Machine Learning Overview (& Examples)

September 8, 2019

AIAA GHS NEXT GEN TECHNICAL SYMPOSIUM

Machine Learning Overview

We'll cover several areas of Machine Learning algorithms / techniques along with some sample applications.

- Regression / Prediction
- Classification
- Clustering
- Optimization

Regression

Regression is a way to predict an outcome based on a set of variables, after training using a set of labeled data.

It is often the first ML method that people learn - also may be the one fallback used.

Regression

Regression is used to answer questions like:

“Given current income, savings, # of late payments, # of kids, marital status, etc—will this person default on a loan of $$$ amount?”

“Given square footage, age, school district, zip code, price of nearby homes—what is the value of a house?”

Regression

There are two parts of the question that will determine which type of regression would be the most useful:

1. The expected answers—Is this a yes/no question? If you plotted the actual answers from historical data, do they look like a straight line or some type of curve? The type of regression method is generally dependent on the type of answer expected.

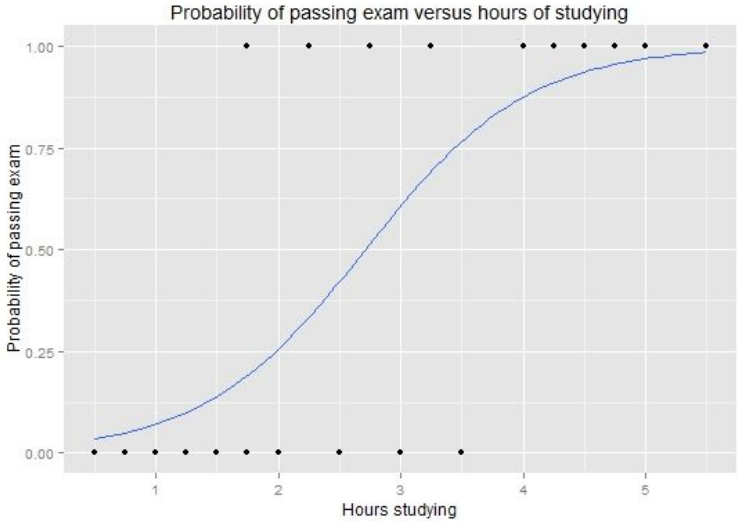
2. The input values—Are they independent of each other (square footage and age)? Are they highly related (zip code and school district)? Based on the dependence between input values, special cases of regression methods can take this into account.

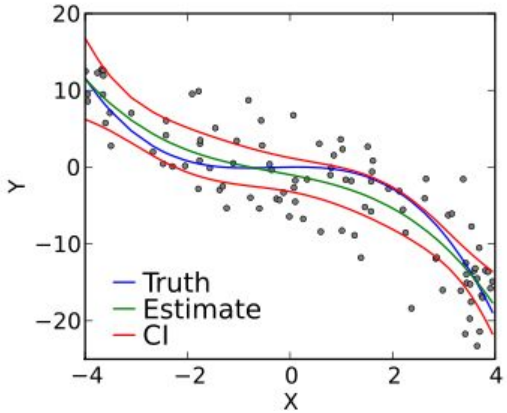
Regression

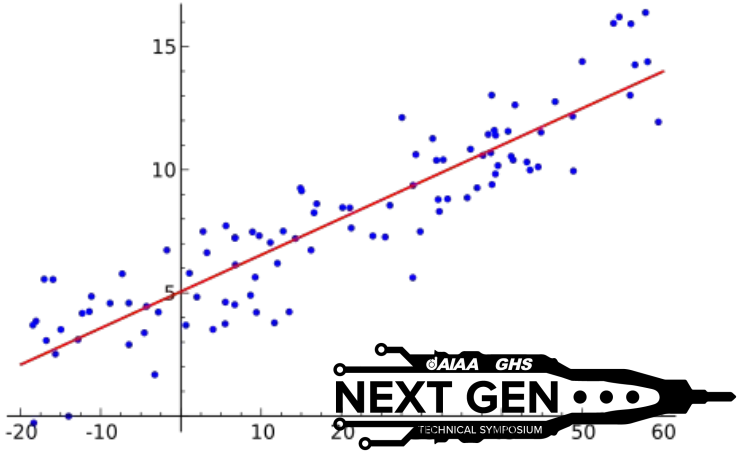
Logistic Regression is used for yes/no answers.

Linear Regression is used when answers follow a straight line if you put them on a graph.

Polynomial Regression is used when answers follow a curve if you put them on a graph.







Regression

Regression Tips:

Always check the data you're training on to see if it is a singular set of data and not actually separate groupings.

For example, a collection of housing data that includes both Madison City and rural Madison County will provide two distinct groupings for price given the same square footage.

Classification

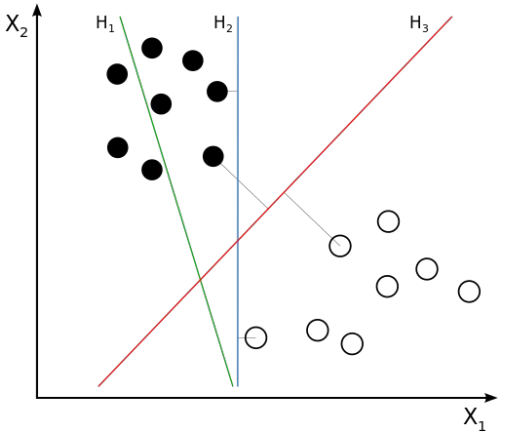
Classification is used to take a piece of data and determine which classification it should belong to, from a set of known classifications.

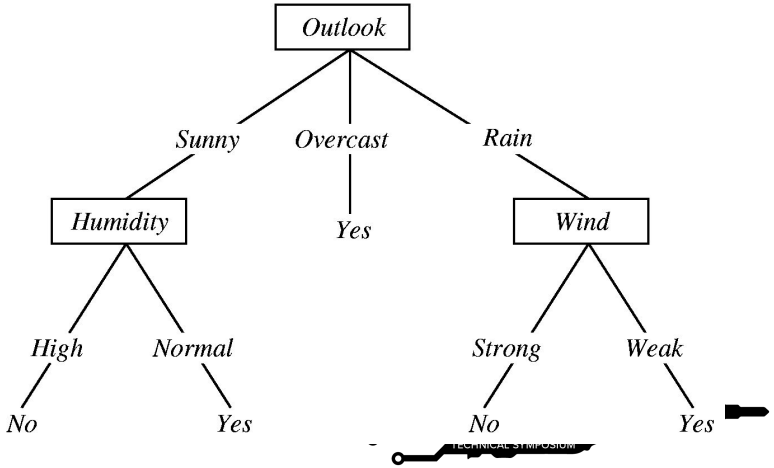
Typically, there are two types of classification problems:

- Binary - there are only two classes - like “Spam or Not Spam”. This is the same as a yes or no answer. This can usually be solved through Logistic Regression.
- Multiple Classes - potentially a large number of classes to choose from. This is used to answer questions like “What part of speech is the word ‘blue’ in the sentence ‘Roses are red, violets are blue’ ?”

Classification

- **Linear Regression** - if the classifications can be grouped separately along some arbitrary line, then linear regression may be the best and most efficient way to solve the problem.
- **Support Vector Machine** - attempts to find a line between two groups in training data.
- **Decision Tree** - used to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Thank you Wikipedia!
- **Naive Bayes** - looks a lot like probability. Actually, it is probability - make sure variables are independent when using it though.





Classification

Let's take a walk through an example of classification using several different methods:

https://colab.research.google.com/github/HSV-AI/presentations/blob/master/2019/190213/MUSK%20Classification.ipynb

Clustering

Clustering can be thought of as the opposite of classification. The classifications (or groupings) are not known ahead of time, and are discovered by using machine learning.

Some algorithms used here need to be given the number of clusters to break the data into, while others take other parameters and identify some set of clusters based on them.

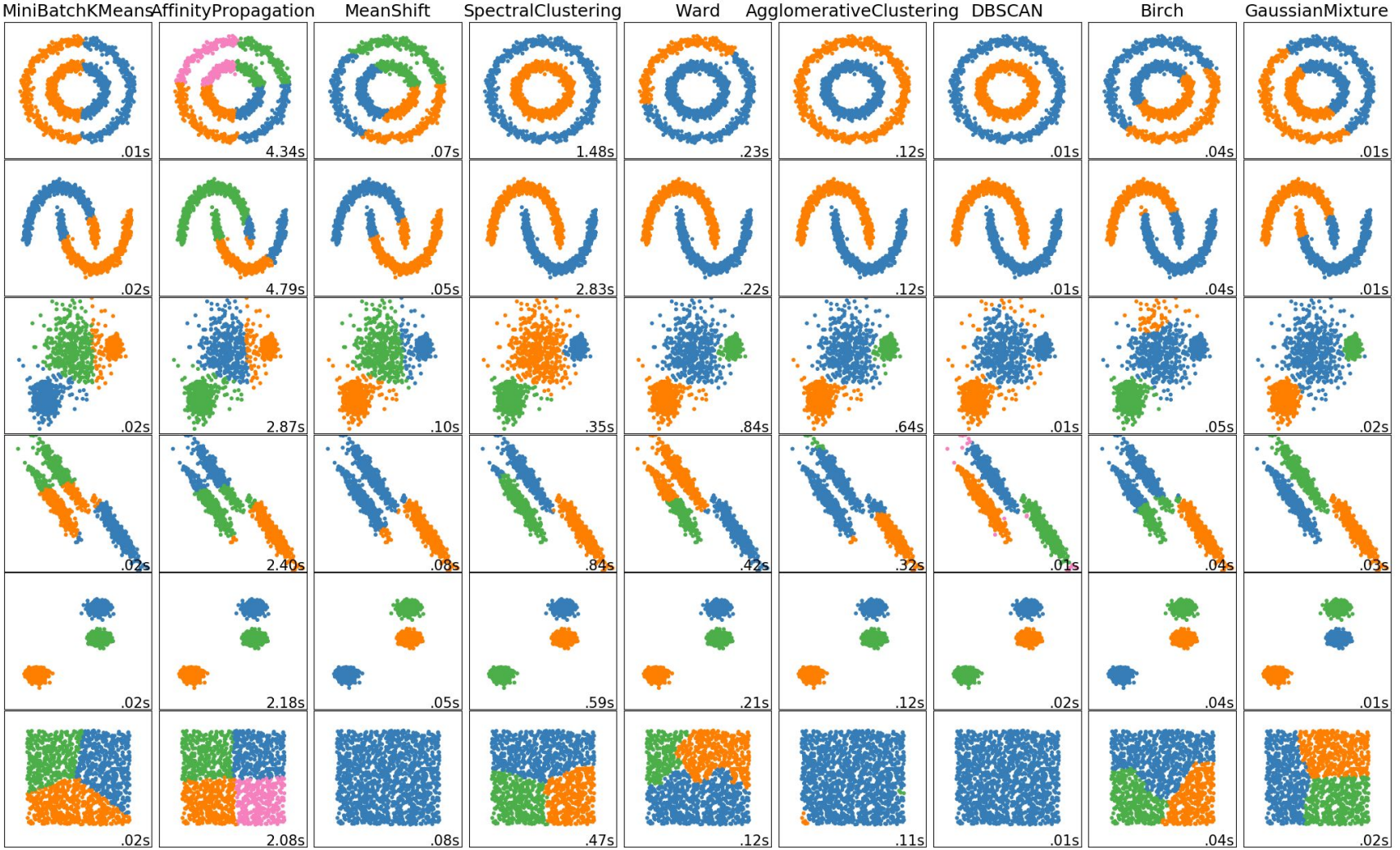
Clustering

Here are a few of the many algorithms used for clustering:

- **K-Means** - takes a number of clusters as an input, and computes the cluster boundaries based on the distance between points.
- **Affinity Propagation** - also known as “Nearest Neighbor”. Takes a set of parameters and computes the cluster boundaries based on the nearest neighbor graph distance.

There are a lot of other algorithms as well, but all seem to use either the distance between points or nearest neighbor graph distance.

From Scikit-learn Website



Optimization

Optimization algorithms helps us to minimize (or maximize) an Objective function (another name for Error function) E(x) which is simply a mathematical function dependent on the Model's internal learnable parameters which are used in computing the target values(Y) from the set of predictors(X) used in the model.

https://towardsdatascience.com/types-of-optimization-algorithms-used-in-neural-networks-and-ways-to-optimize-gradient-95ae5d39529f

Optimization

- **Cross-Validation** - Splits the data into training, test, and validation sets to train the model while avoiding over-fitting.
- **Hyperparameter Tuning** - In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. ...Hyperparameter optimization finds a tuple of hyperparameters that yields an optimal model which minimizes a predefined loss function on given independent data. Thanks again to Wikipedia!