Data Science HW 3

Dimension Reduction

Submission Deadline

- 2018/5/8 23:59
- Submit to E3
- HARD deadline, NO extensions

Goal

Given: Datasets

Goal:

1. Hand-crafted dimension reduction of the datasets with PCA/ICA/SVD/Feature Selection
2. Evaluate the performance(F1-score) of SVM classification after reducing the dimension
3. Find the best dimensionality of each dataset

Datasets

Dataset 1

- # of classes: 2
- # of data: ~10K
- # of features: 68
- Features include:
  - IP Address, Long URL to Hide the Suspicious Part, TinyURL

Datasets

Dataset 2

- # of classes: 2
- # of data: ~1.5k
- # of features: 123
  - 14 features, among which six are continuous and eight are categorical.
  - A categorical feature with m categories is converted to m binary features.

Datasets

Dataset 3

- # of classes: 10
- # of data: ~10k
- # of features: 256

Datasets

Dataset 4

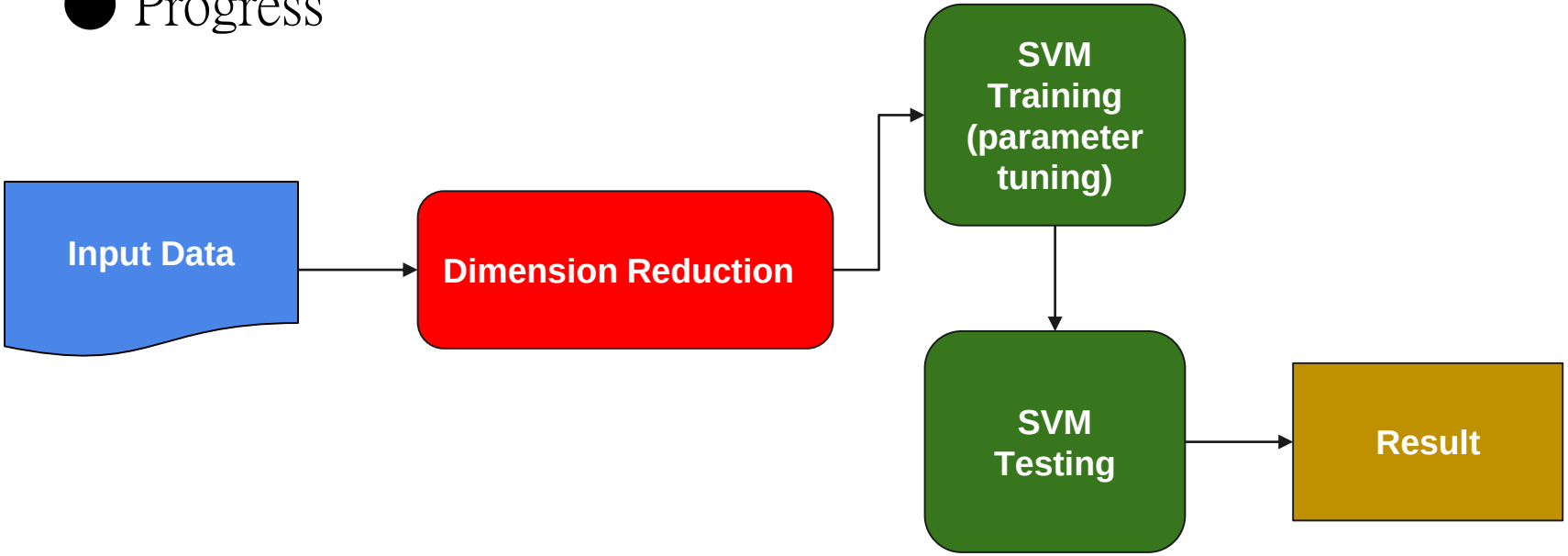
- # of classes: 3
- # of data: ~70K
- # of features: 126

Requirements

- Implement one of the dimension reduction approaches with Python 3.6
- Strictly follow input/output formats
- Do not copy/paste others' codes
- You can refer to the codes on GitHub or anywhere else, but please write your own code
- *eigh can be used

Requirements

- Progress



Requirements

- Input Format
  - The given datasets are in LibSVM Format
  - The labels need to be separated with the features!

Labels

Features

```
5 1:-1.000000 2:-1.000000 3:-1.000000 4:-0.999995 5:-0.999986
8 1:-0.999998 2:-0.999800 3:-0.994136 4:-0.932288 5:-0.673685
4 1:-1.000000 2:-0.999971 3:-0.997452 4:-0.957499 5:-0.801681
7 1:-1.000000 2:-1.000000 3:-0.999851 4:-0.988518 5:-0.820386
4 1:-0.999946 2:-0.995190 3:-0.916669 4:-0.577995 5:-0.041746
2 1:-1.000000 2:-1.000000 3:-1.000000 4:-1.000000 5:-0.999802
1 1:-1.000000 2:-0.999977 3:-0.998420 4:-0.969256 5:-0.774192
2 1:-1.000000 2:-1.000000 3:-1.000000 4:-1.000000 5:-0.999965
8 1:-1.000000 2:-0.999996 3:-0.999617 4:-0.986903 5:-0.839243
1 1:-1.000000 2:-0.999998 3:-0.999760 4:-0.992020 5:-0.901515
2 1:-1.000000 2:-1.000000 3:-1.000000 4:-0.999988 5:-0.998782
2 1:-1.000000 2:-1.000000 3:-1.000000 4:-1.000000 5:-0.999790
8 1:-1.000000 2:-1.000000 3:-1.000000 4:-0.999998 5:-0.999814
8 1:-0.999988 2:-0.998820 3:-0.968097 4:-0.690011 5:-0.034844
5 1:-1.000000 2:-0.999898 3:-0.991689 4:-0.876872 5:-0.635230
```

Requirements

- Submission contains 3 files:
  - File Name: [studentID].py
    - 0680708.py (O) [0680708].py (X)
    - Method: [studentID].py [datasetname].txt
    - Output: [datasetname]_out.txt
  - Parameter file: param.txt
    - C:1,kernek:RBF,gamma:0.01

Grading Policy

- TA will execute your code.
- There are 4 test cases.
- For your convenience to tune the parameters, here is the package of sampled datasets:
  - https://drive.google.com/open?id=1No4iklZY5uwYedBVWjE_e_7643kkAWKS

Grading Policy

分類正確度比拚(交大+清大)～～太慢或是太差就沒得比了! 全部比賽人數平均成四個正確度等級

	Correctness (F1-score)	Effectiveness
Dataset 1	15% (F1-score>=0.86, time < 3 sec)	10%/8%/6%/4%
Dataset 2	15% (F1-score: >=0.76, time < 3 sec)	10%/8%/6%/4%
Dataset 3	15% (F1-score: >=0.86, time < 6 sec)	10%/7%/3%/1%
Dataset 4	15% (F1-score: >=0.81, time < 10 sec)	10%/7%/3%/1%

Contact Information

- If you have any question about HW#3, please email to 蔡睿翊.
  - vincentthunder2011@gmail.com
- FB group:
  - https://www.facebook.com/groups/156477025052136/