Regression vs. Classification Primer

Author: Adam Farley

Supervised Learning

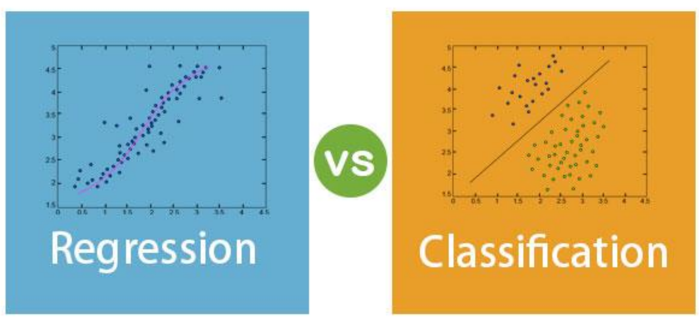
Goal of Supervised Learning

Supervised learning is one of the largest and most common branches of Machine Learning, the goal of Supervised Learning is to approximate a function which models a system based on past input-output pairs. The key is that outputs must be known. A large dataset will be comprised of these input-output pairs and fed into a Supervised Learning algorithm which continually guesses new functions to hopefully minimize the difference between its guesses and the actual output labels.

For a more in depth look at Supervised Learning, click here for a video by Andrew Ng (one of the goats of AI).

Difference between Regression and Classification

There are two main subsections of Supervised Learning: Regression and Classification. Because they're both forms of Supervised Learning, Regression and Classification both aim to approximate functions that map input variables to an output variable. The difference between these two are the type of data they're trying to predict. Regression tries to predict continuous data, while Classification predicts discrete data.



Discrete data is countable data. This means that there's a countable amount of options that the Classification algorithm can predict (think about the natural numbers) for a given set of inputs. Some examples of classification problems include classifying handwritten digits, designating a breast cancer tumour as malignant or benign, or approving/disapproving credit card applications.

Continuous data is measurable data. This means there's an uncountable amount of options that the Regression algorithm can predict (think about the real numbers) for a given set of inputs. Some examples of regression problems include predicting housing prices, predicting stock prices, and predicting the height of a person.

Common Algorithms

The following section is a collection of videos and articles pertaining to some of the most common regression and classification algorithms.

Types of Regression

Linear Regression

The most basic kind of regression. It determines a line of best fit for your dataset. If there are multiple independent variables, you could be predicting a plane or even hyperplane of best fit.

- Andrew Ng Video: https://www.youtube.com/watch?v=kHwlB_j7Hkc
- Wikipedia page: https://en.wikipedia.org/wiki/Linear_regression

Polynomial Regression

An extension of linear regression that determines n^{th} degree polynomial of best fit.

- Mike Cohen Video: https://www.youtube.com/watch?v=QptI-vDle8Y
- Scikit-Learn Article: https://scikit-learn.org/stable/modules/linear_model.html#polynomial-regression-extending-linear-models-with-basis-functions
- Wikipedia: https://en.wikipedia.org/wiki/Polynomial_regression

Ridge Regression (L2 Regularization)

A common form of regularization in linear regression meant to address multi-collinearity in the independent variables.

- Wikipedia (kind of confusing): https://en.wikipedia.org/wiki/Tikhonov_regularization
- Less Rigorous Article on Ridge Regression (covers Lasso regression as well): https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/
- StatQuest Video: https://www.youtube.com/watch?v=Q81RR3yKn30

Lasso Regression (L1 Regularization)

Almost the exact same as ridge regression except it uses the absolute sum (or L1 Norm) instead of the Euclidean norm (L2 Norm).

- Wikipedia: https://en.wikipedia.org/wiki/Lasso_(statistics)
- StatQuest Video: https://www.youtube.com/watch?v=NGf0voTMlcs

ElasticNet Regression (L1 and L2 Regularization)

Literally just Lasso and Ridge Regression at the same time.

- Wikipedia: https://en.wikipedia.org/wiki/Elastic_net_regularization
- StatQuest: https://www.youtube.com/watch?v=1dKRdX9bflo

Types of Classification

Logistic Regression

By finding the logistic function ($f(x) = \frac{1}{1+e^X}$) of best fit, you can determine the best boundary to separate types of data.

- Wikipedia: https://en.wikipedia.org/wiki/Logistic_regression
- Andrew Ng: https://www.youtube.com/watch?v=-la3q9d7AKQ

Naïve Bayes

Using the Bayes Theorem to predict the conditional probabilities of which type each data point is.

- Video Explanation: https://www.youtube.com/watch?v=CPqOCI0ahss
- Article: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Support Vector Machines

An algorithm that calculates the line/plane/hyperplane that separates different types of data while also being the furthest away from all data points.

- TowardsDataScience: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
- Video Explanation: https://www.youtube.com/watch?v=1NxnPkZM9bc&t=147s

Decision Trees

A ton of if-else chains, but it's fancy.

- Wikipedia: https://en.wikipedia.org/wiki/Decision_tree_learning
- Video Explanation: https://www.youtube.com/watch?v=DCZ3tsQIoGU

k-Nearest Neighbor

New data is classified based on the k closest data points in a pre-existing dataset.

- TowardsDataScience: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
- Video Explanation: https://www.youtube.com/watch?v=UqYde-LULfs

Neural Networks

Not going to get into this because there are entire sections of this notebook dedicated to NNs but it's worth pointing out that the most basic neural nets were originally used for classification problems.