機器學習基礎與演算法

Chapter 3　KNN
Chapter 4　Logistic regression

講師投影片Chapter3
講師投影片Chapter4
課程投影片
資料與程式碼
播放清單

「版權聲明頁」

課程內容

3. KNN

-KNN
[實作] KNN

4. Logistic regression

-Logistic regression
-Gradient ascent
-Evaluation (classification)
[實作] Logistic regression實作

Code 放在Hub中的course內

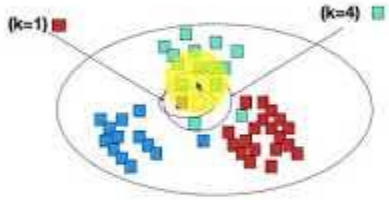
- 為維護課程資料，courses中的檔案皆為read-only，如需修改請cp至自身環境中
- 打開terminal，輸入

cp -r courses-tpe/Machine_Learning <存放至本機的名稱>

Chapter 3　K-nearest neighbors

03-1: KNN

K-nearest-Neighbor (KNN) Classifier



03-2: Overfitting and underfitting in KNN

How to select K?

15-nearest neighbors

1-nearest neighbors

01-1: Introduction to machine learning

Quiz (1/2)

- How to select *k* in KNN (assuming we are dealing with a binary classification problem)?
  - What if *k* is an even number?
  - What if *k* equals 1?
  - What if *k* equals the number of the training instances?
- How fast for model training/testing?

03-4: Answer

Quiz (1/2)

- How to select *k* in KNN (assuming we are dealing with a binary classification problem)?
  - What if *k* is an even number?
  - What if *k* equals 1?
  - What if *k* equals the number of the training instances?
- How fast for model training/testing?

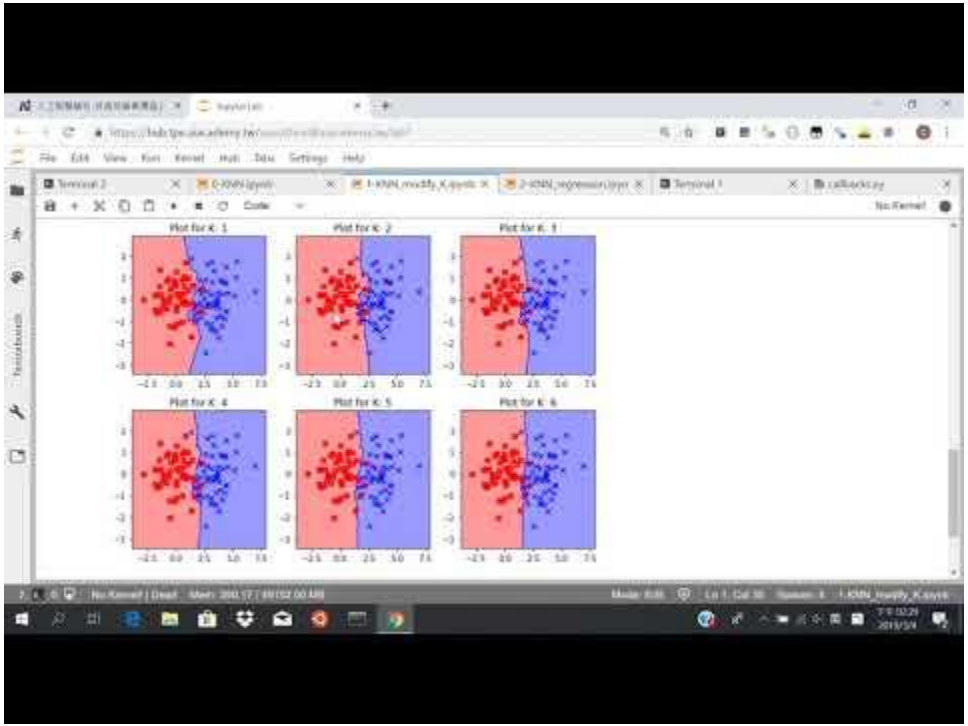
KNN (K Nearest Neighbor)

- K個最近的鄰居
  - 近朱者赤;近墨者黑
  - 孟母三遷
  - You are the average of the five people you spend the most time with.

KNN 優缺點

- 優點
  - 訓練速度快
  - 不用對模型進行泛化
  - 也可以做regression
- 缺點
  - 預測時速度慢，且佔用很大的記憶體空間
  - 需要對feature normalize
  - 不適合用在高維度資料

[實作課程] KNN

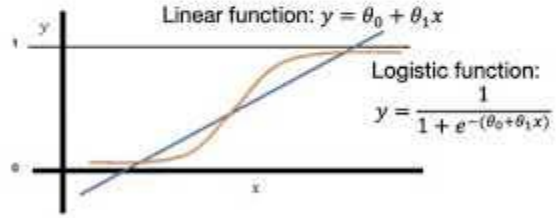


Chapter 4 Logistic regression

04-1: Introduction to logistic regression

Fitting an S-shaped function

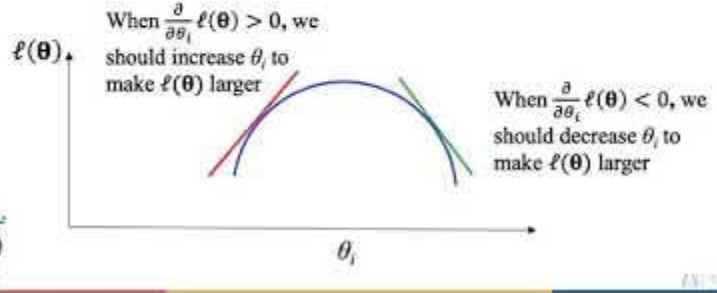
- If the fitting curve is S-shaped, the attributes with extreme (very large or very small) values will have little effect to the fitted curve



04-2: Gradient ascent

Gradient ascent

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \alpha\left(\nabla \ell(\boldsymbol{\theta})\right)^{T}$$



04-3: Quiz

Quiz

- What is logistic regression?
- Why do we usually maximize log-likelihood function (instead of likelihood function)?
- What is the cross entropy loss (for logistic regression)?

04-4: Answer

Quiz

- What is logistic regression?
- Why do we usually maximize log-likelihood function (instead of likelihood function)?
- What is the cross entropy loss (for logistic regression)?

04-5: Evaluation (classification)

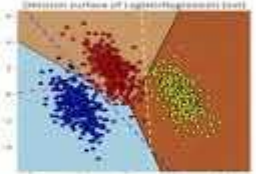
Precision

- Out of the instances I predicted as "positive", how many percentage of them are correct?
- $\text{Precision}(y, \hat{y}) = \frac{I(y_i=\hat{y}_i=1)}{I(\hat{y}_i)} = \frac{\text{TP}}{\text{TP+FP}}$
  - Assuming a binary classification task
- Commonly used to evaluate the quality of a search engine
  - How useful the search results are

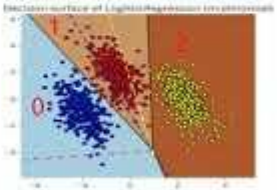
[實作課程] Logistic Regression

Logistic Regression in Scikit-Learn

- multi_class: 'ovr', 'multinominal'
  - 'ovr': 視多元分類為二元分類 (default)



  - 'multinomial': 會考慮到整體的分佈



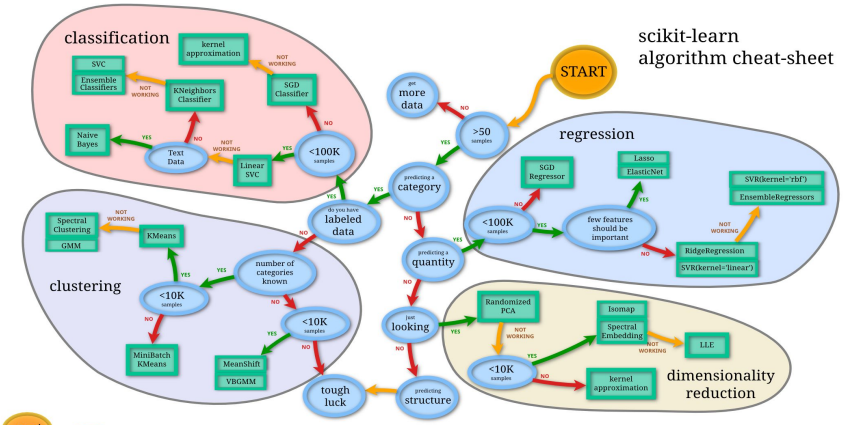
http://www.cnblogs.com/pinard/p/6035872.html

Logistic Regression

- 使用時機:
  - Label為非連續值(二元、多元分類問題)。
  - Features 和 Label之間不必有線性關係，因為Features的值會做non-linear的transform(sigmoid)。

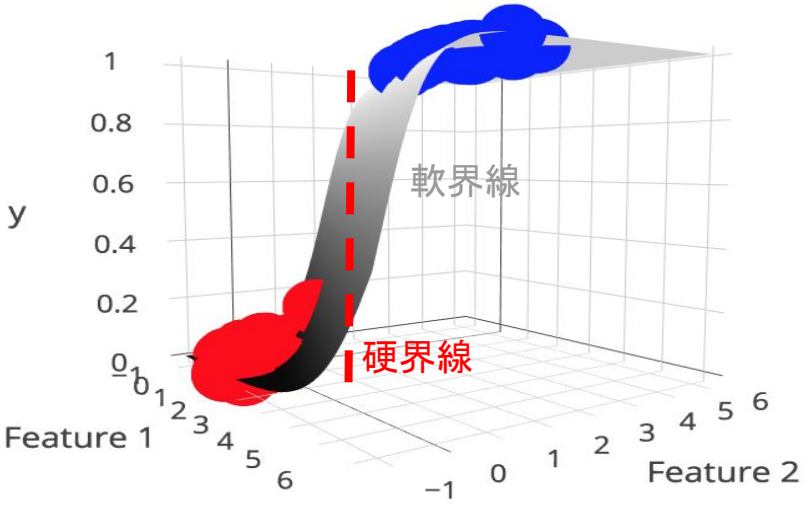
$$P(y|x) = \frac{1}{1 + e^{-yw^T x}}$$

- Prediction:0~1的機率值。



Geometric Meaning

- 利用線性關係算出分數 $yw^T x$
- 將分數代進sigmoid function轉換成機率函數
- 求解最大機率



$$P(y|x) = \frac{1}{1 + e^{-yw^T x}}$$

$$\max_{w} \prod_{i=1}^{k} P(y_i|x_i),$$

$$i = 1, \ldots, k \ (k \ training \ instances)$$

https://florianhartl.com/logistic-regression-geometric-intuition.html

Logistic Regression in Scikit-Learn

- Penalty: L1/ L2 (Lasso/Ridge)
- C: default=1

Logistic Regression: $\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{k} log(1 + e^{-y_i w^T x_i})$

Linear Regression: $Cost = \text{Prediction error} + \alpha \sum (\text{weights})^2$

Note: 和linear regression的alpha不一樣的是C值越大對weight的控制力越弱。

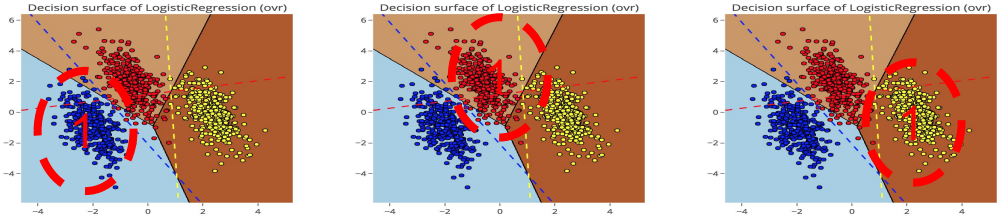
Case	Solver
Small dataset or L1 penalty	liblinear
Multinomial or large dataset	lbfgs, sag or newton-cg
Very Large dataset	sag

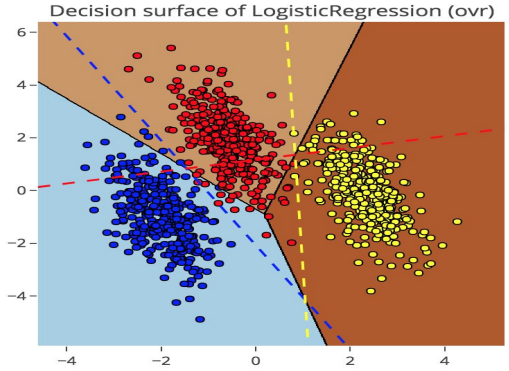
```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(penalty='l2', dual=False,
tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1
, class_weight=None, random_state=None, solver='liblinear'
, max_iter=100, multi_class='ovr', verbose=0, warm_start=False
 , n_jobs=1)
```

http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Logistic Regression in Scikit-Learn

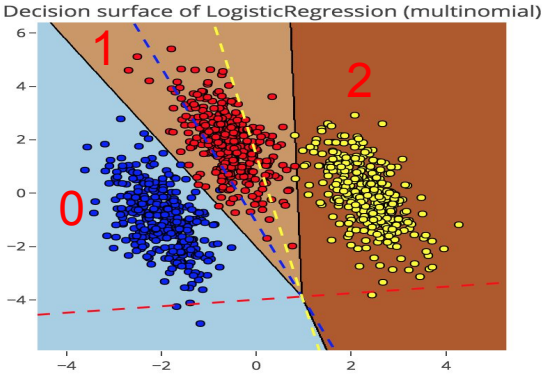
- multi_class: ‘ovr’, ‘multinominal’
  - ‘ovr’: 視多元分類為二元分類 (default)





  - ‘multinomial’: 會考慮到整體的分佈

0,1 分類
0,2 分類
1,2 分類



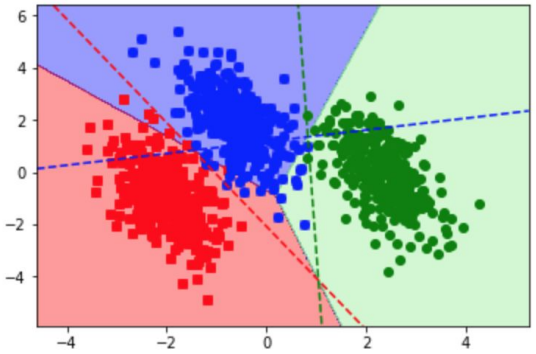
http://www.cnblogs.com/pinard/p/6035872.html

Logistic Regression Example

- 1-Logistic Regression Example
  - 調整C值觀察其對結果的影響
  - 如何解釋結果（Hint: 準確度、擬合outlier的程度）

```
#調整C值
plot_dict = {1:231,10:232,1e2:233,1e3:234,1e4:235,1e5:236}
for i in plot_dict:
    regression_example(plot_dict, i)
```

- 2-Logistic Regression Example
  - 多元分類問題（兩類以上）
  - 比較兩種multi_class的方法: ‘ovr’, ‘multinomial’



[實作課程] Logistic Regression

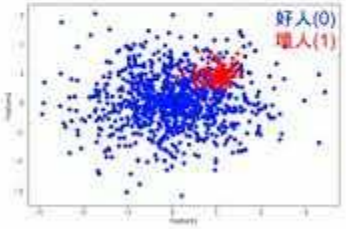
Imbalanced Data Prediction Exercise

動手時間

- 警察抓壞人當例子:
  - 範例中總共有10%被分類為1(壞人)的類別
  - 如果全猜好人準確率就有90%, 但是這樣的做法顯然不合理
- 提示:
  - 用Logistic Regression 看看準確率是否可以超過90%
  - 算出Precision & Recall, 看在這樣的狀況下抓出了多少比率壞人(Recall)？在預測這個人是壞人時你有多少信心(Precision)？
  - 計算F1_score

Hint:可以利用confusion matrix自行算出

  - 如何找到最好的F1_score？



Short Summary

- 用 Logistic Regression 解二元或多元分類問題
- 利用 L1&L2 Regularization 控制模型複雜度

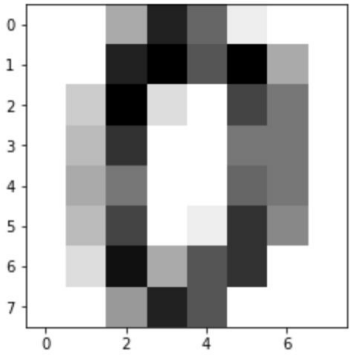
模型	Linear Regression	Logistic Regression
用途	連續值預測	分類問題預測
複雜度控制(L1,L2)	Alpha 越大控制力越強	C 越大控制力越弱

Logistic Regression Exercise動手時間

- 手寫影像為8x8的灰階影像
- 將每個影像展平成64個features的dataset來做預測
- 學習點
  - 調整C值及multi_class、solver
  - 用accuracy來評估模型結果
  - 利用confusion matrix來分析哪些數字較容易分辨哪些較不容易

```
plt.imshow(digits.images[0], cmap=plt.cm.binary, interpolation='nearest')
```

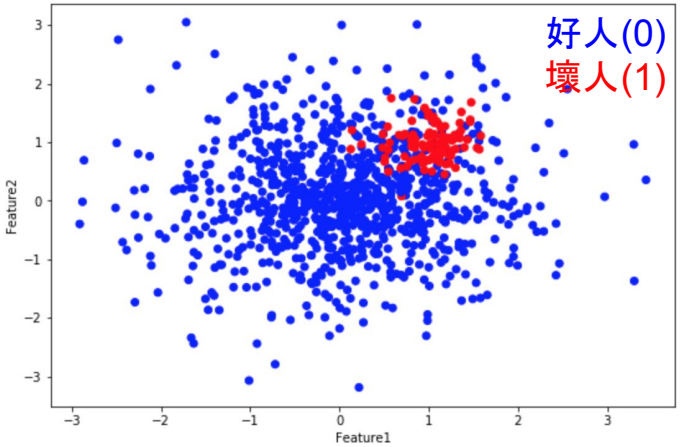
```
<matplotlib.image.AxesImage at 0x116b34e80>
```




```
array([[ 0.,  0.,  5., 13.,  9.,  1.,  0.,  0.],
       [ 0.,  0., 13., 15., 10., 15.,  5.,  0.],
       [ 0.,  3., 15.,  2.,  0., 11.,  8.,  0.],
       [ 0.,  4., 12.,  0.,  0.,  8.,  8.,  0.],
       [ 0.,  5.,  8.,  0.,  0.,  9.,  8.,  0.],
       [ 0.,  4., 11.,  0.,  1., 12.,  7.,  0.],
       [ 0.,  2., 14.,  5., 10., 12.,  0.,  0.],
       [ 0.,  0.,  6., 13., 10.,  0.,  0.,  0.]])
```

Imbalanced Data Prediction Exercise 動手時間

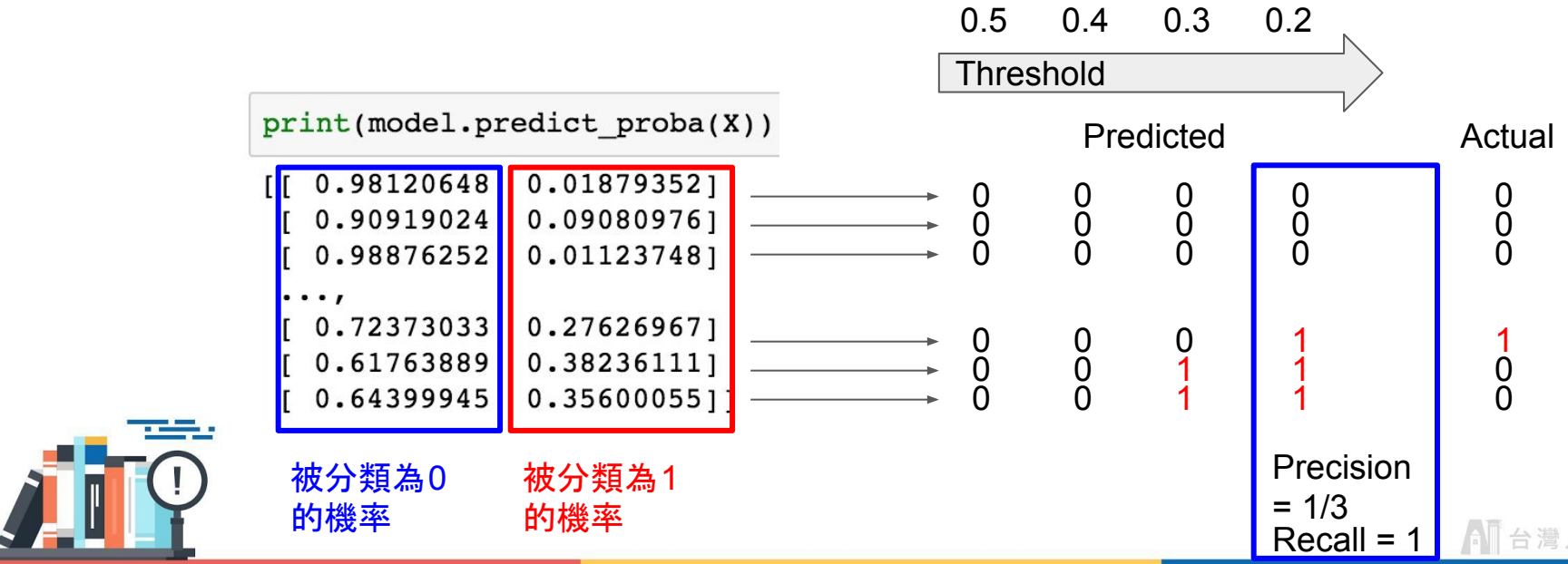
- 警察抓壞人當例子:
  - 範例中總共有10%被分類為1(壞人)的類別
  - 如果全猜好人準確率就有90%, 但是這樣的做法顯然不合理
- 提示:
  - 用Logistic Regression 看看準確率是否可以超過90%
  - 算出Precision & Recall, 看在這樣的狀況下抓出了多少比率壞人(Recall)? 在預測這個人是壞人時你有多少信心(Precision)?
  - 計算F1_score
  - Hint:可以利用confusion matrix自行算出
  - 如何找到最好的F1_score?



[實作課程] F1-Score and AUC ROC

Threshold

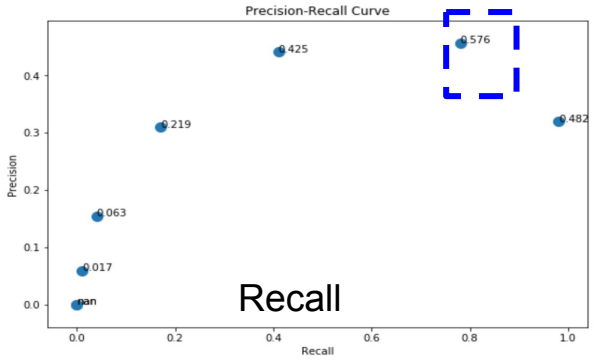
- Logistic Regression和SVM實際是output分類機率。
- 如果利用model.predict會以0.5當作threshold來給出分類預測(二元分類)。
- 針對不同的目的可以調整不同的threshold來得到要的recall和precision。

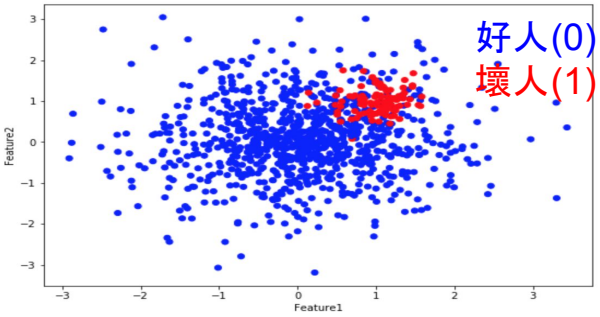


F1_Score

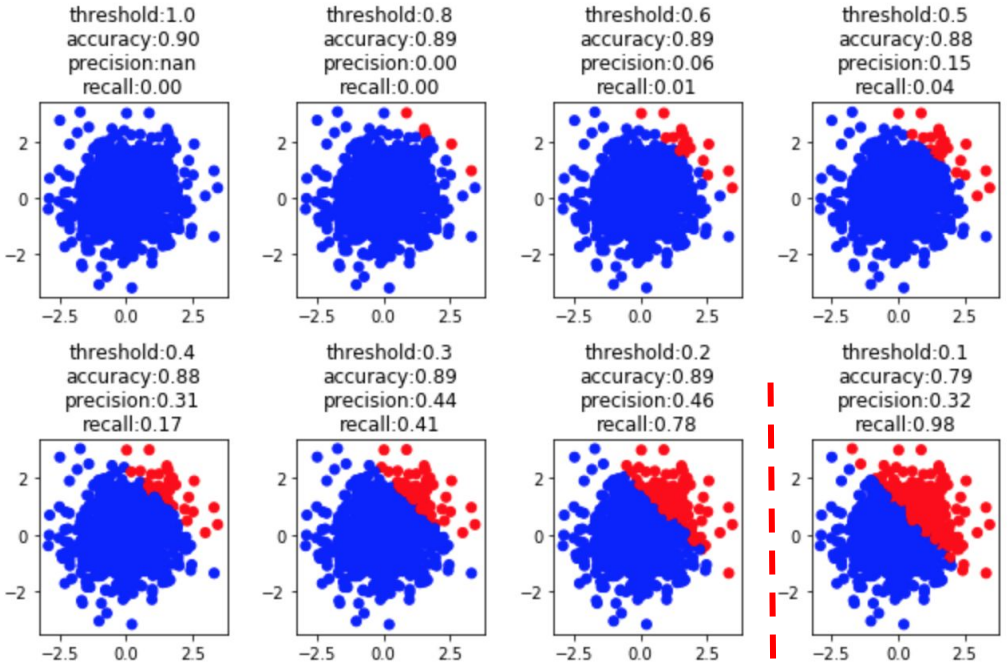
- Threshold調越低就可以抓出越多壞人(recall越高)但就有可能會錯殺越多好人。
- 調整Threshold找到最好的F1_Score。

Precision

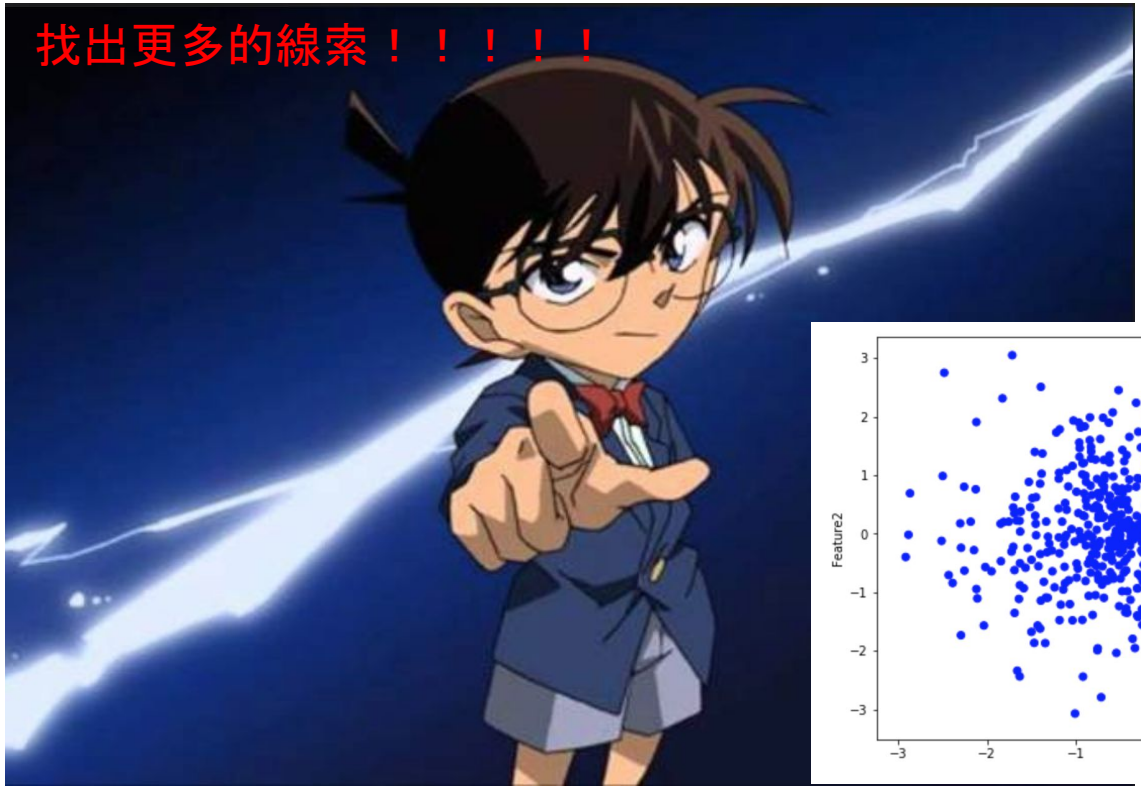


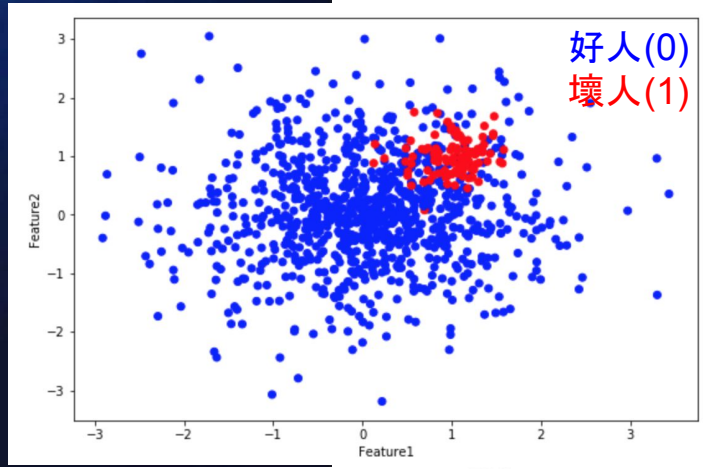


Model Prediction with vaired threshold



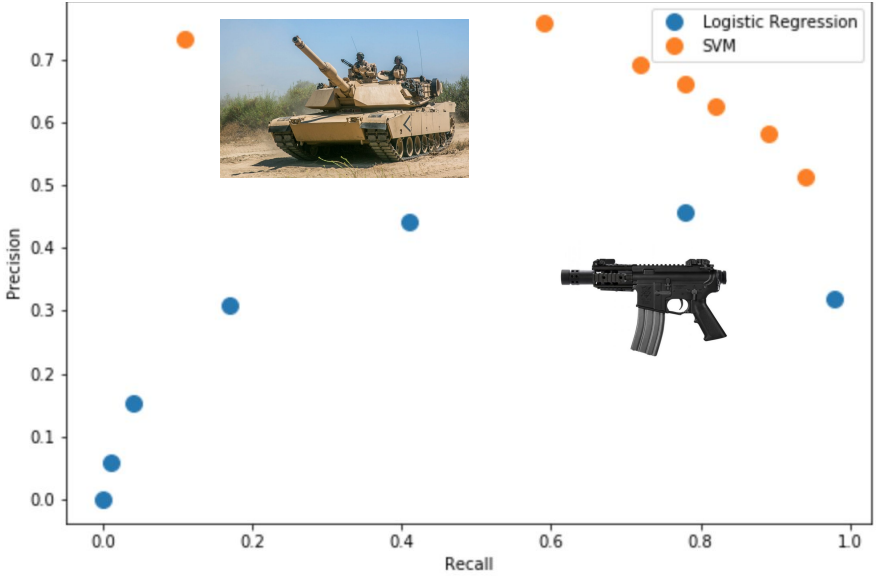
What’s Next?

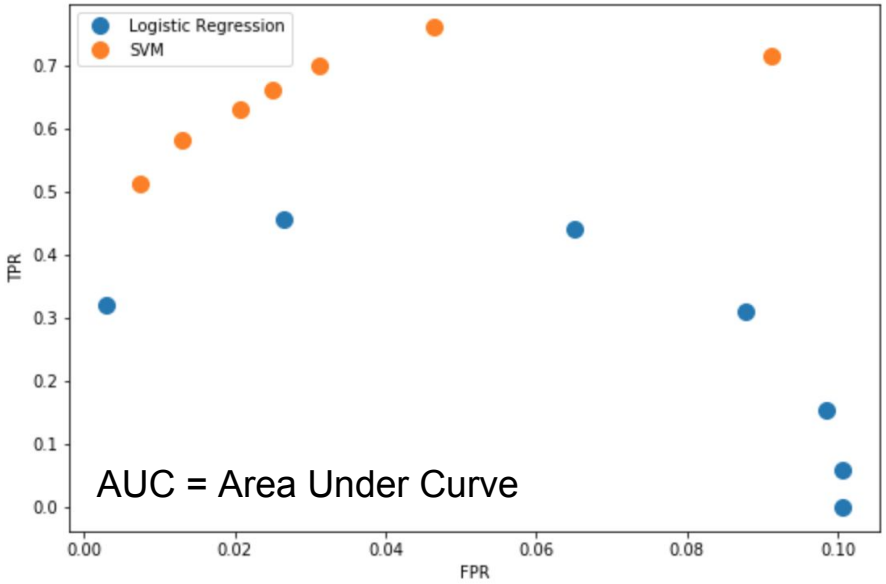




What's Next?

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$





Precision VS. Recall

ROC: TPR VS. FPR

When Precision-Recall and When ROC?

Confusion Matrix	Predicted 0	Predicted 1
Actual 0	True Negative(TN)	False Positive(FP)
Actual 1	False Negative(FN)	True Positive(TP)

偵測問題

Precision = TP/(TP+FP)

TPR(True Positive Rate) = Recall
TP/(FN+TP)

Confusion Matrix	Predicted 0	Predicted 1
Actual 0	True Negative(TN)	False Positive(FP)
Actual 1	False Negative(FN)	True Positive(TP)

分類問題

TPR(True Positive Rate) = Recall
TP/(FN+TP)

FPR(False Positive Rate) =
FP/(TN+FP)

https://www.kaggle.com/general/7517

Classification　動手時間

- Exercise- Classification 

  - 選擇不同的kernel，並調整參數。
  - 利用Accuracy評估預測結果
  - 利用F1 score和confusion matrix評估預測結果