**Regression Versus Classification**

When dealing with a data set, the first thing you want to determine is whether you are dealing with a regression problem or a classification problem and then choose the most appropriate model to your problem. Let's jump into the classification versus regression tutorial.



What we are going to cover:

* Classification
* Binary Classification
* Multiclass Classification
* Algorithms for Classification
* Choosing a Machine Learning Algorithm
* Regression
* Algorithms for Regression
* Choosing a Machine Learning Algorithm for Regression

# Classification Versus Regression - Classification

A classification problem occurs when we want to assign an observation into a predefined group or class. We do that by choosing a classifier. A classifier is a classification technique or a mathematical function that maps input data to a class. It does that by classifying the observation to the class with the highest probability.

# Classification Versus Regression - Binary Classification

As the name suggests, for binary classification there are only two classes to which we can assign our observations. Examples are:

* Medical Diagnosis (Heart Disease or no heart disease, diabetic or not diabetic)
* Email spam detection
* Credit card fraud (being credit worthy or not)
* Titanic survivors (If you were a passenger on the titanic did you survive or die)

# Classification Versus Regression - Multiclass Classification

For a multiclass classification, we have three or more classes for which we can assign our classifications. Examples are:

* We must classify a set of images of vehicles which are bicycles, motor bikes, and motor scooters into one of three possible categories.

# Classification Versus Regression - Algorithms for Classification

* Logistic Regression
* Linear Discriminant Analysis
* Quadratic Discriminant Analysis
* K-Nearest Neighbours
* Tree-Based Methods
* Support Vector machines …

and many more.

If you want to see logistic regression, linear discriminant analysis, k-nearest neighbors, and random forest in action, check out my [titanic tutorial](http://thatdatatho.com/2018/09/18/titanic-data-set-increased-prediction-scores-82/) where I implemented these methods. [Part 2 is here](http://thatdatatho.com/2018/09/25/the-titanic-data-set-and-the-woman-child-model-82-test-set-accuracy/).

# Classification Versus Regression - Choosing a Machine Learning Algorithm

The truth is that there is no single best classifier for which our test set error is smallest. We first must explore our data and see, whether our decision boundaries are linear or quadratic, how many observations we have in our data set and if we have a binary or multiclass classification problem etc. Based on these information, among many others, we either choose our classifier that fits best our assumptions or we apply a couple different classifiers and choose the one with the lowest test set error.

# Classification Versus Regression - Regression

We are interested in regression when wanting to predict a quantitative response. Examples are:

* Predicting the value of a house
* Predicting the college GPA based on a student’s high school GPA, studying habits etc.
* Predicting the crime rate in a certain region

# Classification Versus Regression - Algorithms for regression

* Simple Linear Regression / Multiple Linear Regression
* Ridge Regression/Lasso
* Polynomial Regression
* Regression Splines
* Principal Component Regression…

and many more.

Implementation of a multiple linear regreesion model can be found [here](http://thatdatatho.com/2018/04/30/multiple-linear-regression-model-building-r-tutorial/), and for a lasso model [here](http://thatdatatho.com/2018/05/07/the-lasso-r-tutorial-part-3/).

# Classification Versus Regression - Choosing a machine learning algorithm

Choosing an appropriate regression technique, again, highly depends on the data at hand. Questions we may want to answer is if we have constant variance among the residual. If not, we can try a polynomial regression or some other transformation on the features. When we have a data set that has high variance we may want to consider ridge regression or the lasso which shrinks our variance by shrinking our coefficient estimates. As for classification, the same principle applies for regression. That is, there is no single best algorithm and we have to try a couple in order to see which one is most appropriate.

# Classification Versus Regression - Classification versus Regression Summary

When the response variable is qualitative then we are dealing with a classification problem and when our response variable is quantitative we are dealing with a regression problem. When the response variable is encoded as discrete values (0, 1, 2, 3…) we are dealing with a classification problem. For our Titanic classification problem, passengers who died have a value of 0 and people who survived have value 1 for example. For a regression problem, the response variable takes on continuous values (2.3, 100, 200.9…). When the college GPA is our response variable, it can take on values between 0 and 4. For example 3.65 or 2.453. These values are not discrete anymore.

I hope you have enjoyed this blog post. If you have any suggestions or feedback, write it in the comment sections below. Thank you.