UNIVERSIDAD POLITÉCNICA DE YUCATÁN

MACHINE LEARNING

Solution to most common problems in ML

KAREN VALERIA VILLANUEVA NOVELO
IRC9B
2009146

Solution to most common problems in ML

- Define the concepts of: Overfitting & Underfitting.

Overfitting in Machine Learning

Overfitting refers to a model that models the training data too well, It happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the model's ability to generalize.

Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning algorithms also include parameters or techniques to limit and constrain how much detail the model learns.

For example, decision trees are a nonparametric machine learning algorithm that is very flexible and is subject to overfitting training data. This problem can be addressed by pruning a tree after it has learned in order to remove some of the detail it has picked up.

Underfitting in Machine Learning

Underfitting refers to a model that can neither model the training data nor generalize to new data. An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.

It occurs when a model is too simple, which can be a result of a model needing more training time, more input features, or less regularization. Like overfitting, when a model is underfitted, 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.

- Define and distinguish the characteristics of outliers.

An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.

Many machine learning algorithms are sensitive to the range and distribution of attribute values in the input data.

Outliers in input data can skew and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results.

Even before predictive models are prepared on training data, outliers can result in misleading representations and in turn misleading interpretations of collected data. Outliers can skew the

summary distribution of attribute values in descriptive statistics like mean and standard deviation and in plots such as histograms and scatterplots, compressing the body of the data.
Finally, outliers can represent examples of data instances that are relevant to the problem such as anomalies in the case of fraud detection and computer security.

- Discuss the most common solutions for overfitting, underfitting and presence of outliers in datasets.

For overfitting:

- **Adding more data**

Your model is overfitting when it fails to generalize to new data. That means the data it was trained on is not representative of the data it is meeting in production. So, retraining your algorithm on a bigger, richer and more diverse data set should improve its performance. Unfortunately, getting more data can prove to be very difficult; either because collecting it is very expensive or because very few samples are regularly generated. In that case, it might be a good idea to use data augmentation.

- **Data augmentation**

This is a set of techniques used to artificially increase the size of a dataset by applying transformations to the existing data. For instance, in the case of images, you can flip images horizontally or vertically, crop them or rotate them. You can also turn them into grayscale or change the color saturation. As far as the algorithm is concerned, new data has been created. Of course, not all transformations are useful in every case. In short, data augmentation can be a very powerful tool but it requires a careful examination and understanding of your data.

- **Regularization**

Regularization actually refers to a large range of techniques. The main idea you need to remember is that these techniques introduce a "complexity penalty" to your model. If the model wants to avoid incurring that penalty, it needs to focus on the most prominent patterns which have a better chance of generalizing well. Regularization techniques are very powerful and almost all the models will use them in some way.

- **Removing features from data**

Sometimes, the model may fail to generalize simply because the data it was trained on was too complex and the model missed the patterns it should have detected. Removing some features and making data simpler can help reduce overfitting.
It is important to understand that overfitting is a complex problem.

For underfitting:

- **Increasing the model complexity**

The model may be underfitting simply because it is not complex enough to capture patterns in the data. Using a more complex model, for instance by switching from a linear to a non-linear model or by adding hidden layers to your neural network, will very often help solve underfitting.

- **Reducing regularization**

The algorithms used include by default regularization parameters meant to prevent overfitting. Sometimes, they prevent the algorithm from learning. Reducing their values generally helps.

- **Adding features to training data**

In contrast to overfitting, the model may be underfitting because the training data is too simple. It may lack the features that will make the model detect the relevant patterns to make accurate predictions. Adding features and complexity to the data can help overcome underfitting.

For presence of outliers in datasets:

- **Trimming/Remove the outliers**: In this technique, we remove the outliers from the dataset. Although it is not a good practice to follow.

- **Quantile based flooring and capping**: In this technique, the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value. The data points that are lesser than the 10th percentile are replaced with the 10th percentile value and the data points that are greater than the 90th percentile are replaced with 90th percentile value.

- **Mean/Median imputation**: As the mean value is highly influenced by the outliers, it is advised to replace the outliers with the median value.

- **Transformation**: Transforming the data using mathematical functions can sometimes reduce the impact of outliers. Common transformations include taking the logarithm, square root, or reciprocal of the data. These transformations can help make the data more normally distributed and stabilize the variance.

  Winsorization: Winsorization replaces extreme data values with less extreme values. The process involves capping or truncating the extreme values at a certain percentile (e.g., replacing values above the 95th percentile with the value at the 95th percentile). This approach reduces the influence of outliers while still retaining some information from the extreme values.

  Imputation: Instead of deleting outliers, they can be replaced with estimated values. Imputation techniques include replacing outliers with the mean, median, or another suitable value based on the characteristics of the data. Imputation should be done carefully, as it may introduce bias if not appropriately handled.

- **Robust methods**: Robust statistical methods are designed to be less sensitive to outliers. These methods estimate parameters using robust estimators that are not heavily

influenced by extreme values. For example, the median is a robust measure of central tendency that is less affected by outliers compared to the mean.

- **Model-based approaches**: In some cases, outliers can be detected and treated using specific models. For example, in regression analysis, influential outliers can be identified using diagnostic measures like Cook's distance or studentized residuals. Once identified, the outliers can be downweighted or excluded from the analysis.

- Describe the dimensionality problem.

It refers to when your data has too many features.

In today's big data world it can also refer to several other potential issues that arise when your data has a huge number of dimensions:

1. If we have more features than observations than we run the risk of massively overfitting our model — this would generally result in terrible out of sample performance.
2. When we have too many features, observations become harder to cluster.

Machine learning can effectively analyze data with several dimensions. However, it becomes complex to develop relevant models as the number of dimensions significantly increases. You will get abnormal results when you try to analyze data in high-dimensional spaces. This situation refers to the curse of dimensionality in machine learning. It depicts the need for more computational efforts to process and analyze a machine-learning model.

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. Clustering uses Euclidean distance to measure the similarity between the observations. The meaningful clusters can't be formed if the distances are equidistant.

- Describe the dimensionality reduction process.

Dimensionality reduction is the task of reducing the number of features in a dataset. In machine learning tasks like regression or classification, there are often too many variables to work with. These variables are also called features. The higher the number of features, the more difficult it is to model them, this is known as the curse of dimensionality. This will be discussed in detail in the next section.

Additionally, some of these features can be quite redundant, adding noise to the dataset and it makes no sense to have them in the training data. This is where feature space needs to be reduced.
The process of dimensionality reduction essentially transforms data from high-dimensional feature space to a low-dimensional feature space. Simultaneously, it is also important that meaningful properties present in the data are not lost during the transformation.
Dimensionality reduction is commonly used in data visualization to understand and interpret the data, and in machine learning or deep learning techniques to simplify the task at hand.

- Explain the bias-variance trade-off.

It's important to understand prediction errors (bias and variance). There is a tradeoff between a model's ability to minimize bias and variance. Gaining a proper understanding of these errors would help us not only to build accurate models but also to avoid the mistake of overfitting and underfitting.

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Models with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Models with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but have high error rates on test data.

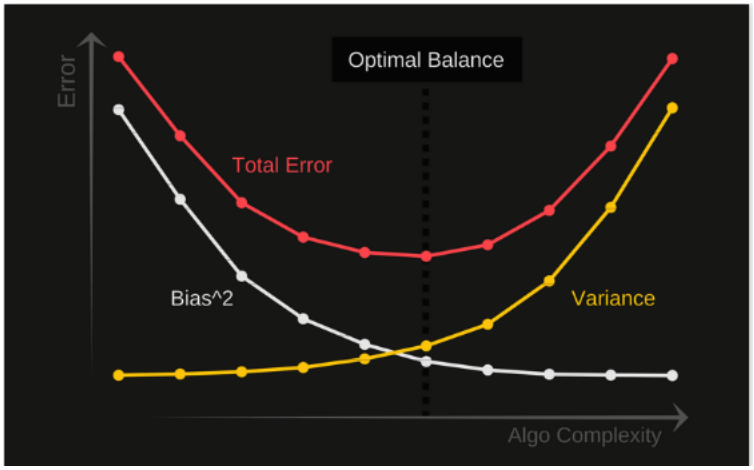
If our model is too simple and has very few parameters, then it may have high bias and low variance. On the other hand if our model has a large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.

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.

Total Error

To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.

Total Error = Bias^2 + Variance + Irreducible Error



An optimal balance of bias and variance would never overfit or underfit the model. Therefore understanding bias and variance is critical for understanding the behavior of prediction models.

References:

Barla, N. (2023). Dimensionality reduction for machine learning. *neptune.ai*. https://neptune.ai/blog/dimensionality-reduction

Bonthu, H. (2023). Detecting and Treating Outliers | Treating the odd one out! *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/

Brownlee, J. (2019). Overfitting and underfitting with machine learning algorithms. *MachineLearningMastery.com*. https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Brownlee, J. (2019). Overfitting and underfitting with machine learning algorithms. *MachineLearningMastery.com*. https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Brownlee, J. (2020). How to Identify Outliers in your Data. *MachineLearningMastery.com*. https://machinelearningmastery.com/how-to-identify-outliers-in-your-data/

Chemama, J. (2020, May 29). *How to solve underfitting and overfitting data models | AllCloud*. AllCloud. https://allcloud.io/blog/how-to-solve-underfitting-and-overfitting-data-models/

GeeksforGeeks. (2020). Machine Learning Outlier. *GeeksforGeeks*. https://www.geeksforgeeks.org/machine-learning-outlier/

Sriram. (n.d.). Top 12 Commerce Project Topics & Ideas in 2023 [For Freshers]. *upGrad blog*. https://www.upgrad.com/blog/curse-of-dimensionality-in-machine-learning-how-to-solve-the-curse/

What is Underfitting? | IBM. (n.d.). https://www.ibm.com/topics/underfitting

Yiu, T. (2021, December 11). The curse of dimensionality - towards data science. *Medium*. https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e