UNIVERSIDAD POLITÉCNICA DE YUCATÁN

BIS Universities

Universidad Politécnica de Yucatán

Ángel Iván Mayo Carrillo

2009093

9no Cuatrimestre

Machine learning

Solution to most common problems in ML

Overfitting, Underfitting and Outliers

Overfitting

Overfitting is a concept in data science, which occurs when a statistical model fits exactly against its training data. When a model is overfitted it will not be able to perform accurately with unseen data and will only work on the data it was trained on.

To train a model we need to divide the dataset in two parts, train sample and test sample. However, when models train for too long on sample data or the model is too complex, it will learn irrelevant information or noise and therefore the model will not be able to generalize well new data which leads to improper classification or prediction that was intended to.

Low error rates and a high variance are good indicators of overfitting.

If the training data has a low error rate and the test data has a high error rate, it signals overfitting.

Underfitting

Is another scenario in data science where a data model is unable to catch the relationship between the input and output variables accurately, giving feedback with high error rates on train data and test data. One of the reasons is when a model is too simple and requires a longer period of time for training, more input features, or less regularization.

When models are underfitting they will not be able to perform properly, it cannot establish the dominant trend within the data, resulting in training errors and poor performance of the model. If a model cannot generalize well to new data, then it cannot be leveraged for classification or prediction tasks.

High bias and low variance are good indicators of underfitting. Since this behavior can be seen while using the training dataset, underfitted models are usually easier to identify than overfitted ones.

Outliers

Outliers are data points that are very different from the rest of the sample. These are errors in measurement, poor or bad data collection, or even show variables not considered when collecting the data.

Many statistical tests are sensitive to outliers and therefore, the ability to detect them is an important part of data analytics. Outliers sometimes can be helpful indicators. For example, in some applications of data analytics like credit card fraud detection, outlier analysis becomes important because here, the exception rather than the rule may be of interest to the analyst.

Solutions to overfitting, underfitting and outliers

For overfitting there are multiple solutions:

- **Early stopping:** This method seeks to pause training before the model starts learning the noise within the model. This approach risks halting the training process too soon, leading to the opposite problem of underfitting. Finding the sweet spot is the task.
- **Train with more data:** Expanding the training set can increase the accuracy of the model by providing more opportunities to parse out the dominant relationship among the input and output variables. Is more effective when clean, relevant data is injected into the model.
- **Data augmentation:** Sometimes noisy data is added to make a model more stable. However, this method should be done sparingly.
- **Feature selection:** Is the process of identifying the most important ones within the training data and then eliminating the irrelevant or redundant ones. This is commonly mistaken for dimensionality reduction.
- **Regularization:** If we don't know which features to remove from our model, regularization methods can be used. Regularization applies a "penalty" to the input parameters with the larger coefficients, which subsequently limits the amount of variance in the model.
- **Ensemble methods:** Ensemble learning methods are made up of a set of classifiers. For example, decision trees—and their predictions are aggregated to identify the most popular result.

For underfitting

- **Decrease regularization:** Used to reduce the variance with a model by applying a penalty to the input parameters with the larger coefficients. However, if the data features become too uniform, the model is unable to identify the dominant trend, leading to underfitting. By decreasing the amount of regularization, more complexity and variation is introduced into the model, allowing for successful training of the model.
- **Increase duration of training:** Extending the duration of training can avoid underfitting.
- **Feature selection:** Specific features are used to determine a given outcome. If there are not enough predictive features present, then more features or features with greater importance, should be introduced.

For outliers

- **Deleting observations:** Delete outlier values if it is due to data entry error, data processing error or outlier observations are very small in numbers. But deleting the observation is not a good idea when we have small dataset.
- **Transforming values:** Transformed values reduces the variation caused by extreme values; Scaling, log transformation, cube root normalization, box-transformation.
- **Imputation:** it's also possible to impute outliers. Use mean, median, zero value in this methods. Since imputing there is no loss of data.
- **Separately treating:** If there are significant number of outliers and dataset is small, treat them separately in the statistical model.

Dimensionality problem

When more dimensions are added to a machine learning model, the processing power required for the data analysis increases. Moreover, adding more dimensions increases the amount of training data needed to make purposeful data models.

The problem 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.

Hughes phenomenon

The Hughes Phenomenon states that with the increase in the number of features, the classifier's performance also increases until the optimal number of features is attained. The classifier's performance degrades when more features are added according to the training sets' size.

Example of Hughes problem

Suppose a dataset consists of all the binary features. We also suppose that the dimensionality is 4, meaning there are 4 features. In this case, the number of data points is 2^4 =16.

If the dimensionality is 5, the number of data points will be 2^5 = 1024. These examples indicate that the number of data points exponentially increases with the dimensionality. So, the number of data points that a machine learning model needs for training is directly proportional to the dimensionality.

From the Hughes Phenomenon, it is concluded that for a fixed-sized dataset, the increment in dimensionality leads to reduced performance of a machine learning model.

Dimensionality reduction process

Dimensionality Reduction is the data conversion from a high-dimensional into a low-dimensional space. The idea behind this conversion is to let the low-dimensional representation hold some significant properties of the data. It is used to solve the Hughes phenomenon.

The reduction process solves the problem by decreasing the dataset and this decreases the storage needed, reduces the computation time of the model and the algorithms trains faster thanks to fewer dimensions, model accuracy gets higher, significant reduction in multicollinearity and simplifies the data visualization, it can also use tools to visualize the data in 1D, 2D or 3D.

Bias-variance trade-off

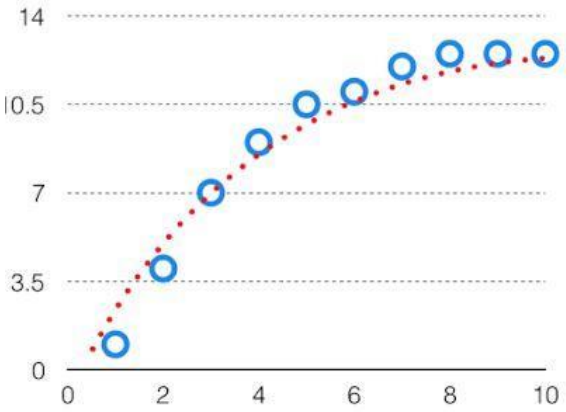
The bias is known as the difference between the prediction of the values by the Machine Learning model and the correct value. Being high in biasing gives a large error in training as well as testing data. Algorithms should always be low-biased to avoid the problem of underfitting. By high bias,

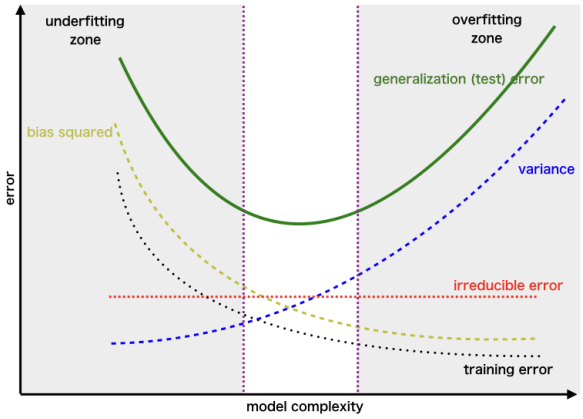
the data predicted is in a straight-line format, thus not fitting accurately in the data in the data set. It is another name for underfitting of data.

Variance is the variability of model prediction for a given data point which tells us the spread of our data is called the variance of the model. The model with high variance has a very complex fit to the training data and thus is not able to fit accurately on the data which it hasn't seen before. When a model is high on variance, it is then said to as Overfitting of Data.

- If the algorithm is too simple, then it may be on high bias and low variance condition and thus is error-prone.
- If algorithms fit too complex, then it may be on high variance and low bias. In the latter condition, the new entries will not perform well.

The combination of the two statements is what is known as Trade-off or Bias Variance Trade-off. The 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. For the graph, the perfect tradeoff will be like this.





References

Follow, P. (2020, February 3). *Bias-variance trade off - machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/ml-bias-variance-trade-off/

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/

Suresh, A. (2020, November 30). *How to remove outliers for machine learning? - analytics Vidhya - medium*. Analytics Vidhya. https://medium.com/analytics-vidhya/how-to-remove-outliers-for-machine-learning-24620c4657e8

Warudkar, H. (2020, September 11). *How to find outliers in data using machine learning*. Express Analytics. https://www.expressanalytics.com/blog/outliers-machine-learning/

What is overfitting? (n.d.). Ibm.com. Retrieved September 13, 2023, from https://www.ibm.com/topics/overfitting

What is underfitting? (n.d.). Ibm.com. Retrieved September 13, 2023, from https://www.ibm.com/topics/underfitting