Splitting the Data into Training and Evaluation Data

The fundamental goal of ML is to *generalize* beyond the data instances used to train models. We want to evaluate the model to estimate the quality of its pattern generalization for data the model has not been trained on. However, because future instances have unknown target values and we cannot check the accuracy of our predictions for future instances now, we need to use some of the data that we already know the answer for as a proxy for future data. Evaluating the model with the same data that was used for training is not useful, because it rewards models that can "remember" the training data, as opposed to generalizing from it.

A common strategy is to take all available labeled data, and split it into training and evaluation subsets, usually with a ratio of 70-80 percent for training and 20-30 percent for evaluation. The ML system uses the training data to train models to see patterns, and uses the evaluation data to evaluate the predictive quality of the trained model. The ML system evaluates predictive performance by comparing predictions on the evaluation data set with true values (known as ground truth) using a variety of metrics. Usually, you use the "best" model on the evaluation subset to make predictions on future instances for which you do not know the target answer.

Amazon ML splits data sent for training a model through the Amazon ML console into 70 percent for training and 30 percent for evaluation. By default, Amazon ML uses the first 70 percent of the input data in the order it appears in the source data for the training datasource and the remaining 30 percent of the data for the evaluation datasource. Amazon ML also allows you to select a random 70 percent of the source data for training instead of using the first 70 percent, and using the complement of this random subset for evaluation. You can use Amazon ML APIs to specify custom split ratios and to provide training and evaluation data that was split outside of Amazon ML. Amazon ML also provides strategies for splitting your data. For more information on splitting strategies, see Splitting Your Data.

Document Conventions « Previous Next »