INDENG 290: Problem Set 3

Due: November 28, 2022 at 3 pm PST

Homework Collaboration Policy. Please feel free to discuss the homework problems in groups if you prefer, however, you must write/code your solutions independently. All code must be written in Python and submitted on Gradescope along with the rest of the assignment.

Problem statement. The goal of this homework is to familiarize students with their choice of either time series prediction methods or generative adversarial networks via performing mini research projects. To get full credit, you need to submit solution to either Problem A (time series prediction methods) or Problem B (generative adversarial networks) described below - not both! Both Problems A and B assume some independent reading and literature research - lecture notes as well as papers provided in the syllabus will be helpful; Problem B assumes independent study of Pytorch or Tensorflow packages for training neural networks (it is only required for extra credit in Problem A). Below are the instructions how to install Pytorch locally and some Pytorch tutorials; similarly, instructions how to install Tensorflow locally and some Tensorflow tutorials.

Problem A. Financial time series prediction (100 points)

- Get daily stock data for AMZN, GOOG, AAL, NCLH from Yahoo Finance for 2016, 2017, 2018, 2019, 2020. You can use below code as an example, more examples of using yfinance library can be found here.

```
import yfinance as yf
import matplotlib.pyplot as plt

data = yf.download(
tickers=['NVDA'],
# use "period" instead of start/end
# valid periods: 1d,5d,1mo,3mo,6mo,1y,2y,5y,10y,ytd,max
# (optional, default is '1mo')
period="1d",
# fetch data by interval (including intraday if period < 60 days)
# valid intervals: 1m,2m,5m,15m,30m,60m,90m,1h,1d,5d,1wk,1mo,3mo
        # (optional, default is '1d')
interval="1m")

# Plot the close prices
data.Close.plot()
plt.show()
```

- Predict daily stock volumes.

  1. (10 pts) Plot time series for stock volumes and close prices for the above time periods. List observations of the data patterns - what kind of properties should a model have in order to be able to predict stock volumes and close prices well? Comment on the distributional shift observations in 2020 - how would you enhance your models for 2020 to improve performance?
  2. (30 pts) Using N-day sliding window, use N-day average and N-day median methods to

predict daily stock volumes for $N+1$st day in 2019 and 2020 for $N = 10, 30, 60$, namely:

$$y_{N+1} = \frac{y_1 + y_2 + \ldots + y_N}{N}$$
$$y_{N+1} = \text{median}(y_1, y_2, \ldots, y_N)$$

Analyze prediction error compared to realized volumes on the same days: compute average mean square error by month. Also, calculate mean square error for banking holidays vs ordinary business days. Do you observe any patterns which N works best? Can you comment why? Do you see any difference across different stocks? Elaborate on your findings. You'll likely notice that mean square error will be smaller for ordinary business days than for banking holidays. You'll also likely notice increase in mean square error during the distributional shift due to the Covid shock in 2020.

3. (30 pts) Daily volumes are often forecast using linear autoregressive models. Using N-day sliding window, find coefficients A, B, C in linear autoregressive models of lag 1 and lag 2 below to predict daily stock volumes for $N+1$st day in 2019 and 2020 for $N = 10, 30, 60$. Specifically:

$$y_{N+1} = Ay_N + B + \epsilon_{N+1}$$
$$y_{N+1} = Ay_N + By_{N-1} + C + \epsilon_{N+1}$$

Do you think models of higher lag would be necessary? Why? Do you observe any patterns which N works best? Do you see any difference across different stocks? Repeat mean square error analysis above and comment on your findings with regard to ordinary business days vs. holidays as well as the distributional shift in 2020.

4. (30 pts) Propose a method to improve volume prediction for banking holidays - you might need to use data for 2016, 2017 and 2018 (and, perhaps, even earlier) for that. Repeat the mean square error analysis and justify why the method that you are proposing is superior to the above.

5. (10 bonus points) Use neural networks to improve daily volume forecast above. Training neural networks can be expensive, therefore, for the purpose of current exercise, we might not need to consider the entire two-year time period - pick a month or two and focus on improving forecast over classic time series models for that time period. When presenting your results, elaborate on the neural network architecture used, training data (eg., the choice of the size of training data), training details (hyperparameters used, etc), training loss, etc. Visualizations will be helpful. Were you able to "beat" the benchmark in prior exercise in terms of prediction error?

Problem B. Generative adversarial networks (100 points)

- Generate synthetic data that mimics distributional properties of the following datasets:

1. (45 points) Two-dimensional rectangular coordinate samples (x_1, x_2), with x_1 in the interval from 0 to 2π and $x_2 = \sin x_1$.
2. (45 points) Two-dimensional rectangular coordinate samples (x_1, x_2), where $x_1^2 + x_2^2 = 1$, and $0 \leq x_1, x2 \leq 1$.
3. (10 points) Two-dimensional polar coordinate samples $(\theta, \cos 2\theta)$, where θ ranges between 0 and 2π.

- For all of the above examples, elaborate on the neural network architecture used for generator and discriminator, training details (data, hyperparameters used, etc), progression of generator and discriminator training loss, etc. Visualizations will be helpful to illustrate the above. How many epochs did it take you to train the GAN to generate "reasonably" good synthetic data? Visualize output of the generator at multiple epochs throughout training. When did you decide to stop training for each case? Try different noise dimensions and see what is the minimal noise dimension (i.e. minimal dimension of the latent space) that is needed to learn the above well?

- (10 bonus points) Take close price time series for one stock of your choice from Yahoo Finance (Problem A) - pick period of time from 2015 to 2019 with reasonably "stationary" behavior before the macro shocks. Partition these time series to N samples size 60 days each - these are going to be your training samples. Now use GAN to generate synthetic time series that looks like your training samples. Elaborate on the training process as above and visualize it. Demonstrate how the 1-day return distributions (i.e. one of the stylized facts!) for synthetic data changed during the training.