Universidad Politécnica de Yucatán

Computational Robotics Engineering

9° A

Machine Learning

Solution to most common problems in ML

Professor:

MSc. Victor Alejandro Ortiz Santiago

Name of the Author:

Ricardo Aron Tzuc Zih

Overfitting.

It is a common pitfall in deep learning algorithms in which a model tries to fit the training data entirely and ends up memorizing the data patterns and the noise and random fluctuations.

These models fail to generalize and perform well in the case of unseen data scenarios, defeating the model's purpose.

The high variance of the model performance is an indicator of an overfitting problem.

The training time of the model or its architectural complexity may cause the model to overfit. If the model trains for too long on the training data or is too complex, it learns the noise or irrelevant information within the dataset. (Baheti, 2023a)

Underfitting.

Underfitting is another common pitfall in machine learning, where the model cannot create a mapping between the input and the target variable. Under-observing the features leads to a higher error in the training and unseen data samples.

It is different from overfitting, where the model performs well in the training set but fails to generalize the learning to the testing set.

Underfitting becomes obvious when the model is too simple and cannot create a relationship between the input and the output. It is detected when the training error is very high, and the model is unable to learn from the training data. High bias and low variance are the most common indicators of underfitting. (Baheti, 2023b)

Characteristics of Outliers.

Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

An outlier cannot be termed as noise or error. Instead, they are suspected of not being generated by the same method as the rest of the data objects.

Outliers are of three types, namely:

1. *Global (or Point) Outliers*

Global outliers are data points that deviate significantly from the overall distribution of a dataset.

2. *Collective Outliers*

Collective outliers are groups of data points that collectively deviate significantly from the overall distribution of a dataset.

3. *Contextual (or Conditional) Outliers*

Contextual outliers are data points that deviate significantly from the expected behavior within a specific context or subgroup. (GeeksforGeeks, 2023b)

Solutions for Overfitting.

Here we will discuss possible options to prevent overfitting, which will help improve the model's performance.

1. Train with more data

With the increase in the training data, the crucial features to be extracted become prominent. The model can recognize the relationship between the input attributes and the output variable.

The only assumption in this method is that the data to be fed into the model should be clean; otherwise, it would worsen the problem of overfitting.

2. Data augmentation

An alternative method to training with more data is data augmentation, which is less expensive and safer than the previous method. Data augmentation makes a sample data look slightly different every time the model processes it.

3. Addition of noise to the input data

Another option (similar to data augmentation) is adding noise to the input and output data.

Adding noise to the input makes the model stable without affecting data quality and privacy, while adding noise to the output makes the data more diverse. Noise addition should be done carefully so that it does not make the data incorrect or irrelevant.

4. Feature selection

Every model has several parameters or features depending upon the number of layers, number of neurons, etc. The model can detect many redundant features leading to unnecessary complexity. We now know that the more complex the model, the higher the chances of the model to overfit.

5. Cross-validation

As mentioned above, cross-validation is a robust measure to prevent overfitting. The complete dataset is split into parts.

In standard K-fold cross-validation, we need to partition the data into k folds. Then, we iteratively train the algorithm on-1 folds while using the remaining holdout fold as the test set. This method allows us to tune the hyperparameters of the neural network or machine learning model and test it using completely unseen data.

6. Simplify data

Till now, we have come across model complexity to be one of the top reasons for overfitting. The data simplification method is used to reduce overfitting by decreasing the complexity of the model to make it simple enough that it does not overfit.

Some of the procedures include pruning a decision tree, reducing the number of parameters in a neural network, and using dropout on a neutral network. (Baheti, 2023b)

Solutions for Underfitting.

To avoid underfitting, we need to give the model the capability to enhance the mapping between the dependent variables.

1. Decrease regularization

Regularization discourages learning a more complex model to reduce the risk of overfitting by applying a penalty to some parameters. L1 regularization, Lasso regularization, and dropout are methods that help reduce the noise and outliers within a model.

More complexity is introduced into the model by decreasing the amount of regularization, allowing for successful model training.

2. Increase the duration of training

Early stopping the training can result in the underfitting of the model. There must be an optimal stop where the model would maintain a balance between overfitting and underfitting.

3. Feature selection

Introducing more features helps make the model more predictive. For example, we might introduce more hidden layers in deep neural networks, or in machine learning algorithms like the random forest, we may add more dependent variables.

This process will inject more complexity into the model, yielding better training results.

4. Remove noise from data

Removing noise from the training data is one of the other methods used to avoid underfitting. The presence of garbage values and outliers often cause underfitting, which can be removed by applying data cleaning and preprocessing techniques on the data samples. (Baheti, 2023b)

Presence of outliers in datasets.

In the machine learning pipeline, data cleaning and preprocessing is an important step as it helps you better understand the data. During this step,

you deal with missing values, detect outliers, and more.

As outliers have very different values abnormally low or abnormally high their presence can often skew the results of statistical analyses on the dataset. This could lead to less effective and less useful models.

But dealing with outliers often requires domain expertise, and none of the outlier detection techniques should be applied without understanding the data distribution and the use case. (C, 2022)

The dimensionality problem.

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.

The dimensionality reduction process.

Choose a dimensionality reduction technique. The choice of dimensionality reduction technique will depend on the specific data set and the desired outcome. For example, PCA is a good choice for reducing the dimensionality of a data set with a large number of correlated features, while LDA is a good choice for reducing the dimensionality of a data set with a categorical target variable.

Preprocess the data. This step is important to ensure that the data is in a format that is suitable for the dimensionality reduction technique. For example, PCA assumes that the data is normally distributed, so the data may need to be scaled before applying PCA.

Apply the dimensionality reduction technique to the data. This step will result in a new dataset with fewer dimensions. The number of dimensions in the new dataset is typically chosen by a trade-off between the amount of information that is preserved and the complexity of the model.

Evaluate the results. This step involves assessing how much information has been lost in the dimensionality reduction process. This can be done by comparing the original data set and the new data set using a variety of metrics, such as the mean squared error or the loss function. (Brownlee, 2020)

Variance-bias trade-off.

Variance-bias tradeoff is basically finding a sweet spot between bias and variance. We know that bias is a reflection of the model's rigidity towards the data, whereas variance is the reflection of the complexity of the data. High bias results in a rigid model. As we increase the capacity, the model tends to increase its flexibility by reducing the rigidity. Essentially, we're transforming an underfitted model towards a statistically good fit model by increasing the capacity.

A good practice is to check the training error and validation error. Because error = bias + variance. If both errors are less and close to each other, then the model has a good fit. (Barla, 2023)

References

Baheti, P. (2023a, abril 24). What is overfitting in deep learning [+10 Ways to Avoid it]. V7. https://www.v7labs.com/blog/overfitting

Baheti, P. (2023b, abril 24). Overfitting vs. underfitting: What's the difference? V7. https://www.v7labs.com/blog/overfitting-vs-underfitting

GeeksforGeeks. (2023b). Types of outliers in data mining. GeeksforGeeks. https://www.geeksforgeeks.org/types-of-outliers-in-data-mining/

C, B. P. (2022). How to Detect Outliers in Machine Learning – 4 methods for Outlier Detection. freeCodeCamp.org. https://www.freecodecamp.org/news/how-to-detect-outliers-in-machine-learning/#:~:text=Outliers%20are%20those%20data%20points,data%20entry%2C%20or%20erroneous%20observations.

Brownlee, J. (2020). Introduction to dimensionality reduction for Machine Learning. MachineLearningMastery.com. https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/

Barla, N. (2023). Overfitting vs underfitting in Machine Learning: Everything you need to know. neptune.ai. https://neptune.ai/blog/overfitting-vs-underfitting-in-machine-learning#:~:text=Underfitting%20is%20when%20the%20training,between%20the%20two%20is%20large.