Logistic Regression

The goal of this machine learning task is to attempt to describe a function that describes the training data, that way a predictions of the dependent variable can be made by inputting the data into that function. Hence, what makes most sense is to use a classifier that implements a regression. As can be expected however, stock data is linear, and nor should it be, as a stock's attribute value often is determined by an interaction with those of other attributes. The classifier our project implements in Weka is a logistic regression. This is similar to a typical linear regression classifier, but with the additional benefit of being able to handle nonlinear attribute behavior.

To help illustrate how the attribute interactions are handled in the regression function, an example is given where there are only two attributes:

$$Y = \theta_1 + \theta_2 X_1 + \theta_3 X_2 + \theta_4 X_1 X_2$$

Here, the nonlinear interaction between the two attributes is represented by X_1X_2. In a model with many attributes, every combination is enumerated, and as can be imagined, some attribute interactions are vital to the accuracy to the function calculation while others are irrelevant. The objective of the machine learning algorithm during training is to decide how important each interaction is to the end classification, and the importance is then reflected by the individual theta values in the theta vector.

Data and Attributes

All data was calculated from Yahoo Finance. To get the data, we implemented methods in the pandas and yahoo-finance module available for Python. The data set consists of data from 2014-2015 for every stock listed on the NASDAQ, NYSE, and AMEX for which data was available. This gave us about 5000 data points with which to train and cross-validate.

Due to the nature of the task, all attributes were required to stem from the market and stock's historical price and volume data. From these two data sets, we were able to generate other attributes that had implications on the stock's statistics and health relative to the market. The other data attributes included were:

- Returns: Price data is obviously not a good attribute because stocks greatly vary in price. In order to standardize this scale, we use a stock's returns over the specified time period. The returns are calculated as log(Price2/Price1). In addition to the weekly returns, specify the returns of the stock over the past year, 9 months, 6 months, and 3 months

- Volatility: This is computed as the standard deviation of the price array. In order to keep this at scale, we use a percentage.

- Sector Movement: Each stock is categorized as being in a given sector such as technology, finance, consumer service, etc. To gage the movement of the sector, we used the S&P500 Sector ETF price data to calculate its yearly, 9mo, 6mo, and 3mo returns.

- Beta: The beta of a stock gages the health of stock against the market's. To get the returns of the overall market, we used the calculated returns of SPY, the S&P500 ETF. The formula for a stock's data is as follows $\frac{cov_{X,SPY}}{\sigma_{SPY}^2}$