Machine Learning HW2

TAs
ntumlta2019@gmail.com

Outline

- Dataset and Task Introduction
- Provided Feature Format
- Requirements
- Kaggle
- Deadlines and Submissions
- FAQ
- Link

Dataset and Task Introduction

1. Task: **Binary Classification**

   Determine whether a person makes over 50K a year.

1. Dataset: **ADULT**

   Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)).

1. Reference:

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

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
```

- For more details please check out Kaggle Description Page

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, 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
25,226802,1,0,0,40,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
```

Submission Format

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

Requirements

1. 請手刻 gradient descent 實作 logistic regression
2. 請手刻實作 probabilistic generative model
3. hw2_logistic.sh、hw2_generative.sh、hw2_best.sh皆須在5分鐘內跑完
4. Only Python 3.6 available !!!!
5. hw2_logistic.sh、hw2_generative.sh 開放使用套件
   a. numpy >=1.14
   b. scipy == 1.2.1
   c. pandas >= 0.24.1
   d. python standard library
6. hw2_best.sh不限做法，開放以下套件（但有版本限制請注意）
   a. pytorch == 1.0.1
   b. tensorflow == 1.12.0
   c. keras == 2.2.4
   d. scikit-learn == 0.20.0
   e. 不可以使用 xgboost, AdaBoostClassifier, ExtraTreesClassifier
7. 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。

Kaggle

1. kaggle url：https://www.kaggle.com/c/ml2019spring-hw2
2. 請使用作業一時創建的kaggle帳號登入。
3. 個人進行，不需組隊。
4. 隊名:學號_任意名稱(ex. b02902000_日本一級棒)，旁聽同學請避免學號開頭。
5. 每日上傳上限5次。
6. test set的16281筆資料將被分為兩份，8140筆public，8141筆private。
7. 最後的計分排名將以2筆自行選擇的結果，測試在private set上的準確率為準。

★ kaggle名稱不符合規定者將不會得到任何kaggle上分數。

Deadlines

1. Kaggle deadline: 2019/3/23 11:59:59 (GMT+8) (revised !!)
2. Github code & report deadline: 2019/3/24 23:59:59 (GMT+8) (revised !!)
3. 助教會在deadline一到就clone所有程式, 並且**不再重新clone任何檔案**

Github Submissions

github上ML2019SPRING/hw2/裡面請至少包含:

1. report.pdf
2. hw2_logistic.sh
3. hw2_generative.sh
4. hw2_best.sh

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

Script Usage

bash ./hw2_logistic.sh $1 $2 $3 $4 $5 $6 output: your prediction

bash ./hw2_generative.sh $1 $2 $3 $4 $5 $6 output: your prediction

bash ./hw2_best.sh $1 $2 $3 $4 $5 $6 output: your prediction

$1: raw data (train.csv) $2: test data (test.csv)

$3: provided train feature (X_train.csv) $4: provided train label (Y_train.csv)

$5: provided test feature (X_test.csv) $6: prediction.csv

上述提供的input大家可以不用全部都使用
批改作業時會cd進同學的資料夾

Script Usage

助教執行範例：

bash ./hw2_logistic.sh /path/to/train.csv /path/to/test.csv /path/to/X_train /path/to/Y_train /path/to/X_test /path/to/prediction.csv

$N 表示第N個argument，助教在執行的時候會輸入相對路徑，**不可將路徑寫死**

Score - Kaggle Rank

- Kaggle Deadline : 03/23/2019 11:59:59 (GMT+8)
- Early Baseline Point - 1%
  - 在 03/14/2019 23:59:59 (GMT+8) 前於 **public scoreboard** 通過 **simple baseline** : **1%**
- Private Score Point - 4%
  - 以 03/23/2019 11:59:59 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%**
- Bonus - 1%
  - (1.0%) private leaderboard 排名前五名且於助教時間上台分享的同學

Score - Reproduce

- Github code & report deadline: 2019/3/24 23:59:59 (GMT+8)
- 除了直接以Kaggle上的資訊評分外, 助教也會clone大家github上的程式來檢查
  - 執行程式時test data順序會shuffle過, 請勿直接輸出事先存取的答案。
  - hw2_logistic.sh 或 hw2_generative.sh的結果, 有一份必須在test set上超過simple baseline, 才會有simple baseline的分數
  - **關於kaggle分數與reproduce結果的關係 , 請參考 Link**

Score - Report

report.pdf：PDF (限制：不能超過2頁、請使用template作答)

- (1%) 請比較你實作的generative model、logistic regression的準確率，何者較佳？
- (1%) 請說明你實作的best model，其訓練方式和準確率為何？
- (1%) 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。(有關normalization請參考：https://goo.gl/XBM3aE)
- (1%) 請實作logistic regression的正規化(regularization)，並討論其對於你的模型準確率的影響。(有關regularization請參考：https://goo.gl/SSWGhf P.35)
- (1%) 請討論你認為哪個attribute對結果影響最大？
- Report template：Link

Score - Policy

- Other policy:
  - 當script格式錯誤，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x0.7。
  - 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。
  - Kaggle超過deadline會直接shut down，可以繼續上傳但不計入成績。
  - Github遲交一天(*0.7)，不足一天以一天計算，不得遲交超過一天，有特殊原因請找助教。**不接受程式or報告單獨遲交**
  - Github遲交表單:Link
    (有遲交的同學才需填寫)，遲交時請「先上傳程式」到Github再填表單，助教會根據表單填寫時間當作繳交時間。

Score - 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

FAQ

- 1. 如果只有做兩個方法是否需要繳交第三份script hw2_best.sh？
  Ans: 是的。請把前兩個方法裡面較好的那份複製一份改名為hw2_best.sh

- 若有其他問題，請po在FB社團裡或寄信至助教信箱，**請勿直接私訊助教**。
- 助教信箱：ntumlta2019@gmail.com

Link

- Kaggle
  - https://www.kaggle.com/c/ml2019spring-hw2

- 網頁
  - https://ntumlta2019.github.io/ml-web-hw2/