Machine Learning HW5

MLTAs
ntueemlta2023@gmail.com

Outline

- HW5 - Income 50K prediction
  - Dataset and Tasks Description
  - Provided Feature Format
  - Sample Submission
- Kaggle
- Grading / Assignment Regulation

Dataset and task introduction

- Dataset : Adult Data Set

- Task : **Binary Classification**
  - **SVM, kernel SVM**

  Determine whether a person makes over 50K a year.

Data Attribute Information

train.csv 、test.csv :

age, workclass, fnlwgt, education, education num, marital-status, occupation
relationship, race, sex, capital-gain, capital-loss, hours-per-week,
native-country, make over 50K a year or not

```
1 39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
2 50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
3 38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
4 53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K
5 28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
6 37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K
7 49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, <=50K
8 52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 45, United-States, >50K
```

Provided Feature Format

X_train, Y_train, X_test

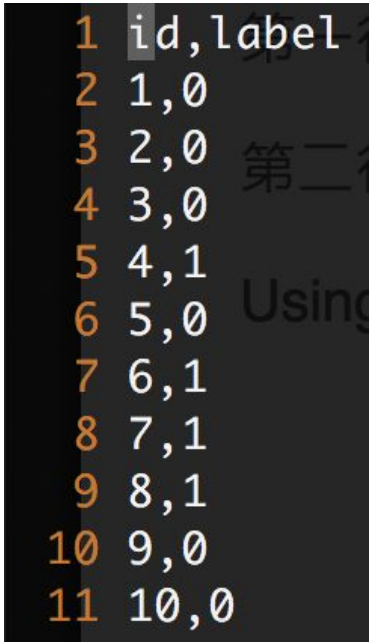
1. discrete features in train.csv => one-hot encoding in X_train (work_class,education...)
2. continuous features in train.csv => remain the same in X_train (age,capital_gain...)
3. X_train, X_test : each row contains one 106-dim feature represents a sample
4. Y_train: label = 0 means “<= 50K” 、label = 1 means “ >50K ”

```
age,fnlwgt,sex,capital_gain,capital_loss,hours_per_week, Federal-gov, Local-gov, Never-worked, Private, Self-emp-inc, Self-emp-not-inc, State-gov, Without-pay,?_workclass, 10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Assoc-acdm, Assoc-voc, Bachelors, Doctorate, HS-grad, Masters, Preschool, Prof-school, Some-college, Divorced, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-married, Separated, Widowed, Adm-clerical, Armed-Forces, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Other-service, Priv-house-serv, Prof-specialty, Protective-serv, Sales, Tech-support, Transport-moving,?_occupation, Husband, Not-in-family, Other-relative, Own-child, Unmarried, Wife, Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, White, Cambodia, Canada, China, Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, England, France, Germany, Greece, Guatemala, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Mexico, Nicaragua, Outlying-US(Guam-USVI-etc), Peru, Philippines, Poland, Portugal, Puerto-Rico, Scotland, South, Taiwan, Thailand, Trinadad&Tobago, United-States, Vietnam, Yugoslavia,?_native_country
25,226802,1,0,0,40,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
```

Sample Submission

請預測test set中16281筆資料

1. 上傳格式為csv
2. 第一行必須為id, label，第二行開始為預測結果
3. 每行分別為id以及預測的label，請以逗號分隔
4. Evaluation: Accuracy



Kaggle Info & Deadline

- Link: https://www.kaggle.com/competitions/ml2023-fall-hw5
- 個人進行、不須組隊
- Team Name:
  - 修課學生: **學號_任意名稱（e.g., b09901666_）**
  - 旁聽: 旁聽_任意名稱
- Maximum Daily Submission: 5 times
- Kaggle Deadline: 12/16/2023 01:30:00 (GMT+8)
- Cool Deadline: 12/17/2023 23:59:59 (GMT+8)
- test set的16281筆資料將被分為兩份，8140筆public，8141筆private
- Leaderboard上所顯示為public score，在Kaggle Deadline前可以選擇2份submission作為private score的評分依據。

配分 Grading Criteria - kaggle (2% + Bonus 1%)

- Kaggle Score Point - 2%
  - 以 12/9/2022 23:59:59 於 **public/private scoreboard** 之分數為準：
    - 超過public leaderboard的simple baseline分數： **0.5%**
    - 超過public leaderboard的strong baseline分數： **0.5%**
    - 超過private leaderboard的simple baseline分數 : **0.5%**
    - 超過private leaderboard的strong baseline分數 : **0.5%**
  - 以上皆須通過 Reproduce 才給分
- Bonus - 1%
  - (1.0%) private leaderboard 排名前五名，並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

配分 Grading Criteria - report(12%)

- Programming Report - 4%
  - https://docs.google.com/document/d/1z0visPuuFZRrBGMvJvm59v9BLOTWmY3QjlU-cLqjl0M/edit?usp=sharing
- Math Problem - 6(+2)%
  - Type in latex(preferable) or take pictures of your handwriting
    - https://ntueemlta2023.github.io/homeworks/hw5/ml-2023fall-hw5-math.pdf
- Write them in report.pdf

作業提示 (Hint)

1. 提供兩份程式檔，一份是自己設計kernel或是使用DNN做feature transformation，另一份是直接使用sklearn套件
   - sample code: https://drive.google.com/file/d/1j4OJxz5xrvhR6cj3IUqCP-EPZ9u3L_cR/view?usp=sharing https://drive.google.com/file/d/1Sxm-HP371DWjDnFOIcVzoXvDDxBG7gBj/view?usp=sharing
   -
2. 只能使用SVM-based方式，不能使用其他類型的分類器
3. 不要思考太複雜的模型，可以自己先做一些feature selection，還有思考normalize的方法
4. 目前使用的是torch跟sklearn，如果有要使用其他模型相關的套件可來信詢問

Cool Submissions

你的cool上至少有下列4個檔案:

1. **handcraft.ipynb** : handcraft kernel or dnn feature transformation

2. **SVM.ipynb** : sklearn SVM approach

3. **model.pth/model.txt** : the model ckpt meets the highest score you choose in kaggle (if you use sklearn, submit a model.txt describing the parameters you use)

4. **report.pdf** : Please refer to report template

將以上四個檔案放入一個資料夾，命名為"學號_hw5”

請不要上傳dataset，請不要上傳dataset，請不要上傳dataset

繳交格式 Handin Format

- Kaggle deadline：12/16/2023 01:30:00 (GMT+8)

  Cool code & report deadline：12/17/2023 23:59:59 (GMT+8)

- 把程式碼和report壓縮成zip檔上傳到cool，檔案名稱為，學號_hw5.zip，包含程式碼及report.pdf(report包含數學題)，你的檔案解壓縮以後應該是以下格式，任何格式錯誤將會扣0.5分

  學號_hw5 (資料夾)

  - report.pdf
  - handcraft.ipynb
  - SVM.ipynb
  - model.pth / model.txt

其他規定 Other Policy

- Lateness
  - Cool 遲交
    - 以最後一次繳交之時間為準
    - 一天: 以小時為單位，線性遞減至七折
    - 兩天: 以小時為單位，從七折線性遞減零分
  - 有特殊原因，請找助教 說明。
  - 不接受程式 or 報告單獨遲交
- 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。
  - 可以更改的部分僅限 syntax 及 io 的部分，不得改程式邏輯或是演算法，至於其他部分由助教認定為主。

Requirements - environment issue

- environment.yaml
- **若需要其它套件，請及早來信詢問。**
  - **套件版本與python版本並沒有強制限制，以colab能跑為主**

    ```
    $pip list
    $python –version
    $nvidia-smi
    ```

    如果助教跑你的code跑不動會寄信與同學確認
  - **強烈建議不要在WSL上嘗試裝設nvidia-driver**

    在Windows/Ubuntu/MacOS環境下直接跑反而會更加順利

    繳交時記得注意一下資料夾結構跟檔名，推薦在.ipynb檔頭用註解寫上跑的系統

    ```
    #@system[Ubuntu-22.04 LTS/Windows10/CentOS/MacOS/Colab….]
    ```

Requirements - file uploading

- 你的上繳至 cool 中的檔案請壓縮在同一個資料夾，並取名為**<學號>_hw5.zip**
  - 請將參數連結(最佳model，或其他reproduce必須的檔案)附在report中
    - **也可以上傳自己的雲端，在 code內用 gdown 指令。**
    - **範例：**

```
import gdown

url = <你的 model 壓縮檔 Google雲端連結>
output = "model.zip"
gdown.download(url=url, output=output, quiet=False, fuzzy=True)
!unzip -o model.zip
```

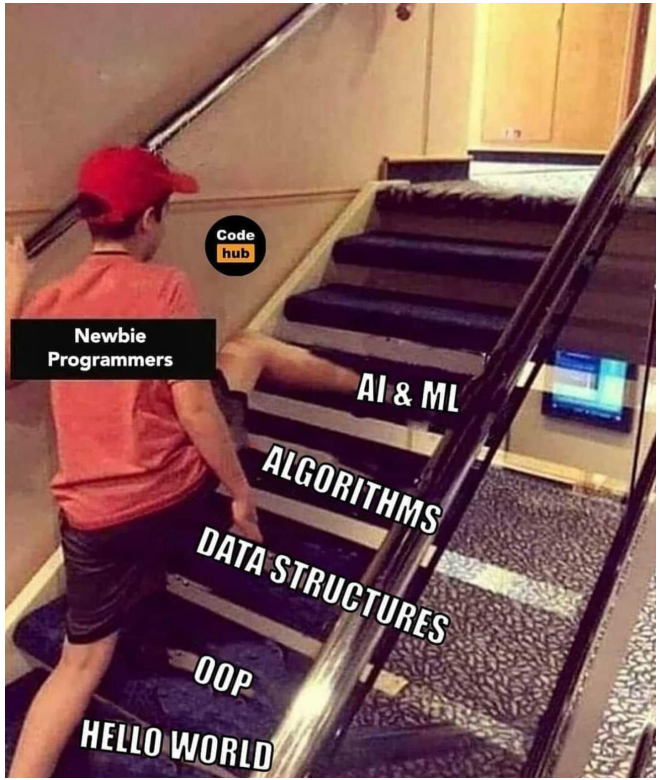
  - Optional: 其他可以幫助說明你的 code 的文件
  - 簡述一下使用到的套件名稱以及套件版本(ex:python3.10 & numpy(a.b.c).....)

FAQ

- **環境問題請善用 google。**
  - **pip install xxx**
  - **apt-get install xxx**

- 有問題建議可以在 FB Group 裡面留言發問，可能很多人都有一樣的問題。

- 若有其他問題，請寄信至助教信箱，**請勿直接FB私訊助教**。
  - Mail title:
    [ML23_hw5_code]{your name}_{title}
    [ML23_hw5_math]{your name}_{title}



TA

- ntueemlta2023@gmail.com or b08209023@ntu.edu.tw
- Title：[ML23_hw5_code]{your name}_{title}
- 關於環境/運行時間/註解問題:

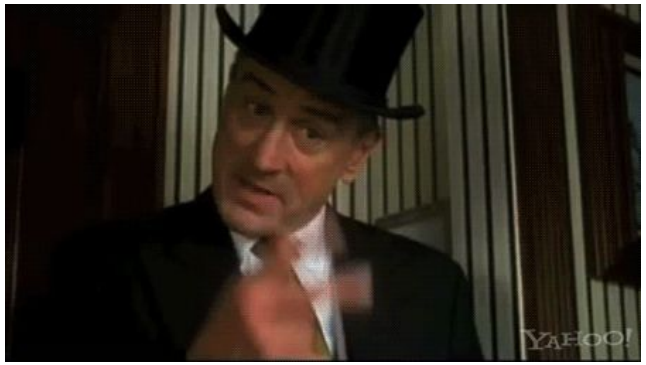
  1. 環境部分基本上以colab可以順利執行為主，推薦使用 python3.7，但python3.10也可以，倘若有套件衝突需要檢 查自己電腦環境有沒有爛掉，強烈不建議用WSL跑，如果都是用助教的範例code為主並且在colab上可以順跑，沒有奇怪的 import的話則不需要特別註解跟提交。
  2. 時間限制非強制，但還是推薦壓在 10min左右可以跑10個epoch，基本上助教檢驗code時不要讓助教的電腦跑太久即可

  (助教顯卡約莫1650等級，不要拿RTX ada A6000跑個十分鐘或者拿工作站的大量GPU unit 去train)

- TA hour:

  週二早上1000~1200 @ IPCS 206，強烈推薦寄信先問，助教最近生病 .....
  跟助教另外約時間/google meet

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
  - 於訓練過程以任何不限定形式接觸到testing data的正確答案
  - 不得上傳之前的kaggle競賽
  - 教授與助教群保留請同學到辦公室解釋coding作業的權利，請同學務必自愛