Todo: continue to add

Found a good compilation of Machine learning definitions at <http://robotics.stanford.edu/~ronnyk/glossary.html>

Unseen data: The data that has not been used to train the model, so during the testing process the model’s predictions are genuine.

Related: Testing dataset.

Online model: The model is considered online when after the initial model is deployed it is able to continue to intake new data which causes the model to dynamically adapt.

Related: Offline model

Offline model: The model is considered offline when after the model is trained then deploy it does not change.

Related: Online model.

Supervised learning: In the dataset it contains a dependent attribute, also known as the ground truth, that is defined by the independent attributes in the instance. With this data available the algorithm is able to learn the relationship between the observations and result for the model. For example, in the Iris dataset the attribute ‘class’ is the result of the other attribute observations. Usually induction algorithms like … use supervised learning.

Related: semi-supervised learning, unsupervised learning.

Semi-supervised learning: In the case that some of the instances have a labeled value or the dataset is so massive that manually labeling each instance is too time consuming semi-supervised learning may be used. A possible strategy is to utilize unsupervised learning to cluster the data into groups, and then perform supervised learning onto each of the groups to accurately label them.

Related: Supervised learning, Unsupervised learning.

Unsupervised learning: The dataset does not contain a pre-specified dependent attribute that is the result of the independent attributes. It is best suited for datasets that are unlabeled with the use of clustering algorithms like …. Based on the clusters and patterns it may be able to classify the data’s results.

Related: Supervised learning, Semi-supervised learning.

Reinforced learning:

Regularization: When a model is very complex and overfitting occurs on the training dataset, a method to address this problem is to increase the amount of data to reduce the variance but this isn’t always feasible. So, another method is to simplify the model by techniques like drop out regularization, data augmentation, and early stopping.

Drop out regularization: In a neural network where there are layers of decisions, remove 50 % of the nodes in each layer to create a simpler model.

Drop out only done during the training process because when done during the testing process it would randomize the outputs and would add noise.

Data augmentation: Consider an image classification problem where it has a deficiency of data so it would be prone to overfitting, a solution would be to create new data by altering an existing image by rotating and cropping.

Early stopping: Stop the training process, for example stopping a decision tree growth.

Clustering:

Association:

Classification model:

Related: Regression model.

Regression model:

Related: Classification model.

Feature vector: Is an n-dimensional vector of numerical features that represent some object.

Attribute:

Also known as field, variable, and feature.

Encoding:

Instances: A row in the dataset.

Training dataset:

Related: Training process

Training process:

Related: Training dataset.

Validation dataset:

Related: Testing process

Testing process:

Related: Validation dataset

Good fit:

Related: Underfit, Overfit.

Under fit: An under fitted model is when the trained model does not fit well, in terms of having high deviation from what the ground truth. The main cause of this is that the algorithm used was too simple for the data provided, for example using a linear equation is more limited than a high powered polynomial when formulating a relationship. In the case of a classification model using a linear model could severely limit its performance. Therefore, if the accuracy of the model during training and when predicting on unseen data is low, then a solution is to choose more complex and flexible algorithms to generate a better fit.

Related: Good fit, Overfit.

Over fit: An over fitted model is one that fits the training data very well, which usually results in low accuracy on the testing/unseen data. In this case, the main cause would be that the algorithm used is so complex that during training it was able to fit every data point and when tested on unseen data it would be likely that it would incorrectly evaluate it. A solution to overfitting would be to use an algorithm that produces a simpler model from the dataset, so it does not take every outlier or data noise into the model. Another solution would be to add a regularization term to the algorithm that is overfitting, where it penalizes the model when it attempts to fit a data point that variates to much from the mean (?).

Related: Good fit, Underfit.

Gaussian Distribution:

Also known as Normal Distribution

Hyper-parameter tuning:

Ground truth: Used in supervised learning where each instance is provided an attribute that is the target result of the non-target attributes.

Also known as target attribute, Labeled data.

k-Fold cross validation: “In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data.”

Source: <https://www.openml.org/a/estimation-procedures/1>