Course: MMO: Introduction to Machine Learning

Video Number:01-04

Estimated Length: 5 min

|  |  |  |
| --- | --- | --- |
|  | **Script text or talking points** | **Visuals / Actions on Screen** |
| 1 | There are several concepts you should know when talking about machine learning.  The idea of data pre-processing is that the data we need to create the model is likely not in good enough condition to be used without working with it first. The data won’t be in the right shape, it will be inaccurate, some will be missing. The data will need to be processed and cleaned to be useful.  The term feature extraction is commonly used in machine learning to identify what parts of the data are useful in training the model. Feature extraction involves filtering and labeling the data in such a way to give our training process what it needs.  Data is commonly split into different sets. Training data is the data that is used in the machine learning process. Validation data is then used to see how accurate the training has been and then used to make adjustments during the training process. Finally, scoring data is used to test the model against data it has never seen before. You may see some slightly different terms used for these concepts.  A model is the result of the training process. New data can then be fed into the model and the model will return what it believes to be the correct result. In some ways, the model can be considered a very complex function.  Prediction is the process of sending new data into a trained model to get a result. | Slide 1 |
|  | While this may sound counter-intuitive, it may not be good for a model to be able to predict the training and validation data with 100% accuracy. What this may indicate is that the model is overfitted. An overfitted model is on that is very good at predicting the data it was trained with but poor at predicting data that it hasn’t seen.  Over fitting is one of the reasons we test afterwards with scoring data. If a model becomes increasingly accurate at predicting its training data, but inaccurate at testing it’s scoring data, that is a sign of overfitting.  The other side of overfitting is underfitting. Underfitting occurs when the model is still very inaccurate at any prediction, including its training data, validation data or it’s scoring data. A model that is underfit likely needs more training. | Slide 2 |
|  | It is important to recognize the implication of underfitting and overfitting, and that is ambiguity. Unlike many functions that you may be used to writing as a developer, there is not 100% certainty that it will correctly predict an outcome.  In many ways this is like how we all see the world. Presented with a picture of a tree, we may not always accurately be able to say what type of tree it is. Maybe the picture is blurry, maybe there is some other type of tree that looks very similar. At the end of the day we won’t always know with 100% certainty what type of tree is pictured.  This brings us to the question when training our models, how accurate is accurate enough. The answer is, well it depends. It depends on the nature of the problem; it depends on the business case and the expectations for accuracy. It is safe to say that a model that needs to identify potentially problematic entries on a tax return form can have a lower standard of accuracy than one that tries to drive a car and has life or death consequences. | Slide 3 |