ML hw2

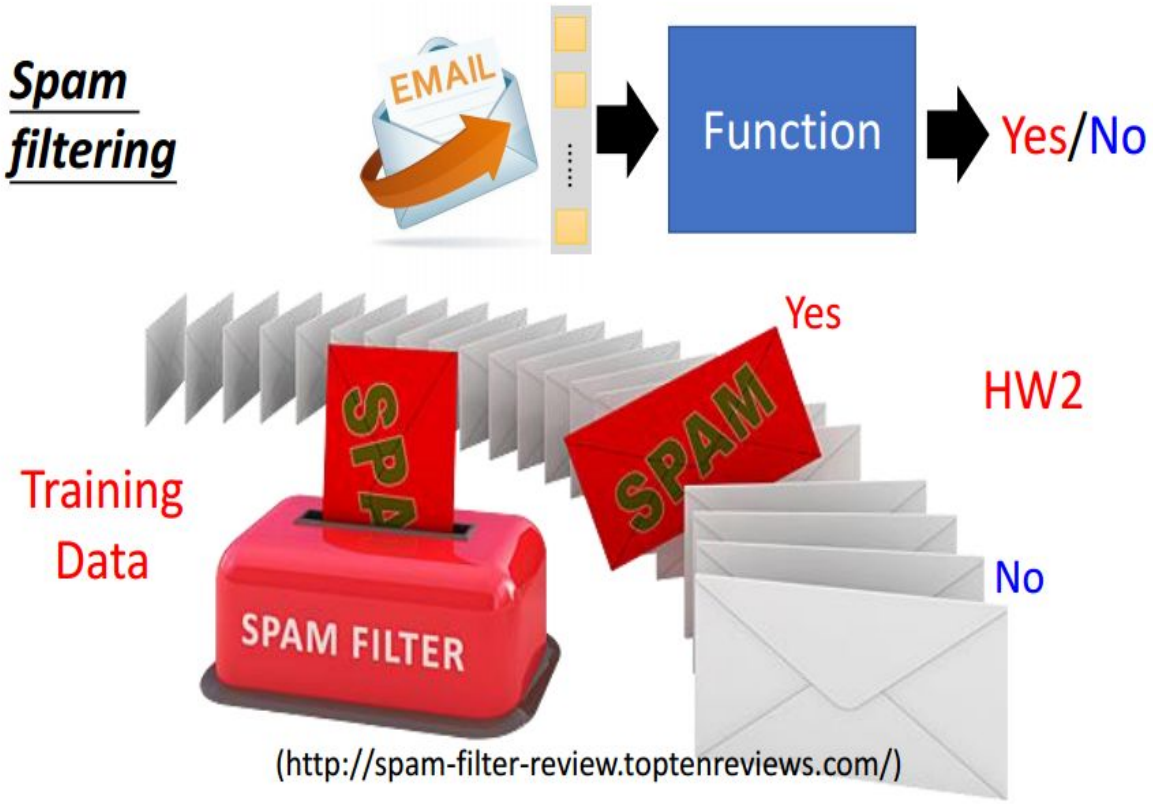
TAs
ml2016ta@gmail.com

Outline

1. Introduction
2. Data format
3. Kaggle
4. Policy
5. FAQ

Introdution

- Spam classification



Data format

- Number of Attributes: 59（data_id , feature , label ）
- data_id: first attributes
- feature: 57 dimensions
- label : 0 or 1 (last attributes)
- data 分為 spam_train.csv，spam_test.csv

```
@ spam_train                                                                                         X
0 1,0,0,1.13,0,0.37,0,0,0,0,0,0,0,0,0,0,0.37,0,0.37,1.13,0,0.37,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.145,0,0.436,0,0,1.792,55,147,0
1 2,0,0,0.6,0,0,0.6,0,0,0.6,0,0,1.82,0,0,0,0.3,0,0,2.74,0,1.21,0,0,0.6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.143,0.047,0.191,0.143,0,2.041,31,196,1
2 3,0,0,0.48,0,0,0,0,0,0,0,0,0,0,0,0,0.48,0,0,2.43,0,0.97,0,0,0.48,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.48,0,0,0,0,0,0,0.45,0,0,1.138,4,41,0
3 4,0,0.51,0,0,0.51,0.51,1.02,0,0,0,0,0,0,0,0,0,0.51,0.51,0,0.51,0,0.51,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.142,0,0.071,1.212,0,7.025,130,281,1
4 5,0,0,0,0,0,0,0,0,0.64,0,0,0,0,0,0,0,0,1.29,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.64,0,0,0,0,0.116,0,0.232,0,0,1.551,6,45,0
```

```
@ spam_public                                                                                        X
0 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3.03,0,3.03,3.03,0,3.03,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5.153,55,67
1 2,0,0,0,42.73,0,0,0.42,0,0,0.42,0,0.42,0,0,0.42,0,0,1.28,2.99,0,2.13,0,1.28,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.42,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.349,0,7,105,441
2 3,0,0.13,0.55,0,0.27,0.13,0,0,0,0.27,0,1.38,0,0,0,0.13,0,0,1.94,0,0.97,0,0,0,0.13,1.11,0,0,0,0,0,0,0,0,0,0.55,0,0,0,0,0,0.13,0,0,0,0,0,0.075,0.025,0,0.025,0,0,5.695
3 4,0.9,0,0,0,0,0,0,0,0,0,0.9,0,0,0,0,0,0,0,0,0,0,0.9,0.9,0,0,0,0,0,0,0,0.9,0,0,0.9,0,0,0,0,0,0,0,0,0,0,0,0,0.15,0,0,0,0,3.225,22,129
4 5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,4.34,0,0,0,0,0,0,0,0,4.34,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.735,0,0.735,0,0,2.571,10,18
```

Data format

- 48 dimensions : word_counting
- 6 dimensions : char_counting
- 1 dimension : capital_sequence_length_average
- 1 dimension : capital_sequence_length_longest
- 1 dimension : capital_number

```
spam_train
1,0,0,1.13,0,0.37,0,0,0,0,0,0,0,0,0,0.37,0,0.37,1.13,0,0.37,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.145,0,0.436,0,0,1.792,55,147,0
2,0,0,0.6,0,0,0.6,0,0,0.6,0,0,1.82,0,0,0,0.3,0,0,2.74,0,1.21,0,0,0.6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.143,0.047,0.191,0.143,0,2.041,31,196,1
3,0,0,0.48,0,0,0,0,0,0,0,0,0,0,0,0,0.48,0,0,2.43,0,0.97,0,0,0.48,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.48,0,0,0,0,0,0,0.45,0,0,1.138,4,41,0
4,0,0.51,0,0,0.51,0.51,1.02,0,0,0,0,0,0,0,0,0,0.51,0.51,0,0.51,0,0.51,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.142,0,0.071,1.212,0,7.025,130,281,1
5,0,0,0,0,0,0,0,0,0.64,0,0,0,0,0,0,0,0,1.29,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.64,0,0,0,0,0.116,0,0.232,0,0,1.551,6,45,0
```

```
spam_public
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3.03,0,3.03,3.03,0,3.03,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5.153,55,67
2,0,0,0,42.73,0,0,0.42,0,0,0.42,0,0.42,0,0,0.42,0,0,1.28,2.99,0,2.13,0,1.28,0,0,0,0,0,0,0,0,0,0,0,0,0,0.42,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.349,0,7,105,441
3,0,0.13,0.55,0,0.27,0.13,0,0,0,0.27,0,1.38,0,0,0,0.13,0,0,1.94,0,0.97,0,0,0,0.13,1.11,0,0,0,0,0,0,0,0,0,0,0.55,0,0,0,0,0,0.13,0,0,0,0,0,0.075,0.025,0,0.025,0,0,5.695
4,0.9,0,0,0,0,0,0,0,0,0,0,0.9,0,0,0,0,0,0,0,0,0,0,0.9,0.9,0,0,0,0,0,0,0.9,0,0,0.9,0,0,0,0,0,0,0,0,0,0,0,0.15,0,0,0,0,3.225,22,129
5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,4.34,0,0,0,0,0,0,0,4.34,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.735,0,0.735,0,0,2.571,10,18
```

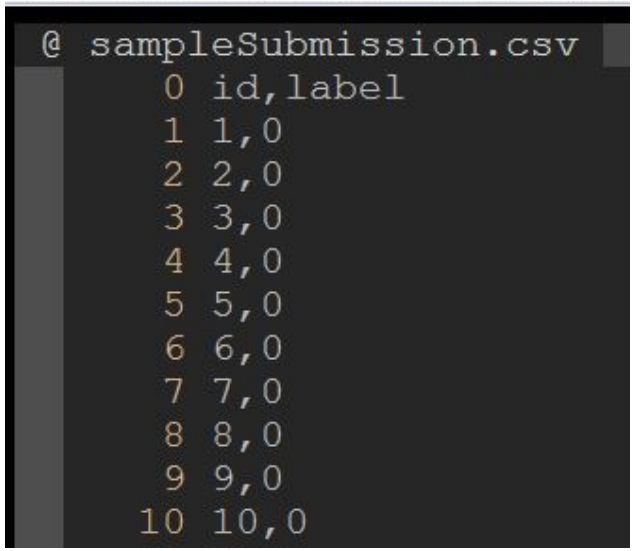
Kaggle

1. https://inclass.kaggle.com/c/spam-classification
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_任意隊名（有修課的同學），旁聽同學請避免學號開頭。
5. 每日上傳上限 5 次
6. test set 的 600 筆將被分為兩份，300筆 public，300筆 private
7. 最後的計分排名將以 private set 上為準。
8. kaggle deadline：2016/10/28 9:00:00 am (GMT+8)
9. github code & report deadline: 2016/10/28 21:00:00 pm (GMT+8)

Kaggle Submission format

預測 test set 中的 600 筆，上傳至 Kaggle。

- 上傳格式為 csv
- 第一行必須是 id,label
- 第二行開始，每行分別為 id 及預測的 label，以逗點分隔。
- Using Accuracy to evaluate



Rules

1. 請實作 logistic regression，並再實作另一種方法
2. logistic regression 方法請放 master，另一種則放branch
   branch name: method2
3. 不能使用現成 package
4. 不能使用額外的資料。
5. train.sh 需 10分鐘內跑完，test.sh 需10秒內跑完
6. Only Python & C/C++ is available.
7. * Please do not upload dataset.*

Rules

branch result

Branch: master ML2016 / hw2 / Create new file Upload files Find file History

poyuwu hw2_final Latest commit 6fca9be a minute ago

..

logistic_regression.py	hw2_final	a minute ago
test.sh	hw2_final	a minute ago
train.sh	hw2_final	a minute ago

Branch: method2 ML2016 / hw2 / Create new file Upload files Find file History

This branch is 4 commits ahead, 2 commits behind master. Pull request Compare

poyuwu hw2_final Latest commit f504bb3 10 minutes ago

..

distribution.py	hw2_final	10 minutes ago
test.sh	hw2_final	10 minutes ago
train.sh	hw2_final	10 minutes ago

Policy

ML2016/hw2/

請至少包含 Report.pdf (only on master), train.sh, test.sh。

Usage:
./train.sh $1 $2 輸出: your model
$1: training data, $2: output model

./test.sh $1 $2 $3 輸出:prediction.csv
$1: model name, $2 testing data, $3: prediction.csv

Policy

Kaggle Rank(4%): top10% 4; top20% 3.5; top50% 3;
beyond baseline 2
Report.pdf(4%): PDF, less than 2 pages, not include code.

1. (1%) Logistic regression function.
2. (1%) Describe your another method, and which one is best.
3. (2%) TA depend on your other discussion and detail.

另外請在 report 上註明討論的對象, if any.

LaTex: https://www.latex-project.org/

Format/github error (2%)

Policy

Other policy:
任一 script 錯誤(0分), 若是格式錯誤, 請來找助教修好(format part*0.5)

遲交每24小時(*0.7); 遲交超過 48 小時不收, 有特殊原因請洽助教。
遲交表單: https://goo.gl/K5aY81 (kaggle無法遲交)

Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.

FAQ

1. Anything about github?

https://github.com/Kunena/Kunena-Forum/wiki/Create-a-new-branch-with-git-and-manage-branches

2. Library 限制之問題

如果有使用到不知道能不能使用的library，請寫信或在FB社團上，跟助教確認並簡述用途，基本上如果是用來處理data的library都是可以的。

3. 請多用 facebook 問問題，你有問題的話，代表別人也有問題 :)