Capstone Project 1 – Inferential Statistics
Springboard Data Science Track
Kamran Ossia
2018-09-12

This project deals with the question of predicting daily stock prices using the Keras deep learning software with a Tensorflow backend and GPU acceleration. The prediction is based on past daily closing values of the same stock price, i.e. no other series such as volatility index, cumulative tick, volume, etc. are used. The following questions are addressed in this report.

1. Are there variables that are particularly significant in terms of explaining the answer to your project question?
   The price data consists of four series: open, high, low, and close. In general, the closing daily price is considered the most significant and most technical analysis is based on it only. High and lows for each day are more volatile than the close, as day traders tend to revert prices to the average by taking profits on the intraday movements. Of course if there are larger forces moving the market, a trend may last through the day and over longer periods.

2. Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?
   Two possible applications of the stock price forecast are the prediction of the closing price of a day from the opening price, and the prediction of the opening price from the closing price of the previous day. Taken as vectors, opening and closing price series may well be correlated.

3. What are the most appropriate tests to use to analyze these relationships?
   The most straightforward measure of the relationship between two time series is the correlation:

$$\rho(x_1,x_2)=\mathrm{E}[(x_1-\mu_{x_1})(x_2-\mu_{x_2})] \,/\, \sigma_{x_1}\sigma_{x_2}$$

   Where μ is the sample mean, σ is the standard deviation, and $\mathrm{E}[]$ is the expectation operator. In Python, the pandas corrcoef() function provides a simple method to calculate the correlation. For example, for the two-year daily TSLA stock data, we have the following correlation matrix for the open, high, low, and close series:

```
array([[1.        , 0.99664986, 0.99649019, 0.99161029],
       [0.99664986, 1.        , 0.99610778, 0.99674407],
       [0.99649019, 0.99610778, 1.        , 0.99635803],
       [0.99161029, 0.99674407, 0.99635803, 1.        ]])
```

   We can see that the four series are highly correlated, but the least correlated are the open and close.