“Dimensionality Reduction” Module Transcript

Chapter 1

Intro, Topics Covered, & Learning Outcomes

Hey. My name is Hayley, and I'm on the One AI team here at One Model. In previous modules, we've explored topics like model refinement and feature selection, emphasizing the importance of balancing the number of features the model selects to ensure robust and accurate predictions without compromising performance and interpretability. With this module, we will begin the advanced configuration module series and focus on dimensionality reduction as a key tool to achieve this balance. We'll dive into how it helps control feature selection methods and determine the optimal number of input features for the model's predictions.

We will cover an overview of dimensionality and dimensionality reduction, how dimensionality impacts machine learning models, available configuration options in One AI, the default configuration for dimensionality reduction in One AI, and we will wrap up with configuring dimensionality reduction in One AI.

After completing this module, you will understand what dimensionality is and why reducing dimensionality is necessary for improving model performance; know the default configuration for dimensionality reduction in One AI and how to configure and customize methods and feature numbers if desired, and you will be able to distinguish between various dimensionality reduction techniques and understand how to apply them effectively to enhance model performance.

Chapter 2

Dimensionality & Dimensionality Reduction Overview

Section 2 - Dimensionality & Dimensionality Reduction Overview

In order to explain dimensionality reduction, first let me tell you about dimensionality. Dimensionality refers to the number of input variables or features in a model dataset used for training a machine learning model. It indicates how many attributes each data point in the dataset has. For example, a dataset with three features, such as height, weight, and age, has a dimensionality of three.

A high-dimensional dataset has a very large number of features. Additionally, high-dimensional data often includes several rows of nearly all null data or constant

values, which are not predictive and add extra noise for the model to sort through when trying to make predictions and learn. We don't want this because it negatively impacts computational efficiency and model performance, among other things we will discuss in the next section.

Dimensionality reduction is a key tool in addressing high-dimensional datasets. This process involves reducing the number of input variables or features to simplify the dataset while retaining as much relevant information as possible. There are several different techniques to achieve this, but I'm only going to focus on the methods that are available for use in One AI.

Before we focus on methods, just to give you an idea of how this applies to One AI models, depending on how much data you have loaded in One Model, model datasets contain anywhere from 50 to 500 features typically for One AI to sort through and consider for selection. While somewhat subjective and dependent on the specific model, we have found that 5 to 15 features is the sweet spot for commonly run One AI models before model performance degrades and understanding and explaining the model becomes more challenging. Therefore, One AI must use dimensionality reduction to go from those 50 to 500 features down to selecting 5 to 15.

One AI does this by optimizing dimensionality through the use of filter and wrapper methods. First, the filter method is used. This evaluates each feature individually using statistical tests to see how relevant each feature is to the target variable, which is what we're predicting. We specifically use univariate tests, which don't consider how features interact with each other and the target value, just how important each one is on its own. The goal is to rank features based on their individual relevance or importance. After this step, we have a group of important features that move on to the next step.

The next step is the wrapper method. In this step, a predictive model, specifically a random forest, tests different combinations of these important features to find the best set. It does this by trying all of the features together and then removing one at a time to see if the model performs just as well without it. If it does, that feature is dropped. This method is more computationally intensive than the filter method, which is why we filter features first.

The main difference between the two methods is that the wrapper uses a learning algorithm itself to evaluate the usefulness of the features, while the filter evaluates features based on general characteristics of the data. Together, they do a great job of presenting the model only with the most powerful features.

Chapter 3

Dimensionality Impacts on Models

Section 3 - Dimensionality Impacts on Models

The One AI dimensionality reduction configuration settings aim to give users control over feature selection, reducing the number of features in a dataset so that the model can focus on the most relevant ones. While it might seem that more data for the model to learn from would always be better, there's actually an ideal number of input variables that improves model performance. Therefore, reducing dimensionality is important for several reasons.

First, it helps avoid the curse of dimensionality and overfitting. The curse of dimensionality refers to the phenomenon where as the number of features or dimensions in a dataset increases, the amount of data needed to effectively cover the feature space grows exponentially. In simpler terms, with more features, the dataset becomes more complex, requiring significantly more data points to adequately explore and represent all possible combinations of feature values and capture the dataset's variability.

This complexity can lead to sparsity where the dataset has many features, but most data points are empty or zero, making it hard to build accurate models because models don't learn well from null data. We want the model to adapt and improve as more data is added, but high-dimensional data often causes overfitting. Overfitting happens when models capture noise or random patterns in the training data instead of the true underlying patterns that help them make accurate predictions.

This makes the model too complex and fit too closely to the training data resulting in poor performance when it's given new data to make predictions on. Reducing dimensionality typically results in more compact and informative representations of the data. By removing irrelevant features and focusing on the important ones, the model's performance improves. It also makes it easier to visualize the data, helping end users understand and interpret complex patterns more effectively.

Finally, high-dimensional data requires more time and computational power to train and run models. Therefore, dimensionality reduction allows us to utilize larger and more complex datasets without excessive run times or computational demands.

Chapter 4

Configuration Options in One AI

Section 4 - Configuration Options in One AI

While the default configuration typically performs quite well, manual configuration can give you more control about how many features are tried and selected, which is important if there are specific features you want to see tested in your model to test your hypotheses.

First, you can disable dimensionality reduction. While this is not typically recommended, it can be helpful in cases of small feature sets, highly curated features, extensive feature engineering, or specific model requirements.

Next, you have the option to configure the filter method and or filter number of features. There are three options for method. First, we have mutual info, which quantifies how much knowing the value of one feature reduces uncertainty about the value of another feature. It helps determine which features are most relevant to the target variable. Higher mutual information between a feature and the target variable indicates a stronger relationship, suggesting that the feature is more informative for predicting the target. It's suitable for both categorical and numerical features.

Next is the chi-square test, which evaluates whether there is a significant relationship between a feature and the target variable. It does this by comparing the observed counts of data points in different categories of the feature against the expected counts if there were no relationships. It helps to identify features that have a strong association with the target variable. Higher chi-square values indicate a stronger relationship between the feature and the target, suggesting that the feature is more informative for the model. It is most suitable for categorical features.

And finally, we have the f-test (ANOVA), which is an analysis of variance that is used for comparing the means of three or more groups to see if at least one of them is significantly different from the others. It is commonly used in feature selection for regression tasks and can handle of The filter number of features can be thought of as the maximum number of features that will be selected and used in the model and is configured by inputting any whole number into the designated fields.

Moving into the wrapper step. You cannot configure the Wrapper method because recursive feature elimination is the only option available. This technique aims to select the most important features by recursively considering smaller and smaller sets of features. It works by fitting a model and removing the weakest feature or features until a specified number of features is reached.

The wrapper minimum number of features can be thought of as the minimum number of features that will be selected and used in the model and is configured by inputting any

whole number into the designated field. It must be less than or equal to the filter number of features because the minimum cannot exceed the maximum.

Chapter 5

Default Configuration in One AI

Section 5 - Default Configuration in One AI

As I mentioned in the previous section, the default configuration in One AI generally performs well and is based on best practices and what we have found to be optimal in most models.

The default filter method is mutual info. For the filter number features, by default 5, 10, and 15 are all tried and the best result is selected. The default wrapper method is recursive feature elimination. And the wrapper minimum features defaults to 5.

In the next section, we will hop over to One Model, and I will demo how to perform these configurations..

Chapter 6

Configuring Dimensionality Reduction in One AI

Section 6 - Configuring Dimensionality Reduction in One AI

To configure dimensionality reduction in One AI, click the ‘Edit’ button for the model of interest, and scroll down to “Dimensionality Reduction” under One AI configuration. Toggle the override slider to 'On' and click the caret next to "Dimensionality Reduction" to reveal your options.

To disable dimensionality reduction, toggle the 'No Selection' override slider to 'On'. Then toggle the slider beneath "No Selection" to ‘On’.

For configurations regarding the filter step of dimensionality reduction, click the caret next to "Filter Methods" to reveal your options. First, you can configure the filter method with the dropdown. Multiple methods can be selected and each selected method will be tried, and the one that results in the highest performance will be chosen and used in the final model. Filter number features can be configured by putting any whole number in the designated field. For example, we can make our feature maximum 12, like so.

Remember, this is the maximum number of features that will be selected and used in the model.

To make configurations in the wrapper step of dimensionality reduction, slide the override toggle to 'On' and collapse by clicking the caret next to wrapper methods. As I previously mentioned, the method is grayed out because you cannot change the method from recursive feature because it's the only option. Wrapper minimum features can be configured by putting any whole number in the designated field. We can make our minimum features 10, like so. If you configured the filter number features, this number must be less than or equal to that value because this is the minimum number of features that will be selected and used in the model, so it can't exceed the maximum without a model error.

With this configuration, One AI will try all of our filter methods and select that of which results in highest performance. The maximum number of features the model can select will be 12, with the minimum being 10.

Once you are happy with your configuration, scroll to the bottom and click 'Save', then rerun your model. Remember, you can always view the final selected dimensionality reduction configuration in the Estimation Details section of the Result Summary for any completed model run.

Chapter 7

Conclusion & Thanks

Mastering dimensionality reduction enhances both the performance and interpretability of your machine learning models. By understanding and applying the concepts and techniques covered in this module, you can effectively control feature selection, number of features, mitigate risks associated with high-dimensional data, and optimize your models for better results. Whether using One AI's default settings or manual configurations, you now have the knowledge to make informed decisions about fine tuning. Happy modeling!