Universidad Politécnica de Yucatán

Machine Learning

Solution to most common problems in ML

Prof. Victor Ortiz

Oswaldo Josue Gomez Gonzalez

- **Overfitting & 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.

- **Characteristics of outliers**

Outliers are data points that deviate significantly from the rest of the data in a dataset. They can have a substantial impact on statistical analysis and machine learning models, so it's important to understand their characteristics. Here are some characteristics of outliers:

- Extreme Values: Outliers are typically values that are much larger or much smaller than most of the data points in the dataset. They are extreme values in relation to the rest of the data.
- Unusual or Rare: Outliers often represent rare events or extreme conditions that are not typical of the dataset. For example, in a dataset of people's ages, an outlier might be over 100 years old.
- Distinctive Appearance: Outliers can be easily spotted in graphical representations of data, such as box plots or scatter plots. They often appear as data points that are far away from the main cluster of data points.
- Influence on Summary Statistics: Outliers can significantly affect summary statistics such as the mean and standard deviation. The mean is particularly sensitive to outliers, and its value can be heavily skewed by the presence of outliers.
- Impact on Models: In machine learning, outliers can have a substantial impact on the performance of models. Some models are sensitive to outliers and may produce inaccurate predictions or classifications when outliers are present.
- Potential Errors: Outliers can sometimes be the result of errors in data collection or data entry. It's important to investigate the cause of outliers and determine whether they are genuine or erroneous.
- Context Matters: Whether a data point is considered an outlier can depend on the context of the analysis. In some cases, what might be an outlier in one context may not be considered an outlier in another.
- Identification Methods: There are various methods for identifying outliers, including graphical methods like scatter plots and box plots, statistical

methods like the Z-score and the IQR (interquartile range) method, and machine learning algorithms designed to detect anomalies.

- Treatment Options: Depending on the nature of the data and the analysis being performed, outliers can be handled in different ways. Options include removing them, transforming the data, or treating them as separate cases in a model.
- Interpretation: Outliers may hold valuable information about the data or the underlying process being studied. It's important to carefully interpret the reasons behind the existence of outliers and consider their potential impact on the analysis.

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

Techniques to Reduce 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.

Techniques to Reduce 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.

- **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.

It becomes harder to collect observations if there are plenty of features. These dimensions make all observations in the dataset to be equidistant from all other observations. The clustering uses Euclidean distance to measure the similarity between the observations. The meaningful clusters can't be formed if the distances are equidistant.

- **Dimensionality reduction process**

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 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.

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.

- **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.

- References:

ML: Underfitting and overfitting (2023) *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/ (Accessed: 13 September 2023).

Curse of dimensionality in machine learning: How to solve the curse? (no date) upGrad blog. Available at: https://www.upgrad.com/blog/curse-of-dimensionality-in-machine-learning-how-to-solve-the-curse/ (Accessed: 13 September 2023).

Introduction to dimensionality reduction (2023) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/dimensionality-reduction/(Accessed: 13 September 2023).