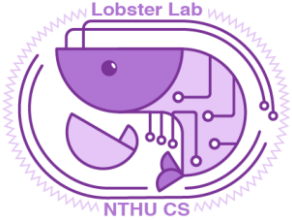
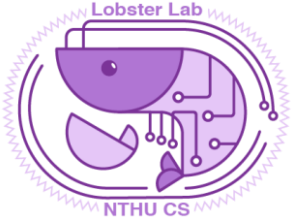


Data science HW3

Department of Computer Science
National Tsing Hua University (NTHU)
Hsinchu, Taiwan

Date: April 24th, 2020

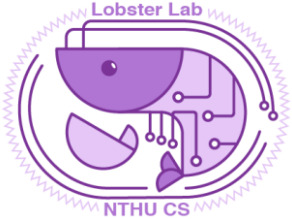


Kaggle

- A platform of
  - Machine learning competition
  - Sharing dataset
- https://zh.wikipedia.org/wiki/Kaggle

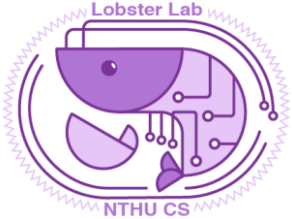
kaggle

- HW3 will be held on kaggle
  - https://www.kaggle.com/c/nthuds2020hw3-1/data
- Deadline: 2019/5/12 23:59



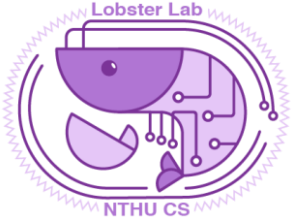
HW3

- Problem description
  - Supervised binary classification problem
  - Given a data set
  - Training set with label, testing set without
  - You need to predict the labels of testing data
- Evaluation
  - F1-score
  - $2 \times \frac{precision \times recall}{precision + recall}$



Dataset description

- The dataset we use is **transformed** from some real dataset
  - Numeric feature are nonlinear transformed
  - 10% data become missing value
- 16 numeric features, 5 nominal features, 1 label
- Our label is '**RainToday**'



Baseline method

- We provide a simple baseline method for your reference
- The steps in baseline are as below
  - Read training/testing data
  - Fill missing value with mode/mean
  - Train a decision tree classifier
  - Output prediction

Output format

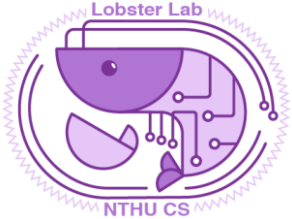
- For each testing instance, there is a unique id
- You need to submit you answer to kaggle with the following format
  第一行請記得也要**output**
  - Id,RainToday
  - Id1, RainToday1
  - Id2, RainToday2
  - …

```
Id,RainToday
0,0
1,0
2,0
3,0
4,0
5,1
6,0
7,0
8,0
```

Evaluation

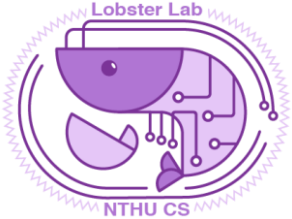
- There are two leaderboards in kaggle
  - Public: can be seen during competition, for reference
  - Private: used to evaluate, can be seen after competition

#	Team Name	Notebook	Team Members	Score	Entries	Last
	baseline 80			0.42890		
	baseline 70			0.36402		
	baseline 60			0.34381		
	baseline 0			0.34267		
	random			0.16401		



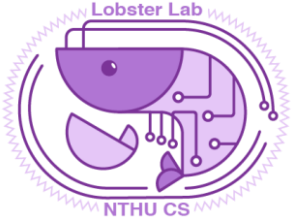
Evaluation

- You will get **60** points, if your private F1-score is between *baseline 60* and *baseline 70*
- You will get **75** points, if your private F1-score is between *baseline 70* and *baseline 80*
- For those have scores better than *baseline 80*
  - TOP 10% (<=10%): 100
  - 10% ~ 30% (<=30%): 92
  - 30% ~ 60% (<=60%): 86
  - 60% ~ 80% (<=80%): 83
  - Other: 80



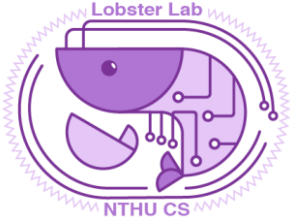
Baseline score in public and private

	Public	Private
Baseline 80	0.38559	0.41335
Baseline 70	0.33626	0.34511
Baseline 60	0.28667	0.30352
Baseline 0	0.26903	0.29067



Hints

- You can try more techniques for better performance
  - Feature selection
  - Dimension reduction (PCA, TSNE)
  - Try different models
  - Data augmentation
  - …
- We use private leaderboard as the final score
  - Use public score to choose your model is dangerous
  - It’s better to perform validation



Packages you may use

- Scikit-learn
  - https://scikit-learn.org/stable/index.html
- Pandas
  - https://pandas.pydata.org/pandas-docs/stable/
- Imbalance learn (for over sampling and down sampling)
  - https://imbalanced-learn.readthedocs.io/en/stable/

Other rules

- You can submit 15 times per day
- You can choose 4 predictions for final scoring