CS-C3240 Machine Learning D

Machine Learning Project – Stage 2 Submission

Changes of stock prices based on COVID-19 infections

Problem formulation

The aim is to see if I can find out a Finnish company's average stock price and the number of trades based on how many people in Finland were infected with coronavirus that week.

The datapoint is a week, the features of which are the reported COVID-19 infections in Finland that week. The labels are the average price of the company's stocks and how many trades were made that week. The labels for training and validating will be calculated from the original dataset. Another possibility is the addition of the company in question as a feature, to train a single model for multiple companies.

The companies whose data will be used will most likely be Finnish, as I doubt Finland's infection rate has any noticeable effect on larger, global companies. The data for the companies' stock prices is acquired from Nasdaq's Nordic site for shares. The data for coronavirus infections in Finland is from THL. I've included the links to these sources in the following chapter.

Methods

Dataset

A data point is a week, the features of which are COVID-19 infections recorded in Finland that week. Another feature that I've considered is the number of COVID-19 tests. Both tests and infections are represented by integers.

The data on COVID-statistics is downloaded from the following link: https://sampo.thl.fi/pivot/prod/fi/epirapo/covid19case/fact_epirapo_covid19case;jsessionid=E9404493A2F291581061106F6320DF10.apps5?row=dateweek20200101-509030&column=measure-444833.445356.492118.&fo=1 (I suggest downloading the excel file and saving in excel as csv, as each feature is separated into different rows when directly downloading a csv-file)

At the time of writing this report, there are a total of 115 different weeks, or datapoints in the file, from the beginning of 2020 to the present day. Although the current week is shown, the number of infections is zero, which likely means the data for that week is incomplete. While there is a value for each case of COVID-infection, the first 8 weeks are missing values for the number of tests for COVID-19. The values for deaths caused by COVID-19 are also missing in the first 9 weeks, but I don't intend to use them anyway. At such an early stage of

the pandemic the missing value could be zero. If test amount is used as a feature, the first 8 weeks will have to be removed, resulting in 106 datapoints.

The labels of a data point are the weekly average stock price of a company, which is a floating-point number, and the number of trades that have occurred during the week, which is an integer. For the dataset, only the weeks that are also present in the COVID-19 file will be looked at.

The data for a company's stock price can be downloaded from www.nasdaqomxnordic.com/shares

The datapoint labels can be formed by downloading the data between 30.12.2019 (the first week of 2020 starts on this day) and 4.3.2020, last week's Friday. This may have to be reordered somehow, since the most recent statistics are presented first. As there is no data for Saturdays and Sundays, the labels are calculated by counting the average for every five days' average price and by summing the trades for every five days. This way there are 116 of each label. More could be downloaded, but there are no features available earlier. The downloaded dataset also has data for the bid, ask, opening price, high price, low price, closing price, total volume and turnover.

The features are chosen to see if stock prices and trades for some companies can be predicted from the number of COVID cases and tests. Because the data is on Finland, the chosen companies to use could be, for instance, Valmet Oyj, Terveystalo Oyj, or Kesko Oyj B.

Multiple linear regression model

In this method the outcome of a dependent variable is predicted from two or more variables. The formula or hypothesis space is a function of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in} + \epsilon$$

Where is the value of the label, β_0 is the y-intercept, the other βs are the regression coefficients and is ϵ the error term. The values of different features are x and the value of the label is y. To determine the coefficients, the mean squared error must be minimized.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

This model was chosen because there are multiple features, and because it is available in scikit.

Validation

The dataset will be divided into a training, validation and testing set. Because the dataset is relatively small, most (or around 50%) of the dataset will be used for training, while the rest of the data will be split equally into validation and testing sets.

For forming these sets, every other data point could be taken into the training set, as both the number of infections and the number of tests seems to grow towards the end of the dataset. This way more variance in the feature values will be covered with the training set.