ACTIVIDAD #4
Tipo actividad: Reading

Reading:

Machine Learning Principles Explained

Learning is the Result of Representation, Evaluation, and Optimization

The field of machine learning has exploded in recent years and researchers have developed an enormous number of algorithms to choose from. Despite this great variety of models to choose from, they can all be distilled into three components.

The three components that make a machine learning model are representation, evaluation, and optimization. These three are most directly related to supervised learning, but it can be related to unsupervised learning as well.

Representation – this describes how you want to look at your data. Sometimes you may want to think of your data in terms of individuals (like in k-nearest neighbors) or like in a graph (like in Bayesian networks).

Evaluation – for supervised learning purposes, you'll need to evaluate or put a score on how well your learner is doing so it can improve. This evaluation is done using an evaluation function (also known as an objective function or scoring function). Examples include accuracy and squared error.

Optimization – using the evaluation function from above, you need to find the learner with the best score from this evaluation function using a choice of optimization technique. Examples are a greedy search and gradient descent.

Generalization is Key

The power of machine learning comes from not having to hard code or explicitly define the parameters that describe your training data and unseen data. This is the essential goal of machine learning: to generalize a learner's findings.

To test a learner's generalizability, you'll want to have a separate test data set that is not used in any way in training the learner. This can be created by either splitting your entire training data set into a training and test set, or to just collect more data. If the learner were to use data found in the test data set, this would create a sort of bias in your learner to do better than in reality.

One method to get a sense on how your learner will do on a test data set is to perform what is called cross-validation. This randomly splits up your training data into a given number of subsets (for example, ten subsets) and leaves one subset out while the learner trains on the rest. And then once the learner has been trained, the left out data set is used for testing. This training, leaving one subset out, and testing is repeated as you rotate through the subsets.

Beware of Overfitting

If a learning algorithm fits a given training set well, this does not simply indicate a good hypothesis. Overfitting occurs when the hypothesis function J(Θ) fits your training set too closely having a high variance and low error on the training set while having a high test error on any other data.

In other words, overfitting occurs if the error of the hypothesis as measured on the data set that was used to train the parameters happens to be lower than the error on any other data set.

Choosing an Optimal Polynomial Degree

Choosing the right degree of polynomial for the hypothesis function is important in avoiding overfitting. This can be achieved by testing each degree of polynomial and observing the effect on the error result over various parts of the data set. Hence, we can break down our data set into 3 parts that can be used in optimizing the hypothesis' theta and polynomial degree.

A good break-down ratio of the data set is:

Training set: 60%
Cross validation: 20%
Test set: 20%

The three error values can thus be calculated by the following method:

1. Use the training set for each polynomial degree in order to optimize the parameters in Θ
2. Use the cross validation set to find the polynomial degree with the lowest error
3. Use the test set to estimate the generalization error

Ways to Fix Overfitting

These are some of the ways to address overfitting:

1. Getting more training examples
2. Trying a smaller set of features
3. Increasing the parameter

Adapted from: https://www.freecodecamp.org/news/machine-learning-principles-explained/

16. Matching Heading Definition:

Match the concepts of column A with their corresponding definitions of column B according to the text:

COLUMN A:

1. Representation
2. Generalization
3. Evaluation
4. Optimization
5. Overfitting
6. Cross validation
7. Test Data Set
8. Polynomial Degree
9. Machine Learning
10. Training Set

COLUMN B:

A. The ability of a machine learning model to apply its findings to new, unseen data, avoiding overfitting to the training data.

B. Occurs when a learning algorithm fits the training set too closely, leading to high variance and low error on the training set but poor generalization.

C. A separate dataset used to assess the generalizability of a learner, not used in training to prevent bias.

D. The field of artificial intelligence that involves the development of algorithms allowing machines to learn from data.

E. The degree of the polynomial used in the hypothesis function, crucial in avoiding overfitting.

F. Describes how you want to look at your data, for example, thinking of data in terms of individuals or in a graph.

G. In the context of supervised learning, it involves assessing or scoring the performance of a learner using an evaluation function.

H. The process of finding the best parameters for a learner using a technique, often based on an evaluation function.

I. A technique where the training data is split into subsets, and the learner is trained on one subset while tested on another, repeated to assess performance.

J. The portion of the data set used to train the machine learning model, typically a majority of the data