

---

---

# Machine Learning HW2

ML TAs

[ntu-ml-2020spring-ta@googlegroups.com](mailto:ntu-ml-2020spring-ta@googlegroups.com)

---

---

# Outline

- Task introduction and Dataset
- Feature Format and Submission Format
- Requirements

# Task Introduction and Dataset

- Task: **Binary Classification**  
whether the income of an individual exceeds \$50000 or not?
- Dataset: Census-Income (KDD) Dataset
  - remove unnecessary attributes and,
  - balance the ratio between positively and negatively labeled data.

# Feature Format

- train.csv, test\_no\_label.csv
  - text-based raw data
  - unnecessary attributes removed, positive/negative ratio balanced.
- X\_train, Y\_train, X\_test
  - discrete features in train.csv => one-hot encoding in X\_train (education, marital state...)
  - continuous features in train.csv => remain the same in X\_train (age, capital losses...).
  - X\_train, X\_test : each row contains one 510-dim feature represents a sample.
  - Y\_train: label = 0 means " $\leq 50K$ " 、 label = 1 means " $>50K$ "

# Submission Format

- 27622 instances of testing data
- First line: "id, label"
- Second line and thereafter: one "id, prediction" per line
- CSV(comma seperated values) format
- Evaluation metric: accuracy

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

# Requirements

- hw2\_logistic.sh: 請手刻 gradient descent 實作 logistic regression
- hw2\_generative.sh: 請手刻實作 probabilistic generative model
- hw2\_best.sh: 不限作法
- hw2\_logistic.sh, hw2\_generative.sh, hw2\_best.sh 皆須在 5 分鐘內跑完
- Please refer to [link](#) for allowed toolkits.
- Any open-sourced code is forbidden (e.g. Implementation of decision tree you find on GitHub).
- Ask if you want to use other toolkits before using them!!!

# Kaggle

- Kaggle competition: <https://www.kaggle.com/c/ml2020spring-hw2>
- Public simple baseline(1%): 0.88617
- Public strong baseline(1%): 0.89052
- Private baselines(2%): will be announced after Kaggle deadline.
- Kaggle scores will be counted if and only if the results can be reproduced by your GitHub code.

# GitHub Submissions

- The “hw2-<account>” directory on GitHub should contain at least (but not limited to) the following files:
  - report.pdf
  - hw2\_logistic.sh
  - hw2\_generative.sh
  - hw2\_best.sh
- Please DO NOT upload the 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 training data (train.csv) \$2: raw testing data (test\_no\_label.csv)
- \$3: preprocessed training feature (X\_train) \$4: training label (Y\_train)
- \$5: preprocessed testing feature (X\_test) \$6: output path  
(prediction.csv)
- You do not need to use all of the arguments in your bash scripts.
- **The TA will cd into the directory of your scripts before executing them.**

# Script Usage

- Example:
  - TA@TA's Computer: ~/...../b08940587\$ 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
- 不要寫死路徑 不要寫死路徑 不要寫死路徑
- 助教會把相對路徑帶入 \$N 所以:  
不要寫死路徑 不要寫死路徑 不要寫死路徑

# Reproducing Results

- Kaggle score will be counted if and only if the results can be reproduced by your GitHub code!!!
- Simple baselines: must be reproduced with hw2\_logistic.sh or hw2\_generative.sh
- Strong baselines: must be reproduced with hw2\_logistic.sh, hw2\_generative.sh, or hw2\_best.sh
- Only error less than **1%** can be accepted
  - For example, if your Kaggle score is 0.87, the accuracy of the result of your GitHub code should be at least  $0.87 \times 0.99 = 0.8613$ .
- Please always fix the random seeds in your code.

# Report

- 請比較實作的 generative model 及 logistic regression 的準確率，何者較佳？請解釋為何有這種情況？
- 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。接著嘗試對正規項使用不同的權重 ( $\lambda$ )，並討論其影響。  
(有關 regularization 請參考 <https://goo.gl/SSWGhf> p.35)

# Report

- 請說明你實作的 best model, 其訓練方式和準確率為何？
- 請實作輸入特徵標準化 (feature normalization), 並比較是否應用此技巧, 會對於你的模型有何影響。

# Links

- Data: <https://bit.ly/2wl4i9n>
- Kaggle: <https://www.kaggle.com/c/ml2020spring-hw2>
- Colab: <https://bit.ly/32D5h6B>
- Report template: <https://bit.ly/32CIs2U>
- 遲交表單: <https://bit.ly/39d2x2m>