Y6380872

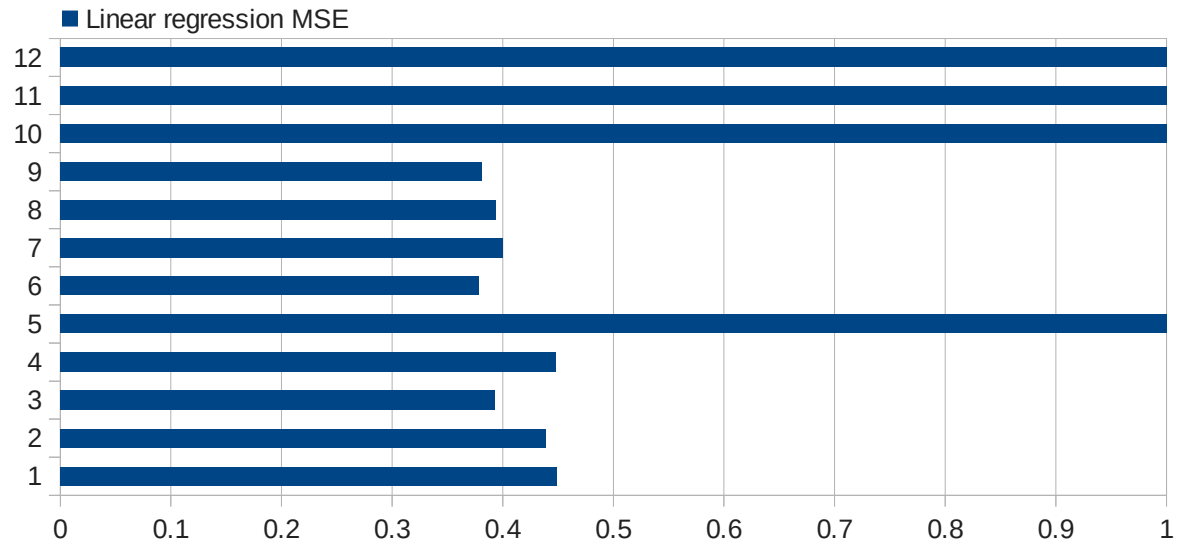
Linear Regression and Logistic Regression Report

Task 1

In my experimentation of searching for different features, I began with using the stock price and stock volume for the past 10 days, which is 20 coefficients in total. I added a constant as a feature and noticed no significant improvement. I decided that having this many features will likely result in overfitting the model and will take much longer to minimise the function. I chose to reduce the number of features three or fewer and investigate the results. As shown in the table, there is a correlation between the size of the feature set and the mean squared error (MSE).

I noticed that features that did not contain any stock price variables were significantly worse than those that did, as shown by the stock volume for the past 10 days resulting in a MSE of over double that of the stock price for the last 10 days.

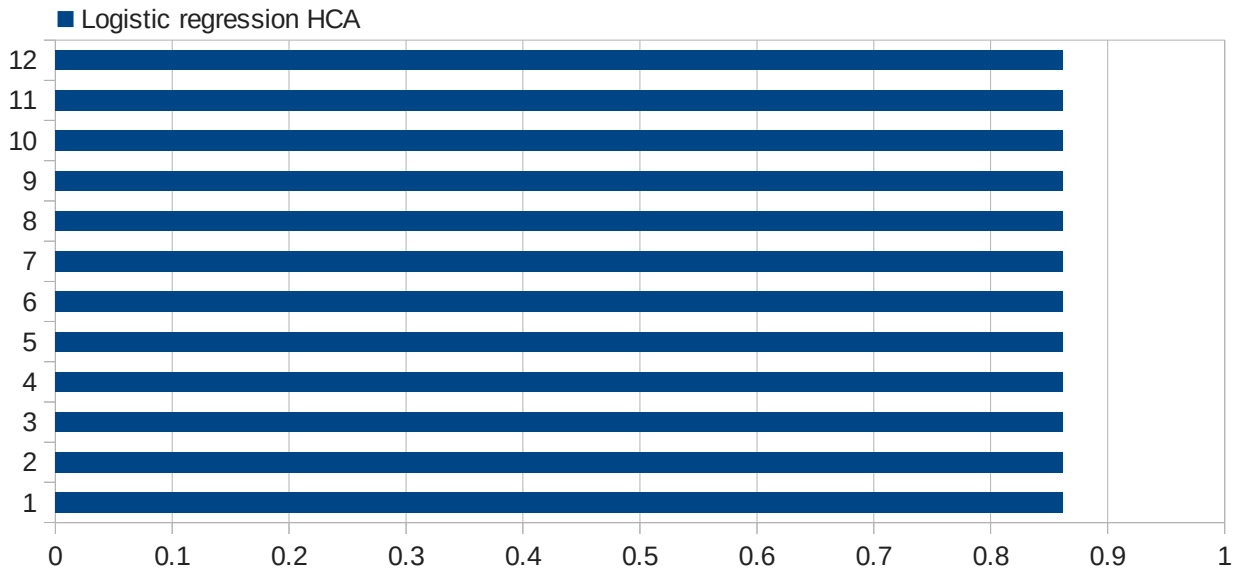
After investigating how stock prices are predicted for real data, I learned that it is generally considered that stock prices obey a random walk model, in which the price cannot be predicted. Therefore the best feature set would be the set consisting of just yesterday's stock price, because yesterday's stock price will be statistically close to today's price. Any further additions to the feature set would waste computation time and add no value.



Task 2

For the logistic regression task, I began with a similar approach to the first task. I noticed that the Hard Classification Accuracy was always the same as I chose different feature sets. I discovered that the classifier was assiging class 0 (no change) to every row. After calculating the actual class for each row, I found that 86% of the rows are in class 0.

This finding is justified by the random walk hypothesis – if stocks are completely unpredictable, then the best classifier is the one which classifies rows into the largest class.



Feature set ID	Features	Feature set size	Linear regression MSE	Linear regression with ridge regularisation MSE	Logistic regression HCA	Logistic regression with ridge regularisation HCA
1	The stock price and stock volume for the past 10 days	20	0.449128355944	0.430808081342	0.861593059937	0.861593059937
2	The stock price and stock volume for the past 10 days, and a constant	21	0.438747935306	0.40384994783	0.861593059937	0.861593059937
3	The stock price and stock volume for the past 5 days	10	0.392850604705	0.406150817447	0.861593059937	0.861593059937
4	The stock price for the past 10 days	10	0.44799268497	0.408629435558	0.861593059937	0.861593059937
5	The stock volume for the past 10 days	10	90.3147484773	87.7251631692	0.861593059937	0.861593059937
6	The stock price and stock volume for	2	0.378970416044	0.379727547325	0.861593059937	0.861593059937

	yesterday					
7	The stock price and stock volume for yesterday, and a constant	3	0.399939022319	0.383059692924	0.861593059937	0.861593059937
8	The stock price for yesterday	1	0.393458954187	0.386082759328	0.861593059937	0.861593059937
9	The stock price for yesterday, and a constant	2	0.38110367946	0.379817557271	0.861593059937	0.861593059937
10	The stock volume for yesterday	1	115.34129917	114.402270993	0.861593059937	0.861593059937
11	The stock volume for yesterday, and a constant	2	44.9960893583	45.2679976773	0.861593059937	0.861593059937
12	A constant	1	68.2875589776	68.2978476654	0.861593059937	0.861593059937

Task 3

I implemented ridge regularisation model for this task. I chose ridge over lasso because I had already determined the optimum feature set.

After running the regularisation code, I found that it usually finds a lower MSE than standard linear regression. I ran the ridge regularisation program multiple times in order to find the best lambda, and found that it was falling between 0.01 and 0.1. This indicates that a better MSE is obtained by giving more weight to the regularisation parameter, which implies that overfitting is occurring.

The ridge regularisation made no difference to the logistic regression results. This is because the best classifier has already been obtained, assuming the data abides by the random walk hypothesis mentioned earlier.