**Task Description**

**Task 1: Price Prediction**

In this task, you are required to predict the midPrice of “Exchange A” in 5 seconds later based on the orderbook data gathered from 3 different exchanges.

**Data:**

There are 3 different exchanges, A, B and C. Each exchange has 3 days orderbook data, shown in the file name.

In each “.csv” file:

1. receiveTs: timestamp for each data records
2. Pa\_n: orderbook ask price at level n
3. Pb\_n: orderbook bid price at level n
4. Va\_n: orderbook ask volume at level n
5. Vb\_n: orderbook bid volume at level n
6. midPrice can be computed as midPrice = (Pa\_1 + Pb\_1) / 2

**Requirement:**

1. Present the end-to-end ML workflow (including data preprocessing and feature engineering). Note that the accuracy of the model is not the top priority in this task, as the dataset provided only covers a few days.

2. You can decide to make this problem as classification problem or regression problem. Explain your labeling mechanism in detail.

3. State the parameters chosen for your ML model clearly (either conventional ML model or Deep Learning model).

4. Python code should be submitted and re-runnable.

***Test with both regression and classification models.***

Utility functions: data\_preprocess, regression\_models, classification\_models, TimeSeriesCV

Main file: main\_midpice.py

**Task 2: User Categorization**

In this task, you are required to do user categorization based on users' data which is multivariate time series data. Note that the total number of user categories are unknown as well.

**Data:**

The "User" folder contains 17967 files. Each file contains a single user's data, and the file name is named by user\_id. Each user has a set of multivariate time series data presented in time order. Column 1-3 are 3 time series; Column 4 are timestamps. Do note that this data set contains missing observations.

**Requirement:**

1. Please design an end-to-end ML model (from data preprocessing to output generation) and explain your work in details.

2. Evaluate the clustering result.

3. The clustering result should be presented with an interactive visualization plot.

***Thoughts:***

*Conventional clustering techniques like KNN is not suitable for two reasons:*

1. *it only works on inputs of fixed size, not suitable for multivariate time series data with various sizes.*
2. *classic K-means based on Euclidean distance is not a good similarity measure for time series data*

*other approaches available:*

1. *spectral clustering to extract a new representation -> K-means*
2. *Deep Embedding Clustering (DEC) (cons: only work for fixed length)*
3. *DTW (Dynamic Time Warping) + K-means*
4. *Variational autoencoder (VAE) + Gaussian Mixture*
5. *Autoencoder (Conv1D + forward & backward GRU)*

*+ K-means, loss function based on Kullback-Leibler Divergence (or reconstruction loss)*

*Visualization:*

*PCA and t-SNE to visualize the clusters*

**Task 3: Trading alpha**

In this task, you are required to construct trading alpha(s) based on level-1 orderbook data.

**Data:**

There are 6 different tickers (A, B, C, D, E, and F) in 2D matrix, information provided are in .csv format with fixed time-interval:

1. ask.csv: ask price
2. bid.csv: bid price
3. askVol.csv: ask volume
4. bidVol.csv: bid volume
5. rtn.csv: price return (time adjusted)

Item 1 ~ 4 can be used for alpha(s) building, while item 5 is for evaluation purpose only.

**Requirement:**

1. Submit re-runnable Python code(s) with [a] input based on data provided and [b] output a target position in 2D matrix (datetime x ticker).

* Constraints:
  + For each interval absolute sum of the target\_position is less or equal to $1mil.
* Example:
  + def alpha\_gen(ask, bid, askVol, bidVol):

…

return target\_position

* + subject to sum( |target\_position| ) <= $1mil for each datetime.

2. You may choose using ML model, regression model, rule based model, and other to construct your trading alpha(s).

3. You can decide targeting multiple tickers or single ticker. Ticker(s) not being used can have zeros target positions.

4. Evaluation maximized utility functions ( and ):

* + Where at time t is defined as:
  + And at time t is defined as:
    - :
    - :

**Bonus Question: Poker Game**

There are two sets of cards – blue and gold. Each set consists of ten cards numbered from 1 through 10. Now two players are playing “poker” with these two sets. Rules are simple

1. Shuffle 20 cards and each player randomly pick up two cards.
2. A player with better combination of cards wins. If two players have the same cards regardless of their color, it is draw. Ex) Gold 9 – Blue 8 = Blue 9 – Gold 8
3. Grade of card combination is as below. All Consecutive numbers are higher than Pairs. And all Pairs are higher than Any Number Cards.

Cards scoring system:

1) Consecutive numbers: 10-1 > 9-10 > 8-9 > … > 2-3 > 1-2 > Pairs(10-10 > …)

2) Pairs: 10-10 > 9-9 > …. > 2-2 > 1-1 > High Number(10-8 > 10-7 > … )

3) Any card with high number: 8-1 > 7-4

4) If the high cards of both players are the same, it is consider as draw.

BQ1. If you have 9-Gold at your hand, what is your odd of winning?

BQ2. If you happen to have one extra card while the counterpart has only two cards, what is your odd of winning? In this case, you have three cards and you need to show the best combination out of three cards. For example, if you have Gold-1, Blue-2, and Blue-10, you may show Blue-10 and Gold-1