機器學習
於交易上的應用

簡介

背景知識

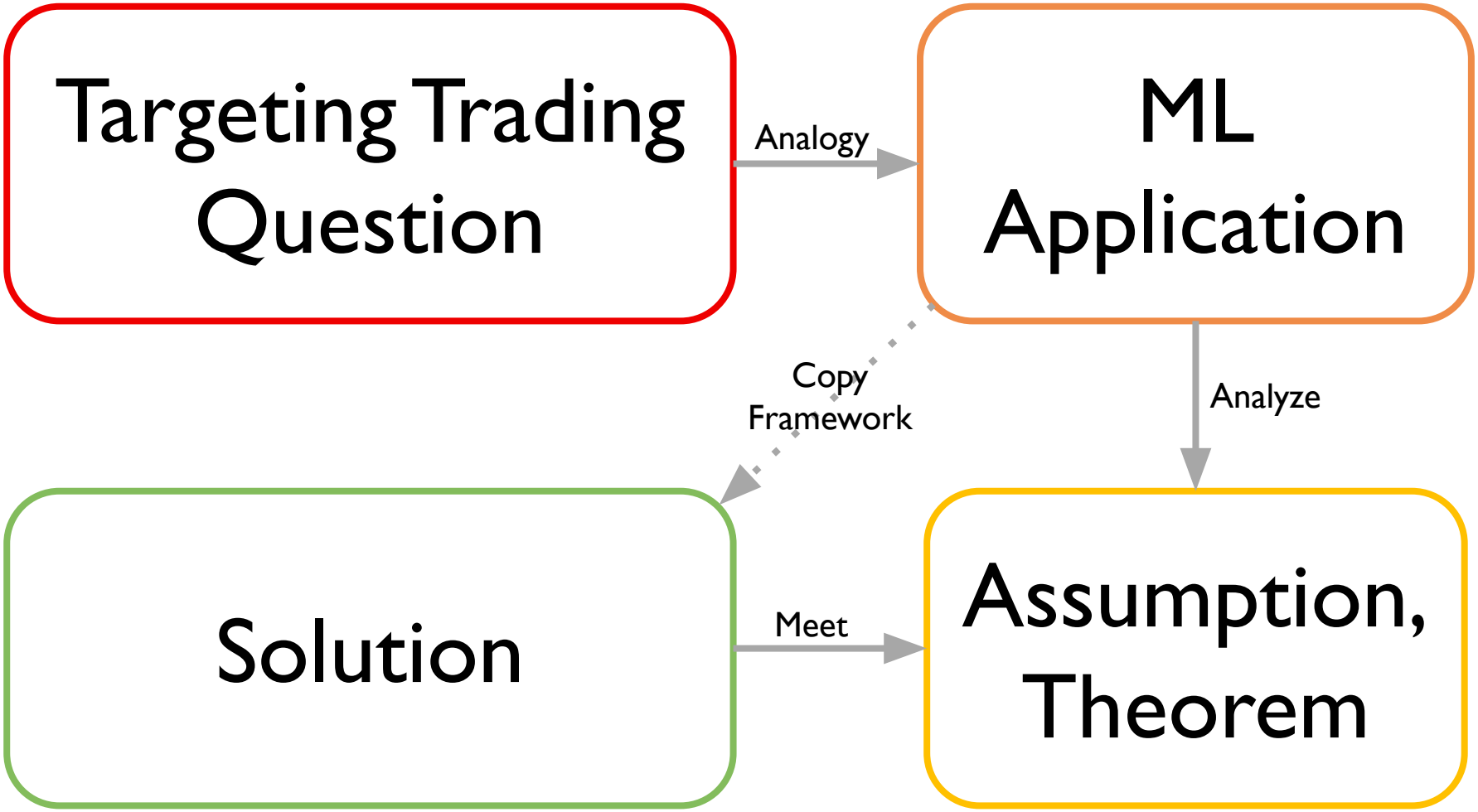
Trading Strategy + Machine Learning = Trash

- Factor Investment
- Asset Allocation
- Statistical Arbitrage
- Market Making
- Event Trading
- Trend Trading

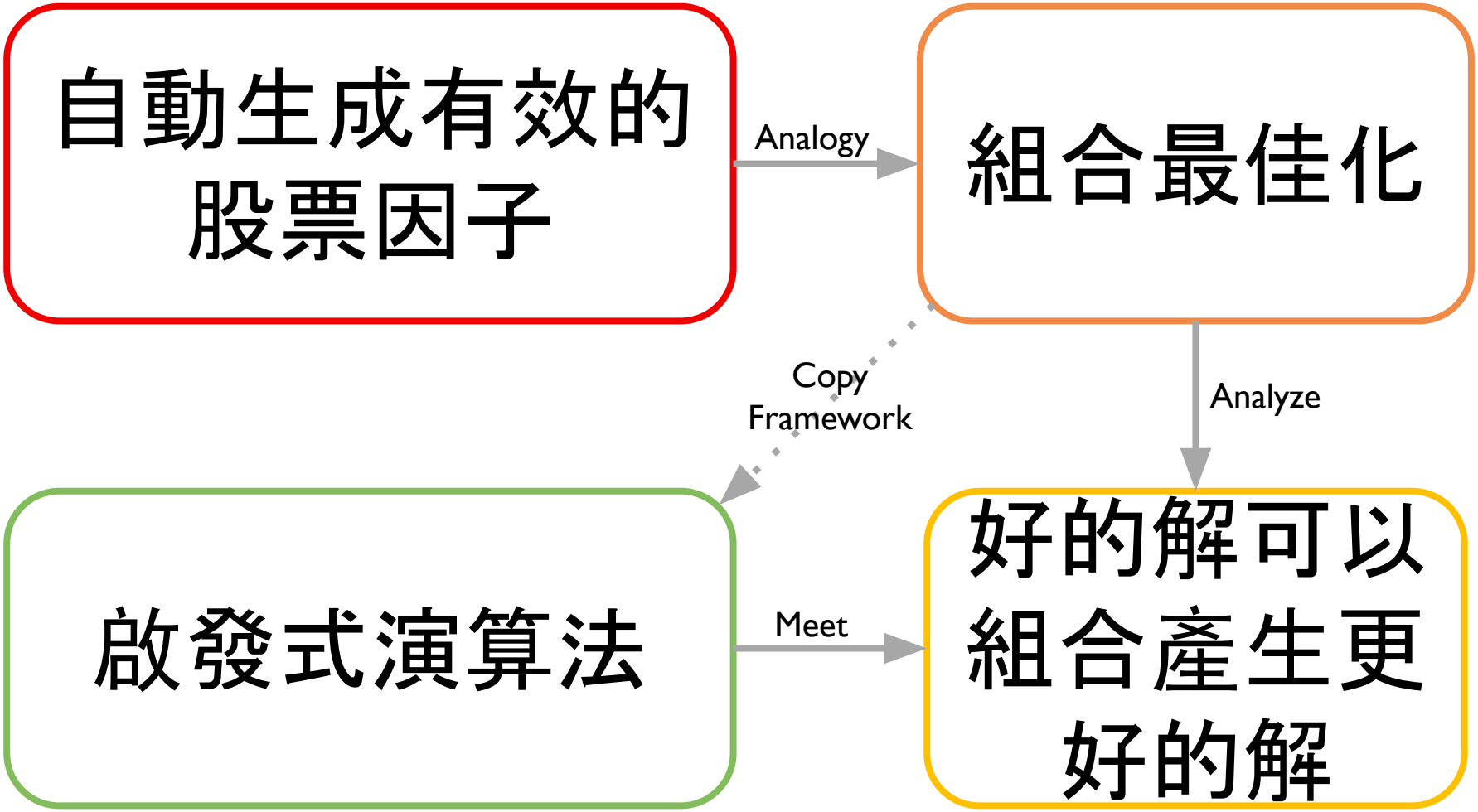
- Metaheuristic
- Unsupervised Learning
- Supervised Learning
- Reinforcement Learning

- Theorem
- Application
- Belief

結合方式



結合例子



常見的交易問題

- **組合問題**
  - 可能的組合有數千兆種，在給定衡量某個組合好壞的方式下，如何快速搜尋出好的組合
  - 例子：自動生成因子、資產池最佳化
- **預測問題**
  - 預測未來特定時點的目標值
  - 例子：預測股價、漲跌、報酬率排名.....
- **分群問題**
  - 給定各樣本的特徵，如何自動依據特徵將樣本分成有意義的類別
  - 例子：給定多個資產的報酬序列，自動將股票分一類、債券一類.....
- **結構化問題**
  - 給定非結構化資料，如何結構化為易於策略、模型使用的資料
  - 例子：財金新聞結構化、產業資訊結構化
- **生成問題**
  - 給定樣本，如何建立一個模型，生成與樣本有相似統計特徵的資料
  - 例子：市場資料合成（Market Data Synthesis），以生成模型替代蒙地卡羅與自定義的微分方程

組合問題範例、入門與實作

組合問題範例

- **目標：自動生成效果好的股票因子**
  - 股票因子：影響未來報酬的變量
  - 效果好：與未來報酬顯著正/負相關的變量
  - 例子：月營收年成長率排名越高，個股未來報酬排名越高
  - 難點：有太多種可能的組合，無法以遍歷的方式慢慢搜尋
    - Data：價量資料、財務資料、籌碼資料 .....
    - Operator：時間序列運算、橫截面運算、中性化 .....
    - Operator Parameter Sets：{1, 2, 4, 8, …}、{5, 10, 15, 20, …}、{20, 60, 120, 240, …} .....
- **方法：借鑑用於解組合最佳化問題的演算法**
  - 基因演算法、退火演算法、交叉熵方法 .....
  - 留下好的組合，剃除壞的組合，雜湊好的組合來生成出更好的組合
- **組合評分：因子投資文獻中常見的檢測**
  - Information Ratio、Fama–MacBeth Regression、Long-Short Sharpe.....

組合問題入門

- **適合對象**
  - 未來想進投信、做股權對沖的避險基金的同學
  - 在意交易策略經濟意涵與背後邏輯的同學
- **背景知識**
  - 因子投資
    - Your Complete Guide to Factor-Based Investing: The Way Smart Money Invests Today
    - Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Selecting Superior Returns and Controlling Risk
    - 因子投資 : 方法與實踐
  - 啟發式演算法
    - 基因演算法 Genetic Algorithm
- **學術論文**
  - Zura Kakushadze. (2016). 101 Formulaic Alphas.
  - 以基因演算法雜湊10種價量資料與30種Operator（> 2^40 ~= 1兆種組合），生成101個有效的價量因子

組合問題實作

- **歷史資料集**（2 ~ 4小時）
  - 台股：TEJ資料庫、TQuant Lab、FinMind、FinLab .....
  - 美股：Compustat、Eikon、Bloomberg、Reuters、FinLab .....
- **表達式回測**（6 ~ 8小時）
  - 以表達式（Data、Operator、Parameter）定義因子
    - 例子：營收成長率 → PercentChange(Revenue, 12 Months)
  - 輸入表達式，輸出因子評分
    - 例子：以Information Ratio作為衡量因子方式
      - 解析表達式 → 每期因子值與對應未來報酬
        → 計算（橫截面相關性平均/橫截面相關性標準差）
  - 程式實作技巧
    - 表達式以字串表示，利用Python內建函數eval或Stack資料結構來解析表達式
      - ChatGPT關鍵字：Parse Expression
    - Data → DataFrame、Operator → Function、
      Parameter → Int/Float

組合問題實作

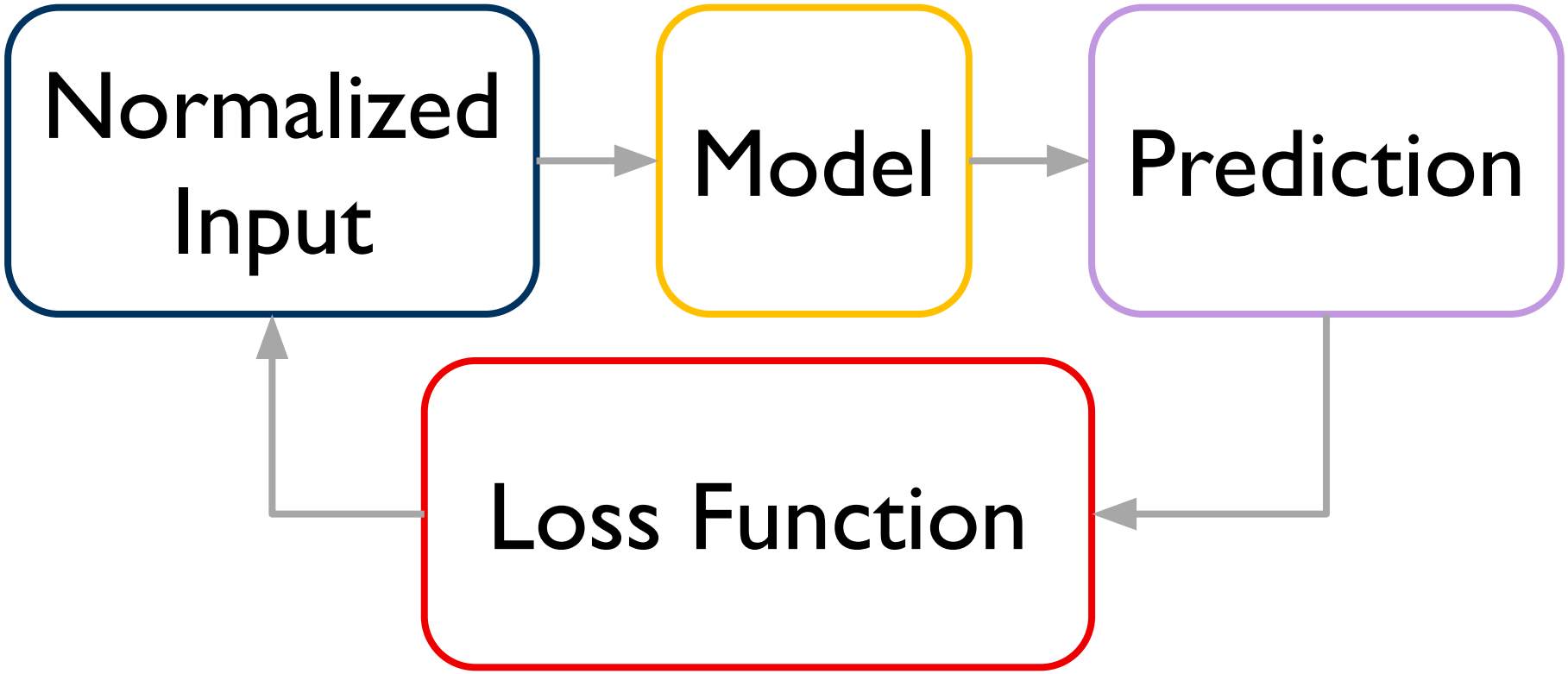
- **以基因演算法搜尋好的表達式組合** （10 ~ 12小時）
  - 生成多個表達數組合 → 衡量每個組合的好壞 → 留下好的表達式
  - 程式實作技巧
    - 使用DEAP這個Python套件實作
      - 表達式分為強表達式與弱表達式，強表達式對於Operator的輸入型別可以加以限制
      - DEAP允許在強表達式上使用基因演算法，常見的基因演算套件GPlearn則不支援
      - ChatGPT關鍵字：DEAP +Genetic Algorithm + Generate Expression
- **可以深入研究的問題**
  - 如何時維持回測速度的同時，增加回測精細度、多樣性
  - 如何設計組合評分表準，來得到一組分散的因子
  - 如何與LLM結合，對搜尋出的好因子，自動賦予經濟意涵

因子自動生成 + LLM範例

Credit: 22th TMBA 陳育理

預測問題要點、入門與實作

預測問題架構



模型輸入要點

- **數值資料標準化**
  - 數值資料例子：價量資料、財務資料 .....
  - 難點：多數的ML Application中，都假設樣本點之間為IID，所以其中常用的標準化方法（Z-score, Min-max .....）不能直接使用
    - 對於非IID的時序資料，如何更好地標準化，使得邏輯上可用、數學上合理、訓練上穩健是預測問題中非常重要的一環
      - 例子：以移動窗格方式標準化，在邏輯上可用，但是在訓練上並不穩健
    - 對於財金時序資料我們很難將樣本點標準化為Identically Distributed，但是可以做到Similar Standard Deviation
- **非數值資料標準化**
  - 非數值資料例子：新聞資料、產業分類 .....
  - 方法：將其轉換成向量或數值
    - 例子：Token Embedding、One Hot Encoding、ChatGPT .....
  - 難點：多數非數值資料無有用資訊
    - 例子：許多新聞都喜歡標籤台積電，即使內文相關度低
    - SNR過低會使模型模型難以學習有效特徵
    - 非數值資料/資料源的清理非常重要

模型輸入要點

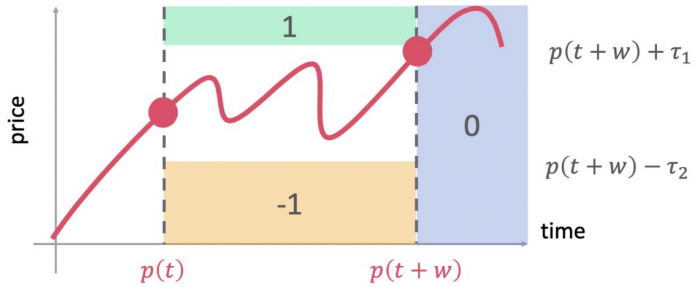
- **訓練樣本數**
  - 難點：日頻率資料對於百萬以上參數的神經網路而已，易過度擬合
    - Curse of Dimensionality：當模型的參數數量增加時，需要的訓練數據量通常也會呈指數增長
    - 即時加入橫截面資料，也會因共線性的問題，使有效樣本不足
  - 解決方法
    - 相信自己是天選之人不會過度擬合
    - 特徵選擇、損失函數中增入正規化項、Dropout
    - 模型簡化、集成學習（Ensemble Learning）
    - 交叉驗證（Cross-validation）（△）
    - 使用日內資料
      - 跳空資料需額外處理，否則模型易對跳空資料擬合
- **跨週期資料**
  - 難點：每種資料的刷新頻率不同，因此在每個時刻，每種資料的剩餘資訊比例差異大
    - 例子：月營收公布當下新資訊最多，離公布時間越遠資訊越少
  - 應對輸入設計Information Decay，幫助模型更好地處理過時信息來提高泛化能力和穩定性

模型本身要點

- **模型選擇**
  - 多數人使用在NLP領域有良好表現的時序模型：RNN（1980）、LSTM（1997）、GRU（2014）、TCN（2016）、Transformer（2017）……
  - 模型無用論：不同輸入對於訓練結果影響顯著，但是使用哪一種模型影響甚微
  - 參數不效率：動輒上百萬、千萬參數得到的策略效果，與幾個參數的簡單策略效果差異不大 (Why ML?)
- **模型汲取資訊的設計**
  - 模型汲取資訊的方式要合理，不然容易過度擬合
  - 不合理例子：以多個資產的報酬序列為輸入，以2D CNN汲取特徵
    - 2D CNN在影像辨識應用上效果顯著，是因為鄰近的像素相對於距離遠的像素對於物件的辨識有更多的資訊
    - 在這個例子中，報酬是否鄰近是因為資料預處理所影響，與要預測的目標無關

模型輸出要點(預測目標)

- **預測標的走勢**
  - 類比回歸問題
    - 價格、價格漲跌比例:因真值樣本的非平穩性，訓練效果極差
    - Position Sizing:預測最佳持有權重(-1 ~ 1)
      - Soft-output → 有助於減少過擬合，特別是在噪聲較多的情況下
  - 類比分類問題
    - 價格漲或跌:二元分類問題，因為漲或跌兩種類別之間的邊界過小，訓練效果不穩定
      - 在一般影像辨識應用中深度學習表現非常好，一個大前提即是兩種分類的邊界明顯(例如:貓跟狗)
    - Triple-Barrier Labeling (2018):三元分類問題，解決漲或跌兩種類別之間邊界過小的問題



模型輸出要點(預測目標)

- **預測資產池相對表現**
  - 資產報酬率排名
    - 類比於推薦系統，預測下個時間點資產報酬率排名
    - 模型即因子
  - 最佳資產權重
    - 預測下個時間點各個資產的最佳權重
    - Soft-output → 有助於減少過擬合，特別是在噪聲較多的情況下
- **預測極端行情**
  - 例子：漲跌停、黑天鵝
  - 類比於異常值偵測(Anomaly Detection)

損失函數要點(最佳化目標)

- **預測標的走勢**
  - 價格、價格漲跌比例：MSE、MAE、MAPE(平均絕對百分比誤差)
  - Position Sizing：Sharpe Ratio
  - 價格漲或跌、Triple-Barrier Labeling：交叉熵(Cross-entropy)
- **預測資產池相對表現**
  - 資產報酬率排名：排名損失(Ranking Loss)
  - 最佳資產權重：Sharpe Ratio
- **預測極端行情**
  - 成本敏感交叉熵(Cost Sensitive Cross-entropy)
- Batch Optimization
  - 問題：因為財金資料Non-IID的特性，使用Batch Optimization會使得模型容易對某一段的歷史資料過度擬合，降低模型的泛用性

預測問題範例(預測標的走勢)

- 目標:每分鐘預測下一分鐘 BTC的最佳持有權重
- 輸入:標準化分鐘報酬序列(過去 n分鐘)
- 輸出:下一分鐘的持有權重
- 模型: Transformer
- 最佳化目標: Sharpe Ratio
- 回測方式: Walk Forward Optimization

預測問題入門(預測標的走勢)

- **適合對象**
  - 喜歡交易加密貨幣的同學
  - 需要呼嚨啥也不會的主管/教授的同學
- **背景知識**
  - 趨勢交易
    - Time Series Forecasting、Position Sizing
    - Walk Forward Optimization
  - 深度學習(Deep Learning)
    - Gradient Descent
    - Transformer、Dropout、Hyperbolic Tangent
    - Pytorch、Numpy
- **學術論文**
  - Bryan Lim, Stefan Zohren, & Stephen Roberts. (2020). Enhancing Time Series Momentum Strategies Using Deep Neural Networks.
  - Zhang, Z., Zohren, S., & Roberts, S. (2020). Deep Learning for Portfolio Optimization. The Journal of Financial Data Science, 2(4), 8–20.

預測問題實作(預測標的走勢)

- **歷史資料集**(1 ~ 2小時)
  - BTC永續合約:幣安API、幣安官網
- **整理歷史資料**(2 ~ 3小時)
  - 將歷史資料整理成Input-Output Pairs
    - Input I_t: $[r^N_{t-n+1}, r^N_{t-n+1}, \ldots, r^N_t]$, r^N_i = 標準化的該分鐘報酬
    - Output O_t: $[r_{t+1}]$, r_i = 該分鐘報酬
- **損失函數**(2 ~ 3小時)
  - Prediction P_t: 在t時間點預測的最佳持有權重
  - Loss = -Mean($[O_i*P_i]$) / Std($[O_i*P_i]$), i = 1, 2, 3, … T
- **模型**(1 ~ 2小時)
  - (Transformer Encoder)*K
    + (Fully Connected Layer + Hyperbolic Tangent)*Q
- **Training Loop**(1 ~ 2小時)
  - Full Sample Optimization
- **Walk Forward Optimization**(1 ~ 2小時)
  - Training Periods:用於更新模型權重
  - Validation Periods:用於選擇超參數
  - Testing Periods:用於評估模型性能和泛化能力

學習地圖

交易策略學習地圖

- 因子投資
  - Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Selecting Superior Returns and Controlling Risk
  - Fama–French Three-factor model, Fama–MacBeth Regression
  - Multivariate Regression, Statistical Inference
- 資產配置
  - Modern Portfolio Theory, Black-Litterman Model, Risk Parity
  - Linear Algebra, Convex Optimization
- 統計套利
  - Pairs Trading: Quantitative Methods and Analysis
  - Arbitrage Pricing Theory, Cointegration
  - Time Series Analysis, Stochastic Process
- 事件交易
  - EventStudyTools
  - Abnormal Returns, Wilcoxon Signed-rank Test, Permutation Test
- 趨勢交易
  - Time Series Forecasting, Position Sizing

ML學習地圖

- **課程**
  - 李弘毅教授Youtube頻道
- **書籍**
  - Advances in Financial Machine Learning (2018)
  - Machine Learning for Asset Managers (2020)
- **期刊、會議**
  - Journal of Financial Data Science
  - Conference on Neural Information Processing Systems
  - International Conference on Learning Representations
  - International Conference on Machine Learning
- **程式資源**
  - mlfinlab（Implementations regarding "Advances in Financial Machine Learning"）
  - Qlib（An AI-oriented Quantitative Investment Platform by Microsoft）

感謝聆聽，敬請指教