A Shortcut to Data Analyst

讓數據說話

Why？

資料可以預測未來

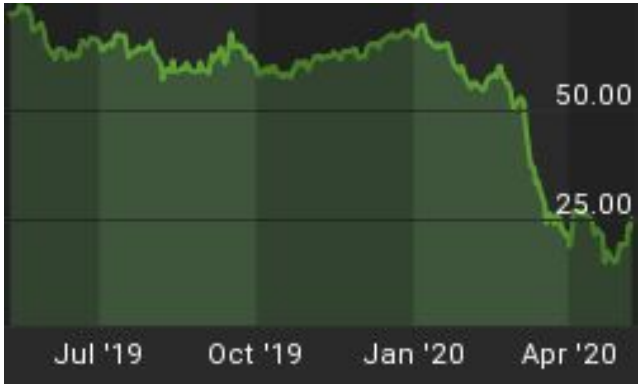
How？

分類分析/迴歸分析

資料可以預測未來

Example

- 大氣數據 ⇨ 颱風路徑
- 國際情勢 ⇨ 油價預測
- 瀏覽頁面 ⇨ 推薦商品



分類v.s迴歸

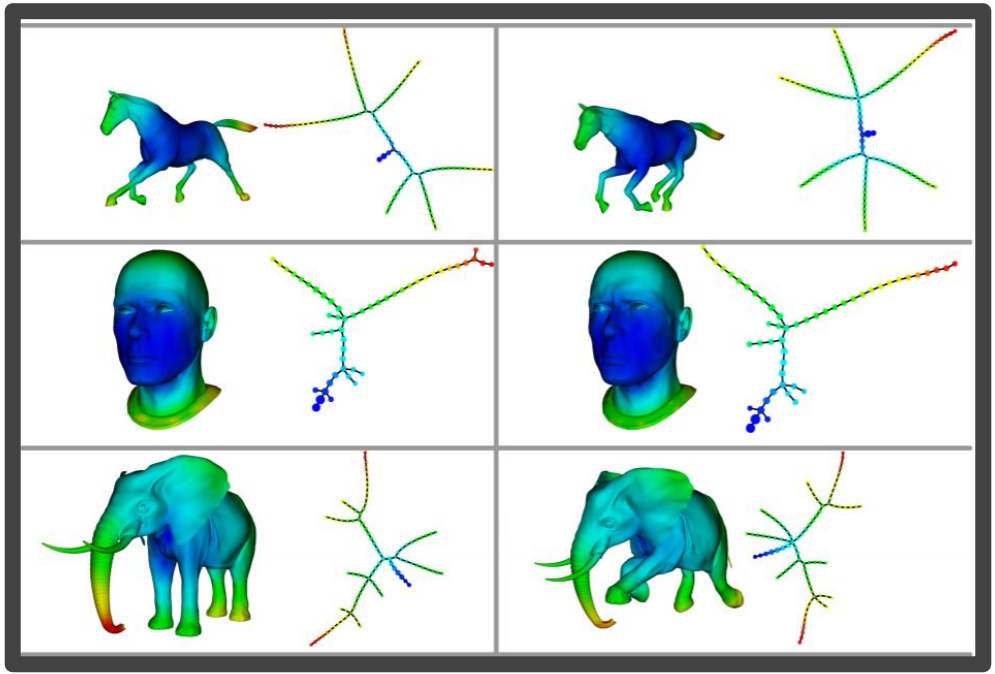
- 分類(Classification)

利用特徵去辨別東西

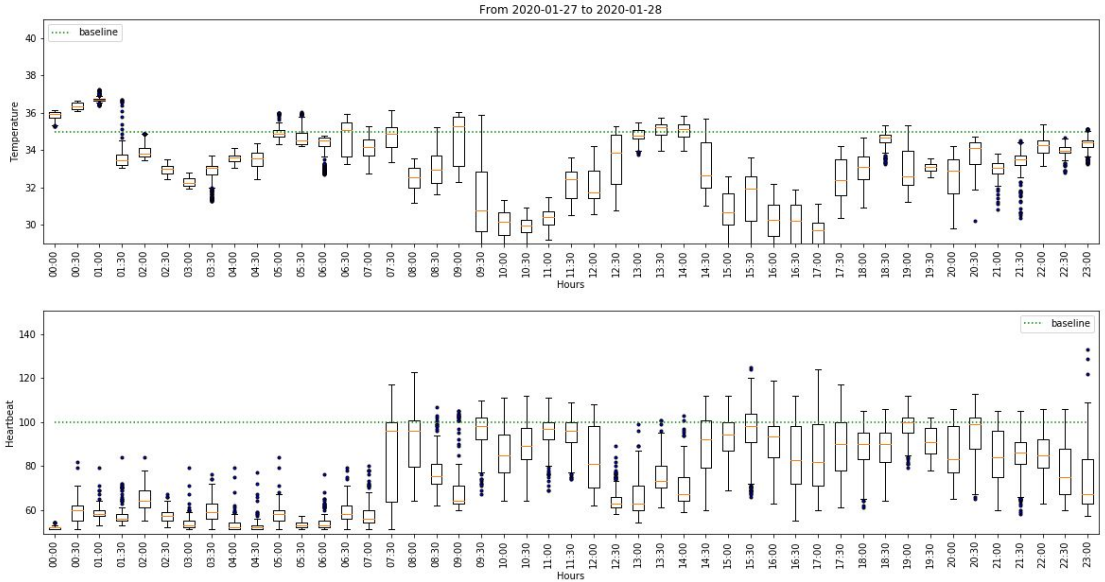
- 回歸(Regression)

利用歷史資料預測未來數據

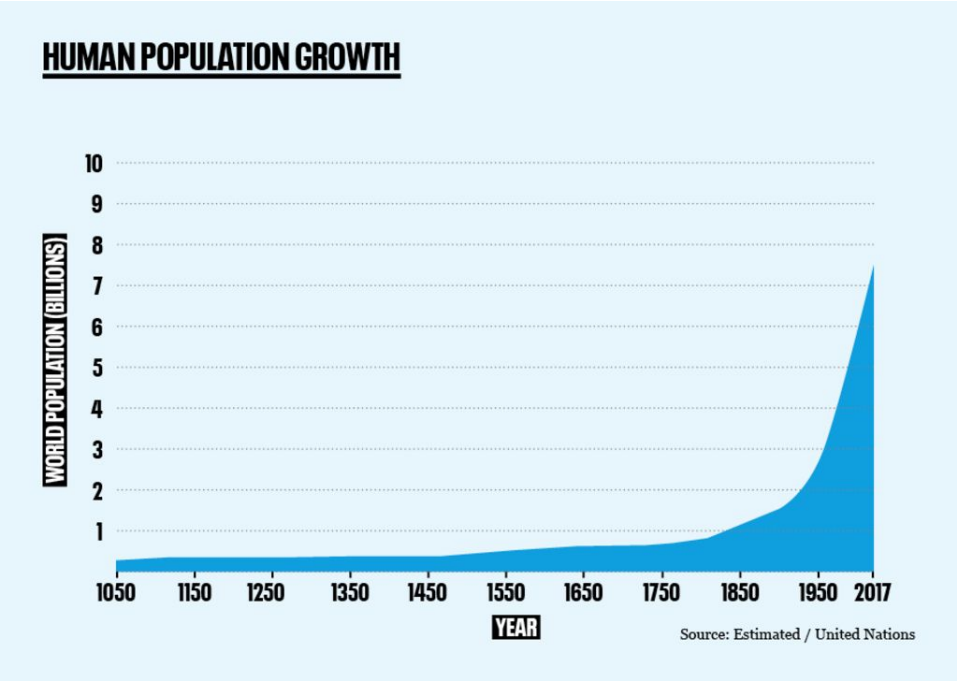
分類(Classification)



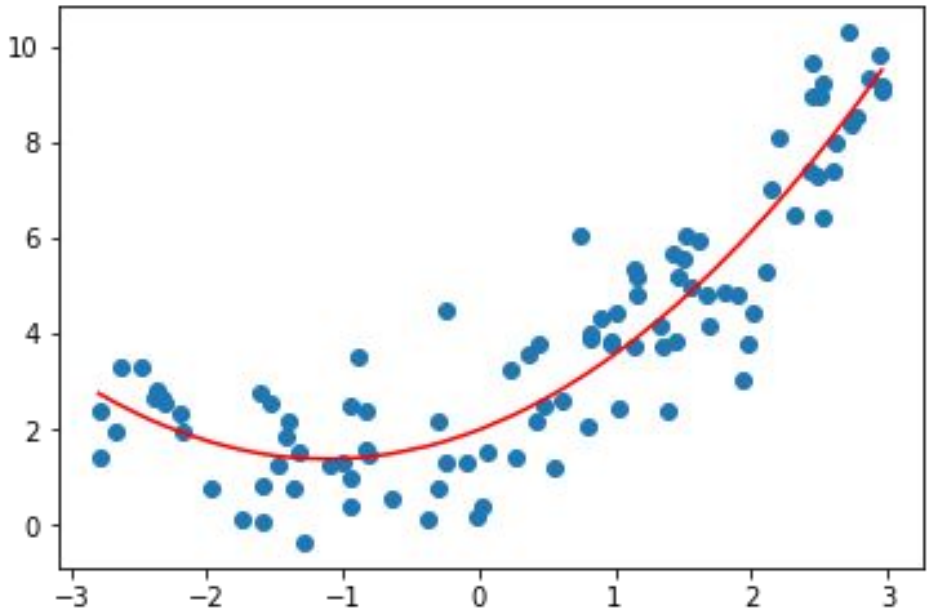
分類(Classification)



廻歸(Regression)



回歸(Regression)



Classification (IRIS DATASET)

花萼長度	花萼寬度	花瓣長度	花瓣寬度	花的種類
5.1	3.5	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
6.3	2.3	4.4	1.3	versicolor
5.8	2.7	5.1	1.9	versicolor
6.9	3.1	5.4	2.1	virginica

Classification (IRIS DATASET)

花萼長度	花萼寬度	花瓣長度	花瓣寬度
5.1	3.5	1.4	0.2
4.7	3.2	1.3	0.2
6.3	2.3	4.4	1.3
5.8	2.7	5.1	1.9
6.9	3.1	5.4	2.1

花的種類
setosa
setosa
versicolor
versicolor
virginica

花的種類
0
0
1
1
2

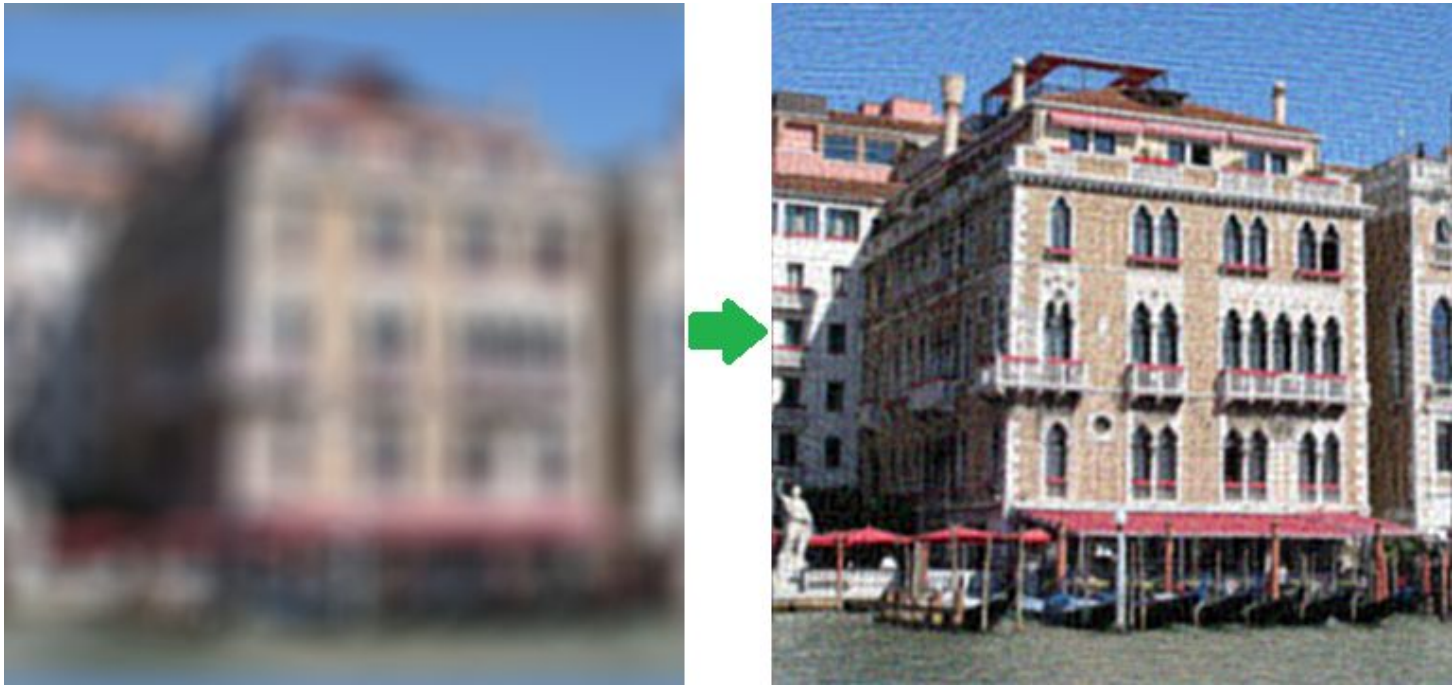
資料分析步驟

- 前處理
- 視覺化
- 特徵提取
- 選擇適合的函數
- 評估好壞

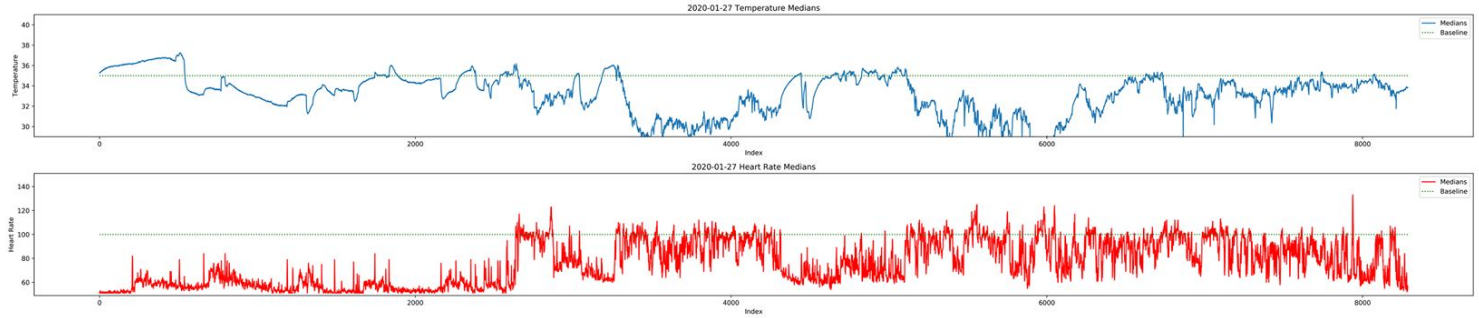
前處理

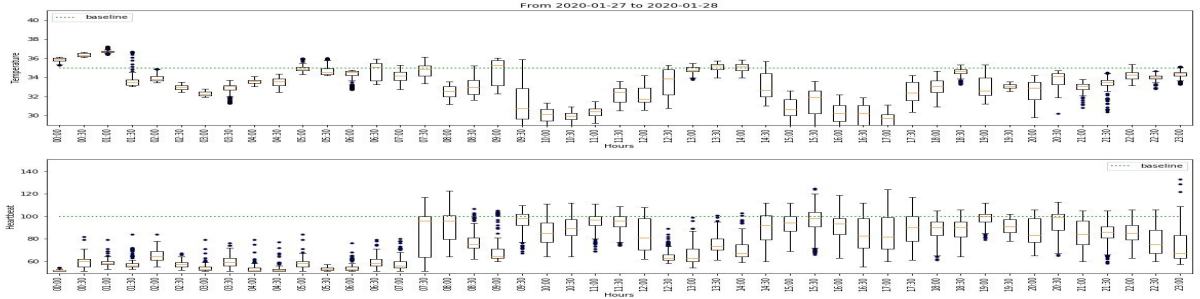
- 資料的雜訊太多 (訊號類型)
- 資料的range過大 (單純數字類型)
- 資料分割 (數字類型比較多)

雜訊過多



雜訊過多


2020-01-27 Temperature Medians
Medians
Baseline
Temperature
Index
2020-01-27 Heart Rate Medians
Medians
Baseline
Heart Rate
Index


From 2020-01-27 to 2020-01-28
baseline
Temperature
Hours
baseline
Heartbeat
Hours

Classification (IRIS DATASET)

花萼長度	花萼寬度	花瓣長度	花瓣寬度
5.1	3.5	1.4	0.2
4.7	3.2	1.3	0.2
6.3	2.3	4.4	1.3
5.8	2.7	5.1	1.9
6.9	3.1	5.4	2.1

花的種類
setosa
setosa
versicolor
versicolor
virginica

花的種類
0
0
1
1
2

資料的range過大

回到iris data

把數值都控制在特定區間，例如[-1, 1], [0,1]之類

- **方法：**
  - 正規化
  - X-min(X)/(max(X)-min(X))

先看看range的大小

花萼長度	花萼寬度	花瓣長度	花瓣寬度
5.1	3.5	1.4	0.2
4.7	3.2	1.3	0.2
6.3	2.3	4.4	1.3
5.8	2.7	5.1	1.9
6.9	3.1	5.4	2.1

```
maxima of column 0 : 7.9
mimima of column 0 : 4.3
maxima of column 1 : 4.4
mimima of column 1 : 2.0
maxima of column 2 : 6.9
mimima of column 2 : 1.0
maxima of column 3 : 2.5
mimima of column 3 : 0.1
```

處理方法

- **from sklearn import preprocessing**

  scaler = preprocessing.MinMaxScaler(feature_range=(0, 1))

  scaler.fit(data)

  data = scaler.transform(data)

MinMaxScaler

```
maxima of column 0 : 7.9
mimima of column 0 : 4.3
maxima of column 1 : 4.4
mimima of column 1 : 2.0
maxima of column 2 : 6.9
mimima of column 2 : 1.0
maxima of column 3 : 2.5
mimima of column 3 : 0.1
```

```
maxima of column 0 : 1.0
mimima of column 0 : 0.0
maxima of column 1 : 1.0
mimima of column 1 : 0.0
maxima of column 2 : 1.0
mimima of column 2 : 0.0
maxima of column 3 : 1.0
mimima of column 3 : 0.0
```

處理方法

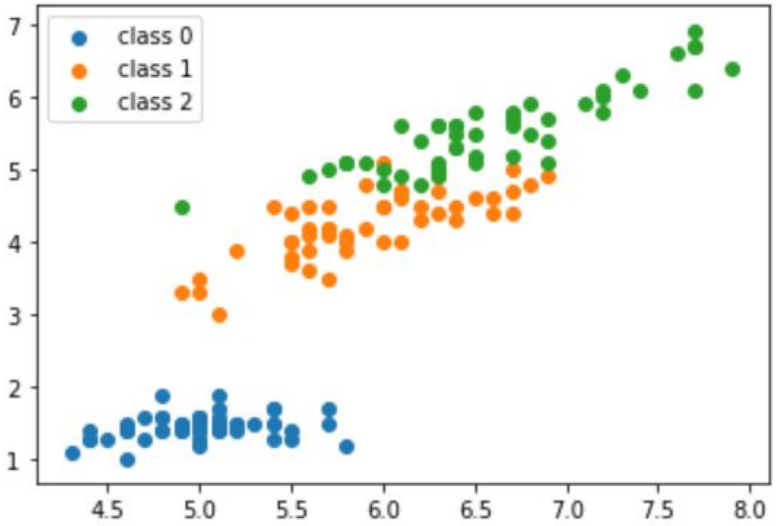
- **from sklearn import preprocessing**

  data_normalized = preprocessing.normalize(data, norm='l2')

視覺化(visualization)

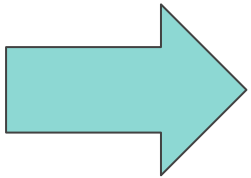
將資料放到R2空間裡面

花萼長度	花萼寬度	花瓣長度	花瓣寬度
5.1	3.5	1.4	0.2
4.7	3.2	1.3	0.2
6.3	2.3	4.4	1.3
5.8	2.7	5.1	1.9
6.9	3.1	5.4	2.1



選擇適合的函數

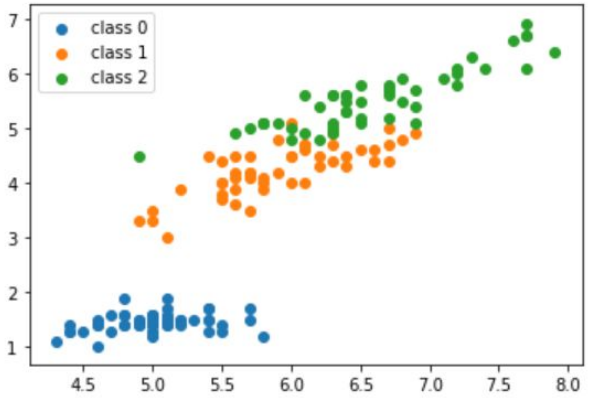
花萼長度	花萼寬度	花瓣長度	花瓣寬度
5.1	3.5	1.4	0.2
4.7	3.2	1.3	0.2
6.3	2.3	4.4	1.3
5.8	2.7	5.1	1.9
6.9	3.1	5.4	2.1



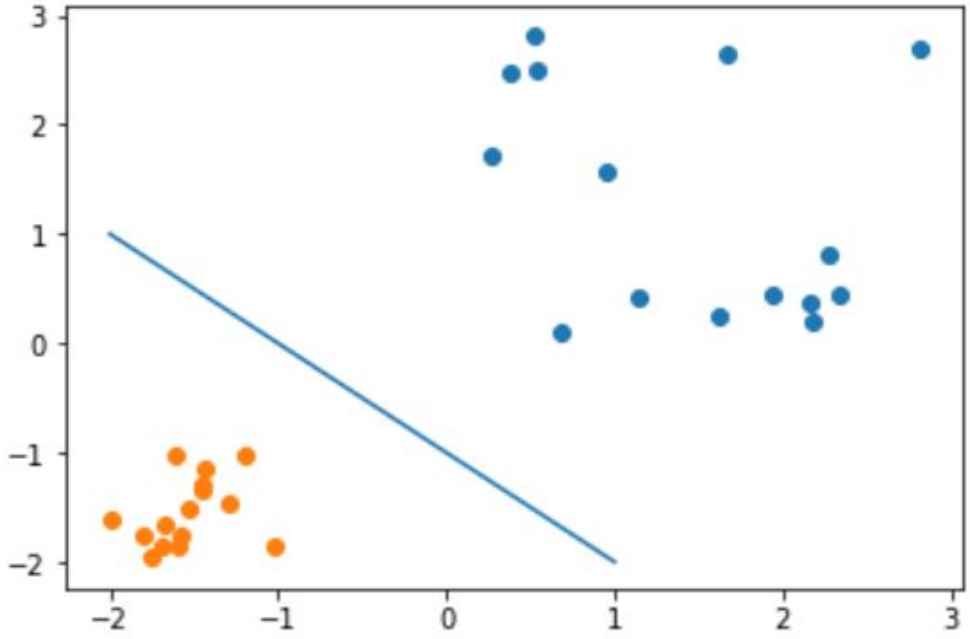
花的種類
0
0
1
1
2

線性分割

- X 有 150筆資料, 4 特徵
- Y 有 150筆對應答案, 0,1,2
- 我們要考慮做一個線性分割



線性分割



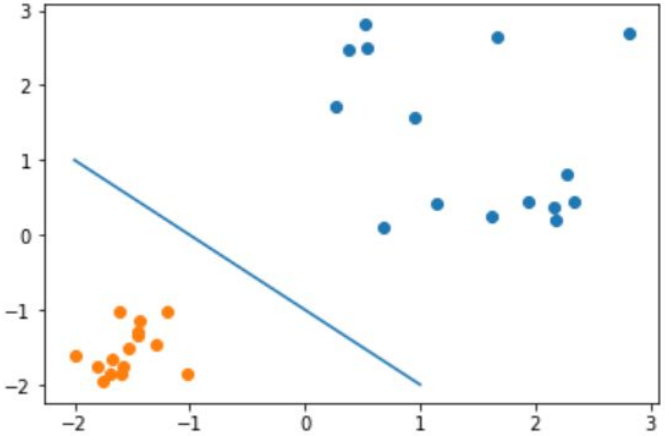
線性分割

$y = w^T x + b$

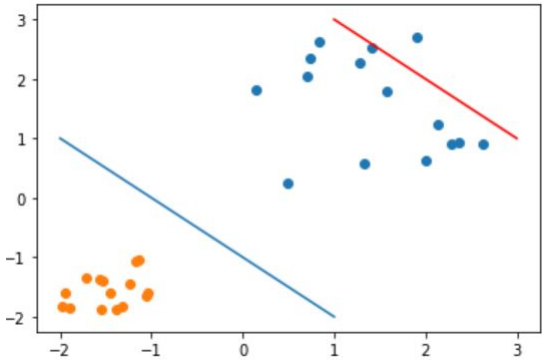
以上一頁的圖為例, 他是2維的資料

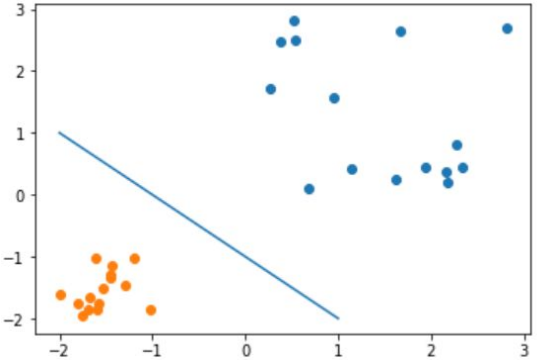
因此我們要討論的$w = (w_1, w_2); b$

然後讓左邊的資料帶進去會是<0, 右邊的>0



線性分割





解釋w和b

- w:權重
- b:調整項，因為我們的目標是要找到一個合適的切割，讓一邊的取值<0，一邊>0，因為權重不能保證讓整體很小，因此可以利用b來調整結果

怎麼決定?

~~靠電腦自己想辦法啊~~

- 討論誤差

  也就是我先給定一個w和b 會得到一個 y_{est}

  然後去算 $y-y_{est}$, 但是他還是向量

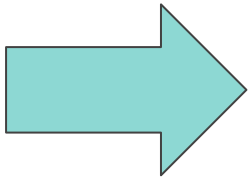
  因此我們要算他的向量大小, 也就是長度

怎麼決定?

- 去算 $||y-y_{est}||$讓他最小, 因此變成求極值問題
- 偏微分!!!
- 剩下的就是留給之後的阿澤來說

線性分割

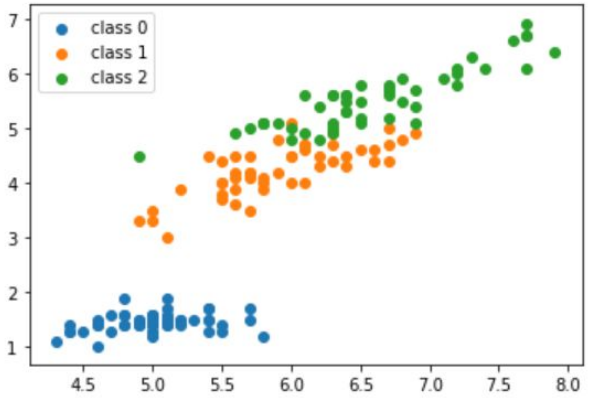
花萼長度	花萼寬度	花瓣長度	花瓣寬度
5.1	3.5	1.4	0.2
4.7	3.2	1.3	0.2
6.3	2.3	4.4	1.3
5.8	2.7	5.1	1.9
6.9	3.1	5.4	2.1



花的種類
0
0
1
1
2

線性分割

- X 有 150筆資料, 4 特徵
- Y 有 150筆對應答案, 0,1,2
- 我們要考慮做一個線性分割



Python

我們只考慮class_0跟class_1, 資料只取[0,2] column 的

```
from sklearn.linear_model import Perceptron
clf = Perceptron(tol=1e-3, random_state=0, verbose=1)
clf.fit(data, label)
clf.coef_
```

Support Vector Machine