

ML in EDU

Homework 2

Regression & Classification

Help: <https://hackmd.io/xjPlheHATE23Ez7qOzaogw?view>

CONTENTS

全國高中學校畢業生數

MEAP93

1-1

1-2

2-1

2-2

**Data Preprocess
(16%)**

**Polynomial
Regression
(24%)**

**Single Regression
&
Multiple Plots
(32%)**

**Multiple Regression
(8%)**

<Hint> You maybe need these packages:

`pandas, matplotlib, statsmodels, scikit-learn`

CONTENTS

Iris

3-1

**Load Iris Data
(5%)**

3-2

**5-fold CV
(5%)**

3-3

**Classification
(5%)**

3-4

**Confusion Matrix
(5%)**

<Hint> You maybe need these packages:

`pandas, matplotlib, statsmodels, scikit-learn`

1-1

Data Preprocess (4 * 4%)

全國高級中等學校畢業生數

學年別 (X)	總計 (Y)
106學年	241,288
105學年	233,642
104學年	250,172
103學年	272,662
102學年	277,047
101學年	277,910
100學年	279,381
99學年	282,605
98學年	278,717
97學年	277,150
96學年	279,320

1. 請讀入給定的資料集 `graduates.csv`
並補上最後一筆資料 {'96學年': '279,320'}

2. 請將資料的中文部分移除

3. 請將千分位的逗號移除，
並確認該欄位的 `type` 並非字串
以便數值運算


4. 將結果輸出成一個 `csv` 檔，命名為
`<student_ID>_graduates.csv`

e.g. `0856029_graduates.csv`

<Hint> 必須使用程式產生此檔案，禁止手打

1-1

Data Preprocess (4 * 4%)



year	graduates
106	241,288
105	233,642
104	250,172
103	272,662
102	277,047
101	277,910
100	279,381
99	282,605
98	278,717
97	277,150
96	279,320

1. 請讀入給定的資料集 `graduates.csv`
並補上最後一筆資料 {'96學年': '279,320'}

2. 請將資料的中文部分移除

3. 請將千分位的逗號移除，
並確認該欄位的 `type` 並非字串
以便數值運算

4. 將結果輸出成一個 `csv` 檔，命名為
`<student_ID>_graduates.csv`

e.g. `0856029_graduates.csv`

<Hint> 必須使用程式產生此檔案，禁止手打

1-1

Data Preprocess (4 * 4%)

year	graduates
106	241288
105	233642
104	250172
103	272662
102	277047
101	277910
100	279381
99	282605
98	278717
97	277150
96	279320

1. 請讀入給定的資料集 `graduates.csv`
並補上最後一筆資料 {'96學年': '279,320'}

2. 請將資料的中文部分移除

3. 請將千分位的逗號移除，
並確認該欄位的 `type` 並非字串
以便數值運算

4. 將結果輸出成一個 `csv` 檔，命名為
`<student_ID>_graduates.csv`

e.g. `0856029_graduates.csv`

<Hint> 必須使用程式產生此檔案，禁止手打

1-1

Data Preprocess (4 * 4%)

jupyter 0856029_graduates.csv	
文件	编辑 查看 语言
1	year,graduates
2	106,241288
3	105,233642
4	104,250172
5	103,272662
6	102,277047
7	101,277910
8	100,279381
9	99,282605
10	98,278717
11	97,277150
12	96,279320
13	

1. 請讀入給定的資料集 `graduates.csv`
並補上最後一筆資料 {'96學年': '279,320'}

2. 請將資料的中文部分移除

3. 請將千分位的逗號移除，
並確認該欄位的 `type` 並非字串
以便數值運算

4. 將結果輸出成一個 `csv` 檔，命名為
`<student_ID>_graduates.csv`

e.g. `0856029_graduates.csv`

<Hint> 必須使用程式產生此檔案，禁止手打



(Polynomial) Regression

Using `processed_graduates.csv`

1. (8%)

Let $y = \beta_0 + \beta_1 x$, use the least squares method to find coefficients β_0, β_1 .
Calculate R squared (R^2).
Predict Y at $X=107, 108, \dots, 111$.

2. (8%)

Let $y = \beta_0 + \beta_1 x + \beta_2 x^2$, use the least squares method to find coefficients $\beta_0, \beta_1, \beta_2$.
Calculate R squared (R^2).
Predict Y at $X=107, 108, \dots, 111$.

3. (8%)

Compare the two models and describe what you found. (Write in your report)

2-1

Single regression (32%)

- 1. 請讀入給定的資料集 `MEAP93.csv`
- 2. 用其餘 16 個 attribute 作單變數迴歸預測 `math10` , 並作圖。

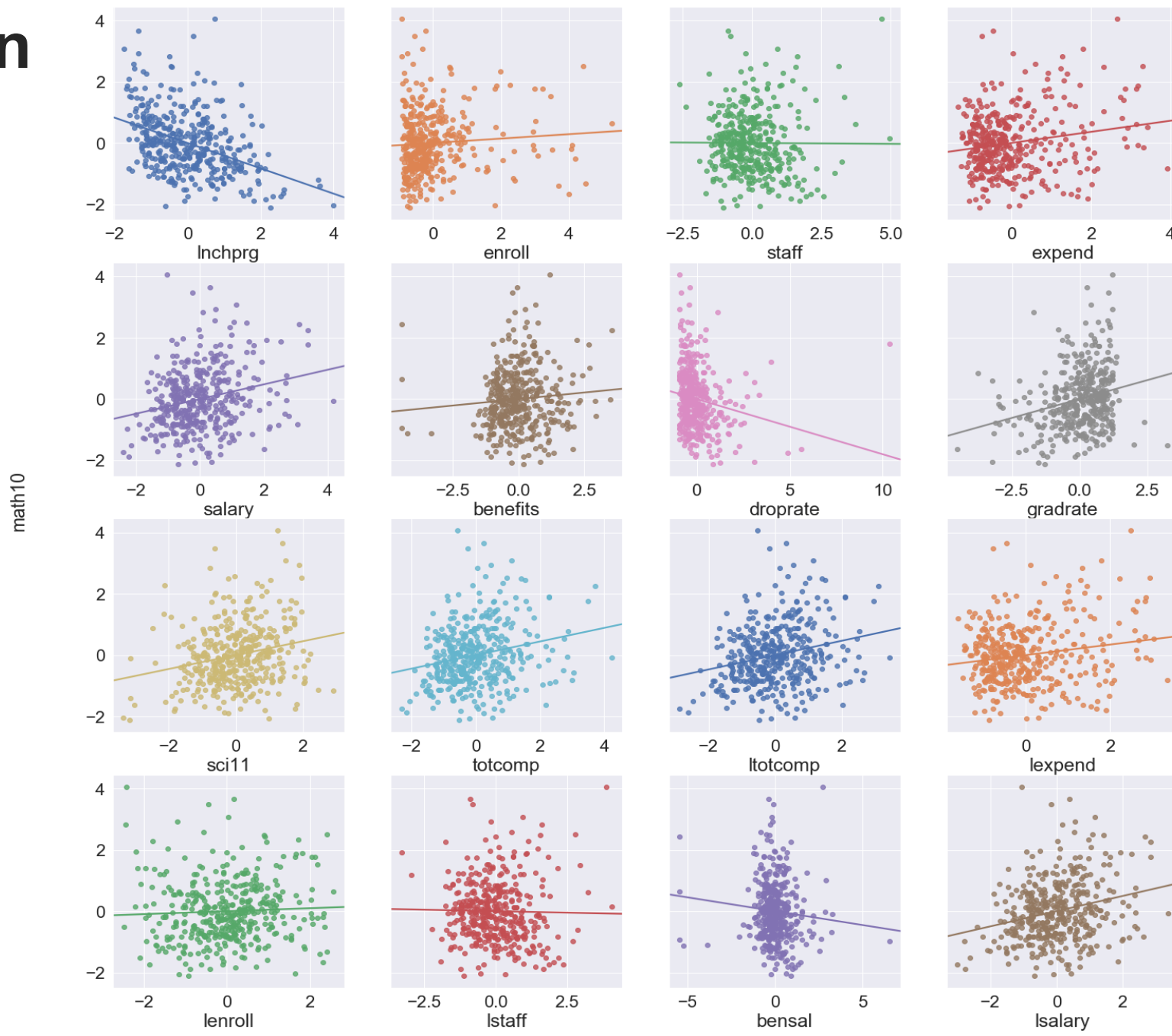
	Inchprg	enroll	staff	expend	salary	benefits	droprate	gradrate	math10
0	1.4	1862	112.599999	5765	37498	7420	2.9	89.199997	56.400002
1	2.3	11355	101.199997	6601	48722	10370	1.3	91.400002	42.700001
2	2.7	7685	114.000000	6834	44541	7313	3.5	91.400002	43.799999
3	3.4	1148	85.400002	3586	31566	5989	3.6	86.599998	25.299999

2-1

Single regression (2% in each plot)

1. 請讀入給定的資料集 `MEAP93.csv`
2. 用其餘 16 個 attribute 作單變數迴歸預測 `math10`，並作圖。

16 Single regression to fit `math10`



2-2

Multiple regression (10%)

1. 使用任意 attribute 和任意 model
使得你的 R-squared 超過 0.25

<Note> 你可以自己決定 Training/Testing set 的分割方式

R2	
lasso1	0.226157
lasso2	0.237468
lasso3	0.252793
lasso4	0.247986
lasso5	-0.513768
R2	
ridge1	0.209356
ridge2	0.231698
ridge3	0.291693
ridge4	-0.0613976
ridge5	-14.6938

3-1

Load Iris data (5%)

1. 使用 sklearn 的 load_iris() 載入資料
2. 使用 5-fold cross validation
3. 使用任意 model 進行分類，並印出 accuracy (分對的比例)
4. 畫出/印出 confusion matrix
請在報告中寫下你分類的結果

Ref:

https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html

```
from sklearn.datasets import load_iris
```

```
raw = load_iris()  
X = pd.DataFrame(raw.data, columns=raw.feature_names)  
y = pd.DataFrame(raw.target, columns=['class'])
```

```
pd.concat([X,y], sort=False, axis=1).head()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	class
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

3-2

K-fold Cross Validation (5%)

1. 使用 sklearn 的 load_iris() 載入資料
2. 使用 5-fold cross validation
3. 使用任意 model 進行分類，並印出 accuracy (分對的比例)
4. 畫出/印出 confusion matrix
請在報告中寫下你分類的結果

Ref: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

```
from sklearn.model_selection import KFold
from sklearn.datasets import load_iris
raw = load_iris()
X = raw.data
y = raw.target
kf = KFold(n_splits=5)
kf.get_n_splits(X)

print(kf)

for train_index, test_index in kf.split(X):
    # split data to 80% training set & 20% testing set
    print("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = X[train_index], X[test_index]
    y_train, y_test = y[train_index], y[test_index]

    # Train your model

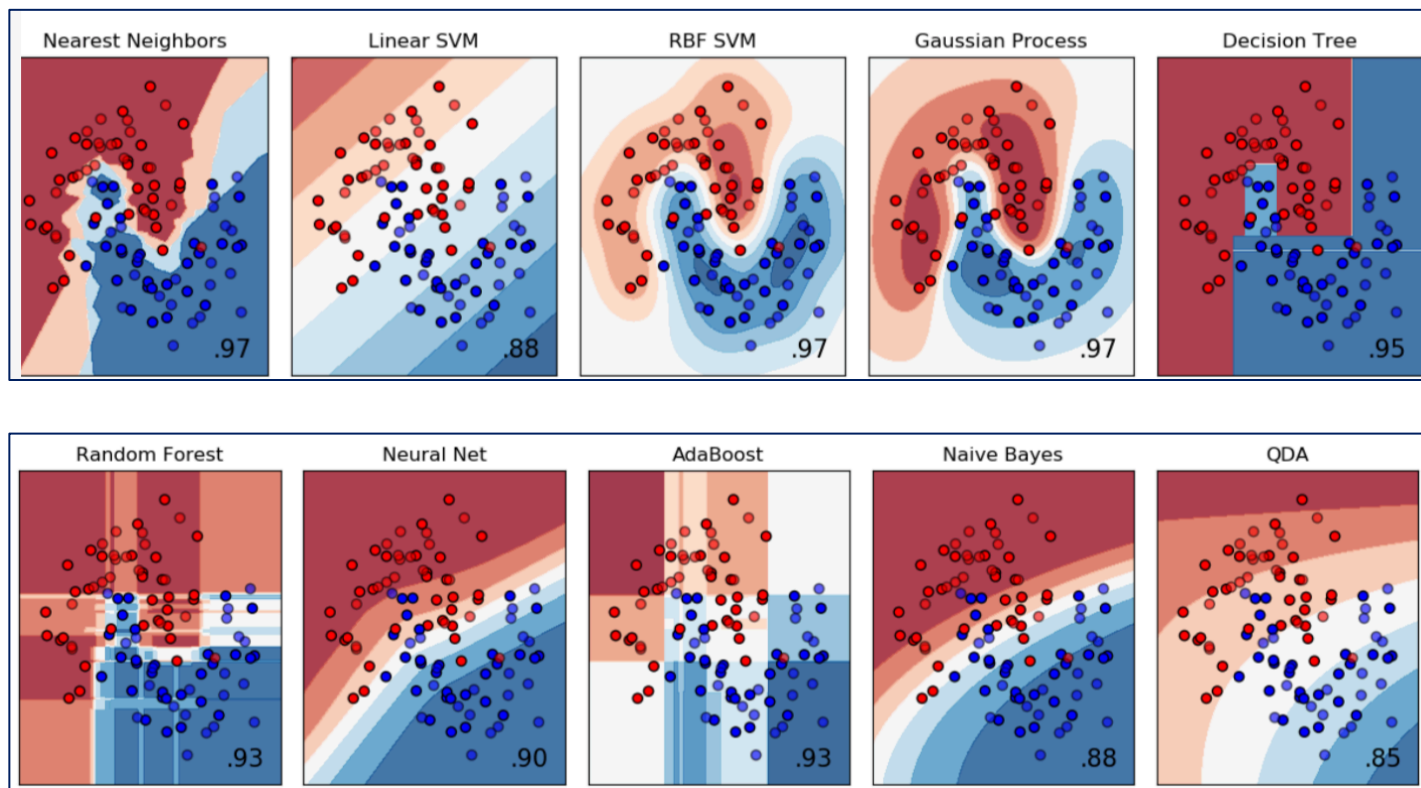
    # Test your model

    # Confusion matrix
```

3-3

Classification (5%)

1. 使用 sklearn 的 `load_iris()` 載入資料
2. 使用 5-fold cross validation
3. 使用任意 model 進行分類，並印出 accuracy (分對的比例)
4. 畫出/印出 confusion matrix
請在報告中寫下你分類的結果



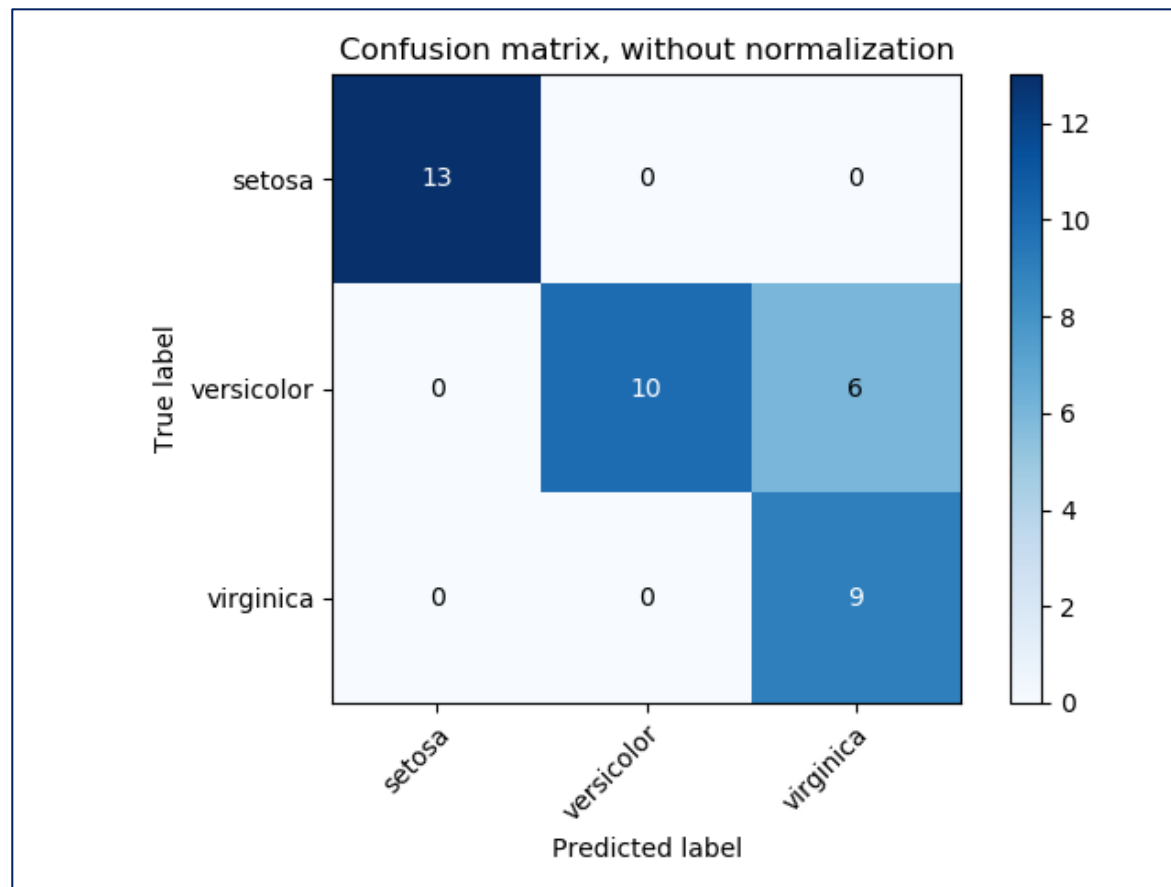
<Note> 你可以使用任意分類器，當然也推薦使用 sklearn package 以外的函式。

Ref: https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

3-4

Confusion Matrix (5%)

1. 使用 sklearn 的 `load_iris()` 載入資料
2. 使用 5-fold cross validation
3. 使用任意 model 進行分類，並印出 accuracy (分對的比例)
4. 畫出/印出 confusion matrix
請在報告中寫下你分類的結果



https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html#sphx-glr-auto-examples-model-selection-plot-confusion-matrix-py



Hand in your homework to e3

Hand in your report & code to e3(<https://e3new.nctu.edu.tw>)

Briefly describe how your code works and show results. Make sure TA could run your code.

You should only hand in 2 files:

`hw2_<student_id>.pdf`

`hw2_<student_id>.zip`

(e.g. hw2_0123456.pdf, hw2_0123456.zip)

Due: 10/22 (Tue.) 11:00 a.m.

TA:

蔡旻均 dollars9256741@gmail.com

劉昱劭 ysl@cs.nctu.edu.tw