieved 16 September 2023, from https://www.javatpoint.com/dimensionality-reduction-technique