Machine Learning HW2

MLTAs
ntueemlta2021@gmail.com

Outline

- HW2 - Income 50K prediction
  - Dataset and Tasks Description
  - Provided Feature Format
  - Sample Submission
- Kaggle
- Grading / Assignment Regulation

Dataset and task introduction

- Dataset : Adult Data Set

  Reference : https://archive.ics.uci.edu/ml/datasets/Adult

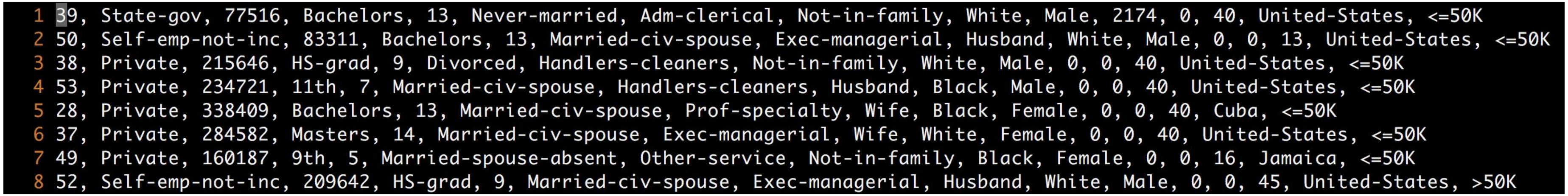
  Please down load data from here（只需要載X_train,Y_train, X_test就好）

- Task : **Binary Classification**
  - **Logistic regression, Probabilistic generative model**

  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



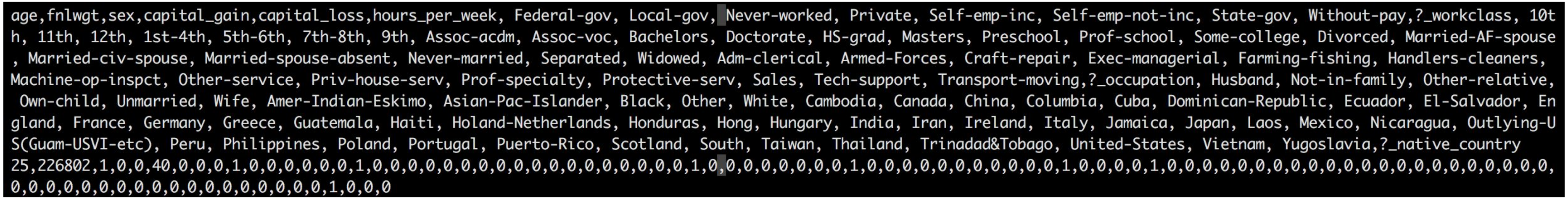
```
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
38, Private, 215646, HS-grad, 9, Divorced,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

- More detail please check out Kaggle Description Page

Provided Feature Format

X_train, Y_train, X_test ： (Please download data here)

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 ”

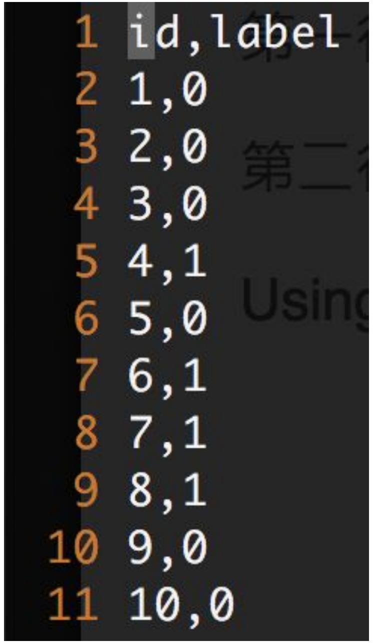


```
age,fnlwgt,sex,capital_gain,capital_loss,hours_per_week, Federal-gov, Local-gov, Never-worked, Private, Self-emp-inc, Self-emp-not-inc, State-gov, Without-pay,?_workclass, 10t
h, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Assoc-acdm, Assoc-voc, Bachelors, Doctorate, HS-grad, Masters, Preschool, 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, En
gland, France, Germany, Greece, Guatemala, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Mexico, Nicaragua, Outlying-U
S(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



```
id,label
1,0
2,0
3,0
4,1
5,0
6,1
7,1
8,1
9,0
10,0
```

Kaggle Info & Deadline

- Link: https://www.kaggle.com/t/93e214f8b5b64978a9e03c923dfd3e8f
- sample code
- 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: 10/28/2021 23:59:59 (GMT+8)
- Ceiba Deadline: 10/30/2021 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 (5% + Bonus 1%)

- Kaggle Deadline : 10/28/2021 23:59:59 (GMT+8)
- Kaggle Score Point - 4%
  - 以 10/28/2021 23:59:59 於 **public/private scoreboard** 之分數為準：
    - 超過public leaderboard的simple baseline分數： **1%**
    - 超過public leaderboard的strong baseline分數： **1%**
    - 超過private leaderboard的simple baseline分數 : **1%**
    - 超過private leaderboard的strong baseline分數 : **1%**
  - 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(5%)

- Programming Report - 3%
  - https://docs.google.com/document/d/1y_5H041452Qu5OtYcFEVK_yAcaFVmc_e6daUv5bLwzE/edit?usp=sharing
- Math Problem - 3%
  - https://hackmd.io/@GfOkB4kgS66YhhM7j6TJew/BJ-wGv8HY
  - Type in latex(preferable) or take pictures of your handwriting
- Write them in report.pdf

作業規定 Assignment Regulation

1. 請手刻 gradient descent 實作 logistic regression
2. 請手刻實作 probabilistic generative model
3. Only Python 3.7 available !
4. hw2_logistic.ipynb、hw2_generative.ipynb 開放使用套件
   a. numpy ==1.19.5
   b. scipy == 1.4.1
   c. pandas == 1.1.5
   d. python standard library
5. hw2_best.ipynb不限做法，開放以下套件（但有版本限制請注意）
   a. pytorch == 1.9.0 (phytorch教學一, pytorch教學二）
   b. tensorflow == 2.6.0
   c. keras == 2.6.0
   d. scikit-learn == 0.22.2
   e. 不可以使用 xgboost, AdaBoostClassifier, ExtraTreesClassifier
6. 若需使用其他套件，請儘早寄信至助教信箱詢問，並請闡明原因。

Ceiba Submissions

你的ceiba上至少有下列4個檔案(格式必須完全一樣):

1. **hw2_logistic.ipynb** : handcraft "logistic regression" using Gradient Descent

2. **hw2_generative.ipynb** : handcraft "probabilistic generative model"

3. **hw2_best.ipynb** : meet the highest score you choose in kaggle

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

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

Report 格式

- 限制
    - 檔名必須為 report.pdf !!!
    - 檔名必須為 report.pdf !!!
    - 檔名必須為 report.pdf !!!
    - 請用中文撰寫report(非中文母語者可用英文)
    - 請標明系級、學號、姓名，並按照report模板回答問題，切勿隨意更動題號順序
    - 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)
- Report模板連結
    - 連結：Link
- 截止日期同 Ceiba Deadline: 10/30/2021 23:59:59 (GMT+8)

其他規定 Other Policy

- Lateness
  - Ceiba 遲交一天(不足一天以一天計算) hw2 所得總分將x0.7
  - 不接受程式 or 報告單獨遲交
  - 不得遲交超過一天, 若有特殊原因請儘速聯絡助教

繳交格式 Handin Format

- Kaggle deadline：10/28/2021 23:59:59 (GMT+8)

  Ceiba code & report deadline：10/30/2021 23:59:59 (GMT+8)

- 把程式碼和report壓縮成zip檔上傳到ceiba，檔案名稱為，學號_hw2.zip，包含程式碼及report.pdf(report包含數學題）

其他規定 Other Policy

- 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

機器學習前測

前測問卷, 請大家幫忙填寫

TA Hour

- 10/22, 10/29 (Fri) @BL B1 系k
- 18:00 ~ 19:00