Machine Learning
HW2

TAs
ntu.mlta@gmail.com

Update (3/23 14:44)

根據下課後同學反應，以下投影片內容有稍作改動：

1. p12,13中，Script Usage中多一個參數Y_train。
2. p14中，第五個分數，時間修改為：3/29 23:59 (GMT+8)前超過。
3. p14中，kaggle前五名的計算，助教會參考public以及private的平均，時間到了之後會寄信通知。
4. p15中，關於程式重新批改部分，也是參考同學的程式在全部的test set的結果與baseline們在public以及private的平均是否相近來決定。
5. hw2_logistic.sh、hw2_generative.sh、hw2_best.sh皆須在10分鐘內跑完。

Outline

1. Dataset and Task Introduction
2. Provided Feature Format
3. Kaggle
4. Rules, Deadline and Policy
5. FAQ

Dataset and Task Introduction

1. ADULT dataset:

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. Task: **Binary Classification**

Determine whether a person makes over 50K a year.

1. Reference:

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

Attribute Information

```
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
```

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

- 資料詳細資訊在kaggle description

Provided Feature Format

```
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
```

X_train, Y_train, X_test ：

1. .csv 格式

2. X_train, X_test: 106 dim feature

3. discrete的資料: one-hot encoding

4. continuous的資料: 保留

Rules

1. 請手刻gradient descent實作logistic regression
2. 請手刻實作probabilistic generative model
3. 不能使用現成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. Only Python, C/C++，建議使用Python 3.4以及Numpy 1.12

Kaggle

1. kaggle_url：https://inclass.kaggle.com/c/ml2017-hw2

2. 請至kaggle創帳號登入，需綁定NTU信箱。

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1筆自行選擇的結果，測試在private set上的準確率為準。

★ kaggle名稱錯誤者將不會得到任何kaggle上分數。

Kaggle submission format

請預測test set中X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

Deadline

1. Kaggle deadline: 2017/04/05 11:59:59 p.m. (GMT+8)

2. Github code & report deadline: 2017/04/06 21:00 p.m.(GMT+8)

Policy

github上ML2017/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

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

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

./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

批改作業時會cd進同學的資料夾

Script Tutorial and Example

Shell Script Tutorial: http://linux.vbird.org/linux_basic/0340bashshell-scripts.php

Example:

```
# 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
```

- 請勿將 data 路徑寫死在.py檔裡，請善加運用 sys.argv

Score - Part1

- Kaggle Rank
  - (1%) 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%) 3/29 23:59 (GMT+8)前超過public simple baseline
  - (BONUS) kaggle排名前五名(且在4/14願意上台跟大家分享的同學)
  - **其中，前五名排名以public以及private平均為準，屆時助教會公布名單**

Score - Part1

- 批改方式
  - 除了在Kaggle上的資訊外，助教會測script中prediction.csv的結果。
  - hw2_logistic.sh 或 hw2_generative.sh的結果必須在test set上超過simple baseline，才會有simple baseline的分數。
  - hw2_best.sh的結果必須在test set上超過strong baseline，才能得到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 - Part2

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

- (1%) 請說明你實作的generative model，其訓練方式和準確率為何？
- (1%) 請說明你實作的discriminative 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。
- (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。
- (1%) 請討論你認為哪個attribute對結果影響最大？

Score - Part2

- 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kaggle分數*0.7。
  - Kaggle超過deadline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：
    https://docs.google.com/forms/d/e/1FAIpQLSdle52DaVorU6i5_llnlVByurPqDf4qqFRfVg-1AhFS3hM0AA/viewform?usp=pp_url&entry.1754524972(遲交才必需填寫)
    遲交請「先上傳程式」Github再填表單，助教會根據表單填寫時間當作繳交時間。

Report 是否可以用中文

有不少同學寄信詢問可否使用英文作答，在這裡統一回答同學。
除了母語人士或是國際生，助教**「強烈建議」你使用中文作答**，以下有幾點原因：

1. 屏除不必要的爭議以及方便批改，因為此次修課人數眾多，中文作答可以更快的讓同學拿到成績。
2. 同學們用英文作答可能會沒辦法完整表達意思，助教批改時也可能無法完全明白，可能造成同學們成績上的損失。

因此還是建議同學們**以中文作答，但如果仍想使用英文作答**也是可以的。

TAs

FAQ

- 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： ntu.mlta@gmail.com