COM502: Machine Learning

Term Paper

Question 1. You are given a set of stock price returns with varying time span. Three of which are named by first, last, and full. In all the attached data sets, the first column is for the composite index, and all other 10 columns are individual stocks. In other words, those 10 columns should explain some part of the composite return. A quick glance confirms that, for some period, returns have been distributed in approximation of Gaussian, as is argued by many finance papers, but the Gaussianity does not cover for other periods.

Since you are aware that Gaussian distributed target variable does not require extensive non-linear fitting, you expect the simple regression between the composite index and all individual stocks should be nearly identical to Maximum Likelihood estimation. On the other hand, for the non-Gaussian case, you believe there are rooms to exploit the non-linearity of the fitted regression by techniques that you learned from COM502: Machine Learning, along with other courses at SIAI.

1. For which period, among three attached files' time span, do you find the Gaussianity assumption fails? For the test of Gaussianity, provide Q-Q plots for an eye-ball test and Kolmogorov-Smirnov test statistics for more scientific conclusion.

2. You notice that three files have different length of time, so you want to double check if the length of stock return period may affect Gaussianity. How would you construct the test? First, you are recommended to divide the periods into sub-intervals of your choice that you may find it critical between Gaussian and Non-Gaussian, as is for Independent Component Analysis. Then, create a rolling statistics of 1st to 4th moments to see if your conclusions are identical, or at least similar, to sub-interval tests. In terms of cost-benefit analysis (i.e. computational efficiency), which one do you prefer?

3. For the period where you find the non-Gaussianity, you question if it is because of some common outlier across the economy that may have created spiral effect, i.e. strong negative correlation during downturn (such as COVID-19 pandemic, global financial crisis). Given the mathematical proof in the class that sum of Gaussian distributions are also Gaussian, you have constructed a hypothesis if Gaussian Mixture Model (GMM) can be used to separate 1.underlying Gaussian, and 2.negative correlation. What do you need to estimate for this hypothesis testing? How do you link this construction with (G)ARCH models, a financial volatility model that captures movements of variance?

4. Among many other required information for constructing the test, you would like to apply PCA to your data. Are some of the stocks have significant joint movements? If so, can you consider them in the same industry? If not, what are the causes of such correlation? Whether it is industry effect or some unknown source of correlation, are they consistent in all your sample time span? How sporadic is it? If you have to choose K number of variables for PCR, upto what order of eigenvectors in the model? Why?

5. To back up your argument above, you would like to use LDA (Linear Discriminant Anaylsis) for re-mapping the data into a proper hidden dimension. However, the time series itself does not help you to apply LDA, or at least it does not give you what you intended. Instead, you can create a set of representative values (such as mean, variance of sub-intervals) for LDA. Defend your modification as much as you can.

6. By above two steps, you have more understanding of the return structure. Although there are individual stock specific components, you believe, there are common trends, which you can call a composite macro-economic factor and industry specific factors. Since you do not have industry naming, you can name the factors by industry 1, 2, 3 and so on.

7. In finance, portfolio choice theory says a good diversification can help removing idiosyncratic risks, or company specific risks for non-financial experts. By above, can you confirm the theory? If so, assuming that you have a well-diversified portfolio, upto what order of principal components do you carry? Relate

this results to factor analysis in words. For non-Gaussian intervals, can you rationally back-up your argument? In your logic, you are recommended to follow the idea that PCA is a special case of factor analysis.

8. Now with way better understanding of the data structure, you would like to build a multi-task learning model for stock return prediction for varying scenarios with specific considerations for which standard Gaussianity assumptions are challenged. Construct the model with non-linear regression and regularization parameters. Make sure to choose your non-linearity in conjunction with Gaussian vs. Non-Gaussian arguments. First construct the model with raw data and compare it with PCR with your choice of number of PCs. Which one do you prefer?

9. Instead of targeting for the perfect fitting for a continuous variable, what would happen if your target variable is simple + and - (essentially 1 and 0)? In addition to that, what if you change the target value as + and - by margins to the previous stock trading day (think of it as momentum), how would you change your models above? Can logistic regression be outperforming than other models? What about SVM or tree-based models? Why or why not? In your comparison, provide a confusion matrix of each model result.

10. Can you achieve any better conclusion by artificial neural network in terms of predicting stock price movements? Instead of wasting huge computational cost, to find some insights, 1.you normalize (not scaling) all your data, 2.PCA, 3.de-normalize, and 4.use it for prediction. For step 4, you have to differentiate the kernel (or activation) function depending on the target variable, whether continuous or discrete for 8) and 9). In comparison to sub-interval modeling choices above, what do you expect to earn from ANN? In which sub-intervals do you find the model perform better? How do you interpret the performance differences? Do you think any deeper network can outperform in terms of predictions? In words.

Bonus. After all the hard work, for non-Gaussian sub-intervals, does your model perform any better than simple ARMA(p,q)?, for example, $y_t = \alpha y_{t-1} + \epsilon_t$?

MSc only. (This is 30% of your Regression Analysis III final exam.) In the data, all stock return processes suffer from common shocks to the aggregate economy, and since some companies in our universe are competitors in the same industry, there might be joint effect that above model discussions have missed. You hypothesize if VAR (Vector AutoRegression) can be of any help. Given the imaginary industry division that you have chosen above, construct an SUR for each regression representing each industry. In other words, you use company stock returns of the same industry for regressors and the composite index as the dependent variable for one industry. If you have 4 industries, for example, you will have 4 equations. In your argument, make sure that all unit root or co-integration processes are dealt with. In terms of predictability, do you find any gain against above multi-task learning model with raw data? or PCR version?

In your answer, provide necessary arguments and graphs. For questions that mathematical derivation can sharpen your argument, precise intuitive reasoning can be sufficient. Good luck with the term paper.