Machine Learning HW4

MLTAs
ntueemlta2024@gmail.com

Links

- Code template
- Kaggle
- Report template
- Math problems

Outline

- HW4 Intro - Prediction of Postpartum Depression (PPD)
  - Tasks Description
  - Training/Testing Data
  - Criteria: ROC and AUC
- Prerequisites
  - Ensemble (optional)
  - Surrogate Loss
  - Analyzing Feature Importance
    - Ablation study
    - Saliency Map
- Regulations & Grading

Prediction of Postpartum Depression (PPD)

Task Description

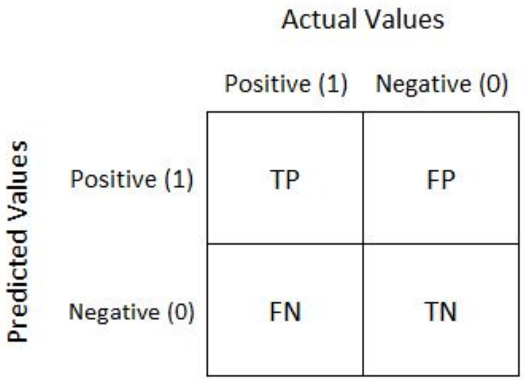
- 產後憂鬱症預測 (binary classification)
- 樣本:同時期母嬰組, 共21202人
- 疾病:產後6個月 憂鬱症(17.2%)
- 預測因子(共103個):
  - 嬰幼兒:出生方式、第一(幾)胎、季節、出生體重、早 產、健康狀況、母乳…
  - 母親:生產年紀、教育程度、壓力、安胎、抽菸、喝酒、體態 (BMI)…
  - 家庭:婚姻關係、經濟收入、家庭支持、非預期懷孕 …
  - 環境:居住區(都市鄉鎮)、空污濃度(PM2.5, CO, 氮氧化物, 臭氧等)、溫濕度、綠地…
  - 詳細表格

Training/Testing Data

- Total :
  - 21202 samples
- Training and validation splits :
  - 16962 samples (80%)
  - You can determine the ratio for the two sets
- Testing splits :
  - Public leaderboard: 2120 samples (10%)
  - Private leaderboard: 2120 samples (10%)

Criteria: ROC and AUC

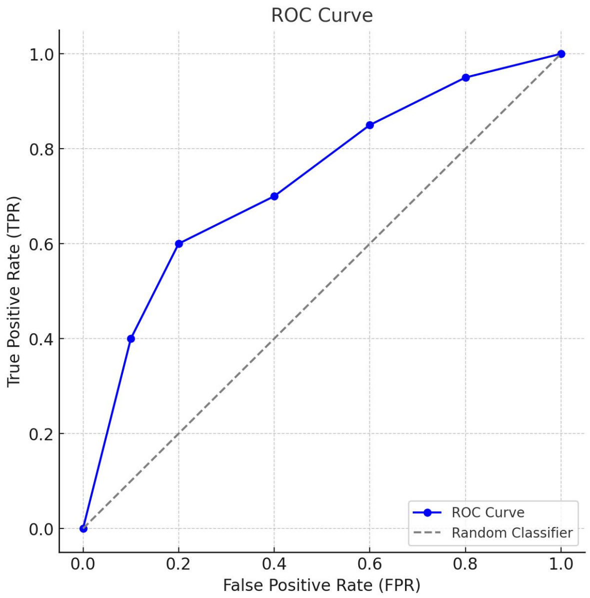
1. Confusion Matrix (Foundation for ROC)
    - True Positive (TP)
    - False Positive (FP)
    - True Negative (TN)
    - False Negative (FN)



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

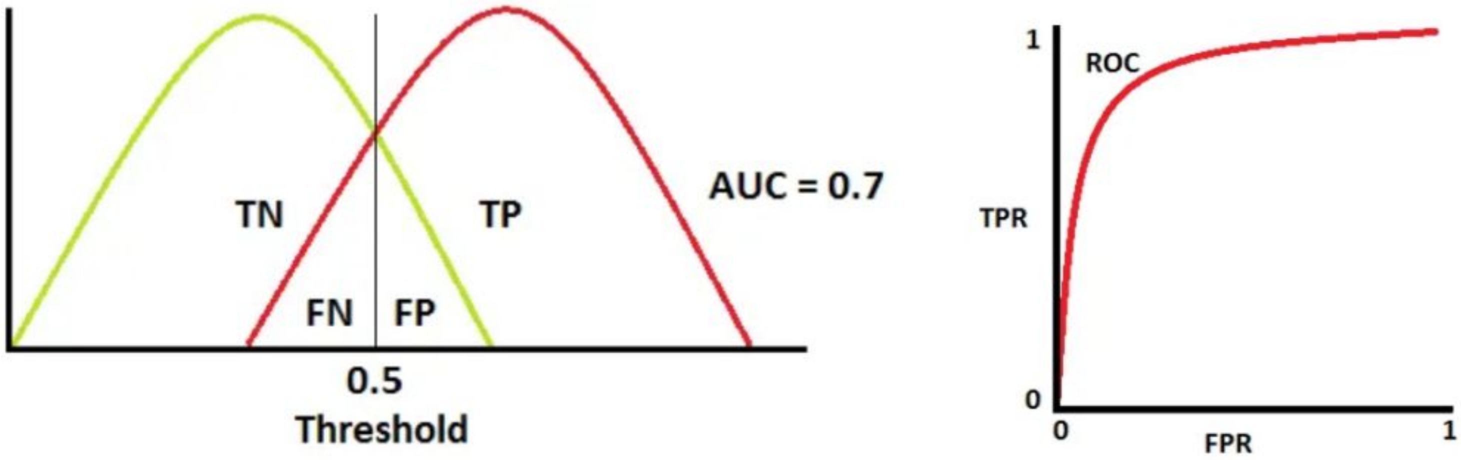
Criteria: ROC and AUC

2. ROC Curve
   - The ROC curve is a graphical plot that shows the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) for different threshold values.
   - TPR = TP / (TP+FN)
     - 又稱recall，代表陽性中被發現的比例
   - FPR = FP / (FP+TN)
     - 代表測出的陽性中，屬於偽陽性的比例
3. AUC (Area Under the Curve)
   - The AUC represents the area under the ROC curve.
   - Ranges from 0 to 1
     - = 1 : Perfect classifier
     - = 0.5 : Random classifier
     - < 0.5 : Worse than random



Criteria: ROC and AUC

4. ROC and AUC



Prerequisites

Ensemble

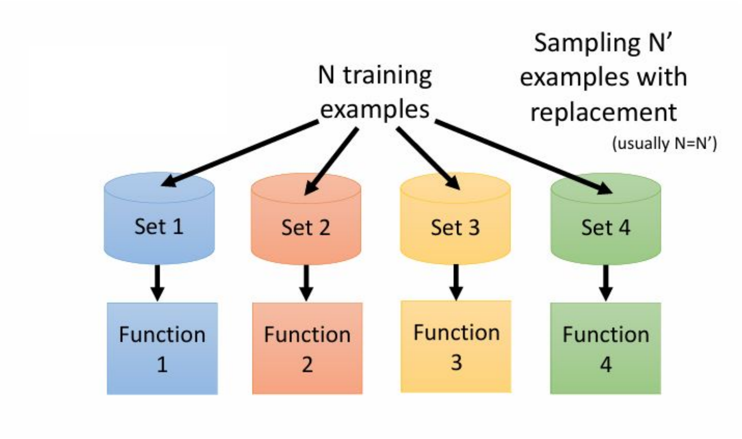
- Ensemble techniques: **combine multiple models** to improve performance, accuracy, and robustness over a single model. It reduce variance and handle noise by aggregating the predictions of several base models. Some popular methods:
  - Bagging
  - Stacking
  - Boosting
  - …

- In this homework, we will use bagging to deal with imbalanced data (17.2% positive). You are also encouraged to use other ensemble techniques.
- More information about ensemble can be found here.

Ensemble

- Bagging

Training



N training examples

Sampling N' examples with replacement

(usually N=N')

Set 1

Set 2

Set 3

Set 4

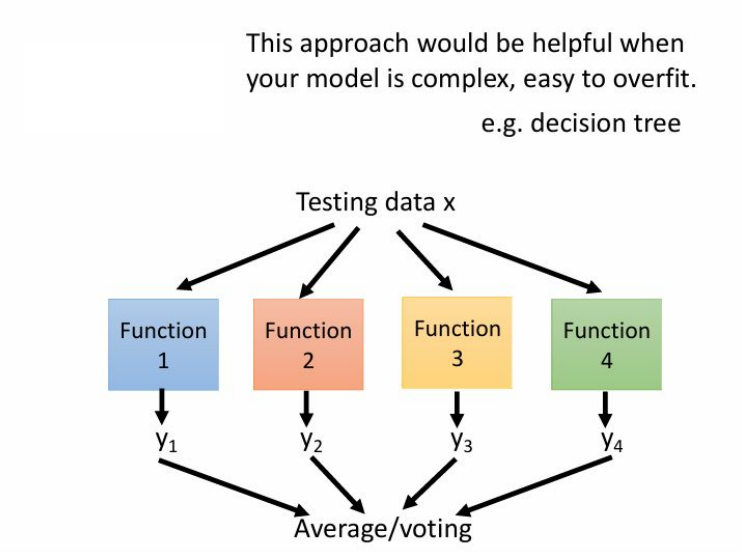
Function 1

Function 2

Function 3

Function 4

Testing



This approach would be helpful when your model is complex, easy to overfit.

e.g. decision tree

Testing data x

Function 1

Function 2

Function 3

Function 4

y_1

y_2

y_3

y_4

Average/voting

- (optional) Bagging variants for imbalanced data:
    - include all positive samples, randam sample comparable amount of negative samples
    - other imbalanced classification

Surrogate Loss

- Optimizing Cross Entropy Loss vs. Optimizing AUC
  - AUC = P($f(x^+) > f(x^-)$)
    - Why? Prove it in the report.
  - Gradient-based optimization vs. Rank-based optimization
- Surrogate loss imitates the AUC objective in a differentiable way, making it easier to optimize with standard gradient-based methods. Examples:
  - Pairwise Hinge Loss

    $$\mathcal{L}_{\text{hinge}} = \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} \max(0, 1 - (f(x_i^+) - f(x_j^-)))$$

  - Squared Hinge Loss

    $$\mathcal{L}_{\text{squared hinge}} = \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} \max(0, 1 - (f(x_i^+) - f(x_j^-)))^2$$

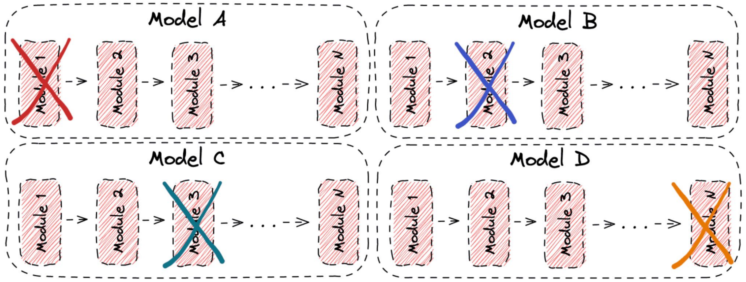
  - Exponential Loss

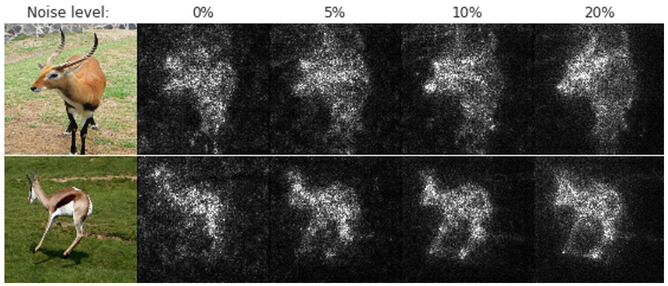
    $$\mathcal{L}_{\text{exp}} = \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} \exp(-(f(x_i^+) - f(x_j^-)))$$

  - More

Analyzing Feature Importance

- There are several ways to analyze and understand the importance of each input feature. In this homework, we perform the following two techniques:
    - **Ablation study**: An ablation study is a method used in machine learning to assess the contribution of specific components by removing or altering them and measuring the impact on performance.
    - **Saliency map**: A saliency map is a visualization technique that uses the gradient of the model's output with respect to the input to highlight which input features most strongly influence the model's prediction.





- More information can be found here (Explainable AI).

Regulations & Grading

Kaggle - Info

- Kaggle 連結 : https://www.kaggle.com/competitions/ml-2024-fall-hw-4/overview
- 個人進行，不需組隊
- 隊名:
  - 修課學生 : **學號(底線)任意名稱 (e.g., b09901105_謝博揚喜洋洋)**
  - 旁聽 : 旁聽_任意名稱（旁聽請於期限過後再上傳）
- 每天上傳上限 5 次
- 在Kaggle Deadline前可以選擇2份submission作為private score的評分依據。如果未勾選，系統會自動選擇Public Leaderboard中表現最佳的兩次。
- Bonus（Optional）- 1%
  - 修課生 private leaderboard 排名前五名可繳交。
  - 繳交投影片描述實作方法，另外需錄製一份講解影片（少於三分鐘）作一個簡單的presentation，助教將公布給同學們參考。

Regulation

- 可以使用任何套件
- 可以使用額外資料

Grading Policy - Deadline

- Kaggle Deadline: 2024/11/22 23:59:59 (GMT+8)
- Cool Deadline: 2024/11/22 23:59:59 (GMT+8)

Grading Criteria

- Kaggle - 3%
    - 超過public leaderboard simple baseline分數： **0.5%**
    - 超過private leaderboard simple baseline分數： **0.5%**
    - **排名分數:**
        - score = 1 - rank*0.01
        - public leaderboard的排名分數： **1%**
        - private leaderboard的排名分數 : **1%**
    - code template
- Programming report - 3%
    - report template
- Math problem - 6%
    - math problem
    - 若有和其他修課同學討論，請務必於題號前標明 collaborator（含姓名、學號）

Cool Submissions

在Cool上分別繳交以下檔案:

1. **report.pdf**
2. **math.pdf**
3. **code.ipynb**

Grading Policy - Others

- Lateness
  - Cool 遲交每小時分數*0.95，兩天後歸0
  - 有特殊原因請找助教
- Runtime Error
  - 當程式錯誤，造成助教無法順利執行，請在公告時間內寄信向助教說明，修好之後重新執行所得kaggle部分分數將x0.5。

學術倫理

- Cheating
  - 抄code、抄report（含之前修課同學）
  - 開設kaggle多重分身帳號註冊competition
  - 教授與助教群保留請同學到辦公室解釋coding作業的權利

References

1. https://miro.medium.com/v2/resize:fit:712/1*Z54JgbS4DUwWSknhDCvNTQ.png
2. https://www.baeldung.com/wp-content/uploads/sites/4/2022/06/ablation_study.png
3. https://arxiv.org/pdf/1706.03825
4. https://complex-systems-ai.com/en/learning-supervises/auc-and-rock/
5. https://ar5iv.labs.arxiv.org/html/2203.15046
6. https://arxiv.org/pdf/1608.06048