Machine Learning HW2

MLTAs
ntumlta2018@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.

2. 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)).

3. 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's Description Page

Provided Feature Format

train_X, train_Y, test_X ：

1. discrete: one-hot encoding
2. continuous: remain the same
3. train_X, test_X: each row contains one 123-dim feature represents a sample
4. train_Y: 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
25,226802,1,0,0,40,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0
```

Requirements

1. 請手刻gradient descent實作logistic regression
2. 請手刻實作probabilistic generative model
3. 不能使用binary classification有關的現成package
4. 不能使用額外data
5. hw2_logistic.sh、hw2_generative.sh、hw2_best.sh皆須在10分鐘內跑完
6. Toolkit Versions:
    a. Only Python3.5+
    b. hw2_logistic.sh、hw2_generative.sh僅可使用numpy, pandas以及python standard library
    c. hw2_best.sh可額外使用tensorflow1.3, keras2.0.8, pytorch0.3.0
    d. hw2_best.sh若還有任何想用的額外套件請在社團詢問或寄信到助教信箱

Kaggle

1. kaggle
   url : https://www.kaggle.com/t/3092b8e4d3fd4d63ac2011c9b2904965
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。

Kaggle Submission Format

請預測test set中16281筆資料並將結果上傳Kaggle

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

Deadlines

1. Kaggle deadline: 2018/4/4 23:59:59 (GMT+8)
2. Github code & report deadline: 2018/4/5 23:59:59 (GMT+8)
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ML2018SPRING/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) $4: provided train label (Y_train)

$5: provided test feature (X_test) $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

Example Script

```
# using TA's feature
python hw2_logistic_train.py $3 $4
python hw2_logistic_test.py $5 $6
# feature extraction by yourself
python my_feature_extraction.py $1 $2
python hw2_logistic_train.py
python hw2_logistic_test.py $5 $6
```

```
# hw2_logistic_train.py
import sys
f_train = open(sys.argv[1], 'r')
f_label = open(sys.argv[2], 'r')
```

- 請勿將 data 路徑寫死在.py檔裡，請善加運用 sys.argv

Score - Kaggle Rank

- Kaggle Rank
  - (0.8%) 超過public leaderboard的simple baseline分數
  - (0.8%) 超過public leaderboard的strong baseline分數
  - (0.8%) 超過private leaderboard的simple baseline分數
  - (0.8%) 超過private leaderboard的strong baseline分數
  - (0.8%) 3/28 23:59 (GMT+8)前超過public simple baseline
  - (BONUS1%) kaggle排名前五名，且願意上台分享
  - **其中，前五名排名以private為準，屆時助教會公布名單**

Score - Reproduce

- 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
  - hw2_best.sh必須與kaggle上分數接近，才會有strong baseline的分數
  - **其中，上述提到的baseline皆以public以及private平均為準，重跑程式只是為了確認同學的程式可以正常執行，output部分會容許random造成的誤差，請同學不必特別擔心**

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：
  https://docs.google.com/document/d/1mhLEVu0OgKdbXgFdJpg0QMEipLET6fe4KiZGBITBugY/edit

Score - Policy

- Other policy:
  - script 錯誤, 直接0分。若是格式錯誤, 請在公告時間內找助教修好, 修完此次作業分數*0.7。
  - 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, 不得遲交超過兩天, 有特殊原因請找助教。
  - Github遲交表單:
    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。

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：ntumlta2018@gmail.com

小老師制度 (手把手教學)

- 在3/28以前超過simple baseline並願意在3/29在上課時間教導同學撰寫作業一程式，請填寫一下表單 ：

  https://goo.gl/forms/Rbj4VSoAiaMocPOl1

- 3/29將公布小老師名單在作業網頁，人數太多將以符合以下標準的同學為主：
  1. 沒有當過小老師
  2. Kaggle Public Leaderboard成績排名較高 (但請不要因此想overfit public set)

- 小老師當次成績+1%

Link

- Kaggle
  - https://www.kaggle.com/t/3092b8e4d3fd4d63ac2011c9b2904965

- 網頁
  - https://ntumlta2018.github.io/ml-web-hw2/