



巨匠線上真人

Python 資料科學應用開發

# 第十五堂：整體學習 ( Ensemble Learning )

## 同學，歡迎你參加本課程

- ☑ 請關閉你的FB、Line等溝通工具，以免影響你上課。
- ☑ 考量頻寬、雜音，請預設關閉攝影機、麥克風，若有需要再打開。
- ☑ 隨時準備好，老師會呼叫你的名字進行互動，鼓勵用麥克風提問。
- ☑ 如果有緊急事情，你必需離開線上教室，請用聊天室私訊給老師，以免老師癡癡呼喚你的名字。
- ☑ 軟體安裝請在上課前安裝完成，未完成的同學，請盡快進行安裝。

# 課程檔案下載

The screenshot displays the homepage of the Juei Computer Online Live website. The header features navigation links: 巨匠電腦線上真人 (Juei Computer Online Live), 開課查詢 (Class Inquiry), 免費體驗專區 (Free Experience Area), 課程總覽 (Course Overview), 專業師資 (Professional Faculty), 學員專區 (Student Area), 講師專區 (Instructor Area), and 最新消息 (Latest News). Social media icons for 360, Facebook, and YouTube are also present. A user greeting '您好!' and a '登出' (Logout) button are shown. The main banner area contains the text '程式語言好難學?' (Programming Language is so hard to learn?), '那是因為你還沒學過Python!' (That's because you haven't learned Python!), and '(線上老師 LIVE 直播教學 · 搶先看)' (Online Teacher LIVE Streaming Teaching · Sneak Peek). A dropdown menu is open, listing various services, with '課程檔案下載' (Course Archive Download) highlighted. The background features a blue and orange abstract design with circuit-like patterns.

巨匠電腦線上真人 開課查詢 免費體驗專區 課程總覽 專業師資 學員專區 講師專區 最新消息

360 f YouTube

您好! 登出

點數卡產品兌換  
APCS檢測專區  
公告專區  
我的課表  
IT真人課程劃位  
電腦分校課程劃位  
外語真人課程劃位  
美語分校課程劃位  
取消劃位  
**課程檔案下載**  
上課權益查詢  
教學平台測試  
學習諮詢  
常見問題  
個資維護  
忘記密碼  
登出

課程檔案下載

程式語言好難學?  
那是因為  
你還沒學過Python!  
(線上老師 LIVE 直播教學 · 搶先看)

巨匠電腦真人課程

# ZOOM 學員操作說明

The screenshot shows the Zoom interface with several callouts:

- 5 查看選項/共同註記/筆 (連連看)**: Points to the '共同註記' (Annotate) button in the top toolbar.
- 2 共享螢幕 (指導演練; 點評作品)**: Points to the '共享螢幕' (Share Screen) button in the bottom toolbar. A text box explains: '老師須先停止共享螢幕 才能請學生共享螢幕' (The teacher must first stop sharing the screen before asking the student to share the screen).
- 1 聊天**: Points to the '聊天' (Chat) button in the bottom toolbar.
- 3 與會者/舉手**: Points to the '與會者' (Participants) button in the bottom toolbar. A text box explains: '舉手' (Raise hand).
- 4 解除靜音**: Points to the '解除靜音' (Unmute) button in the bottom toolbar.

The interface also shows a '查看選項' (View Options) menu with '原始大小' (Original Size), '請求遠端控制' (Request Remote Control), '共同註記' (Annotate), and '退出全螢幕' (Exit Full Screen). The bottom toolbar includes buttons for '解除靜音', '啟動視訊', '邀請', '與會者', '共享螢幕', '聊天', and '錄影'.

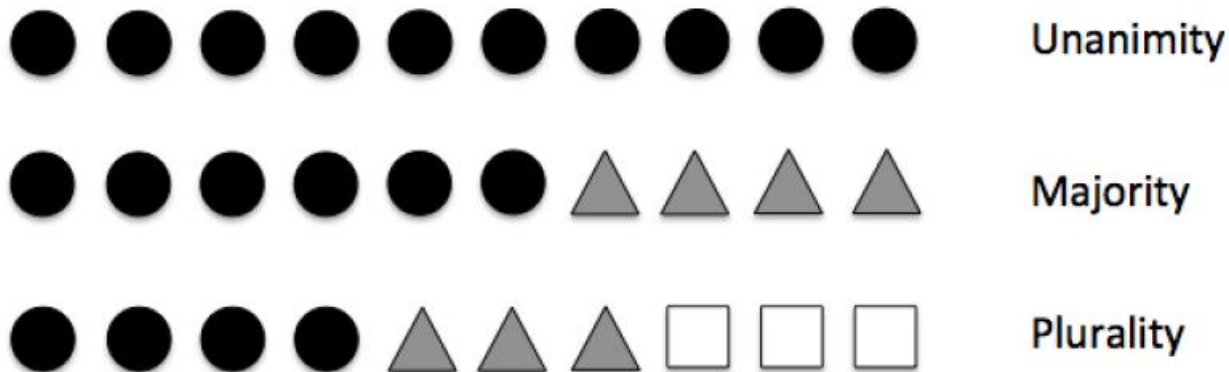
# 課程內容

## 整體學習方法

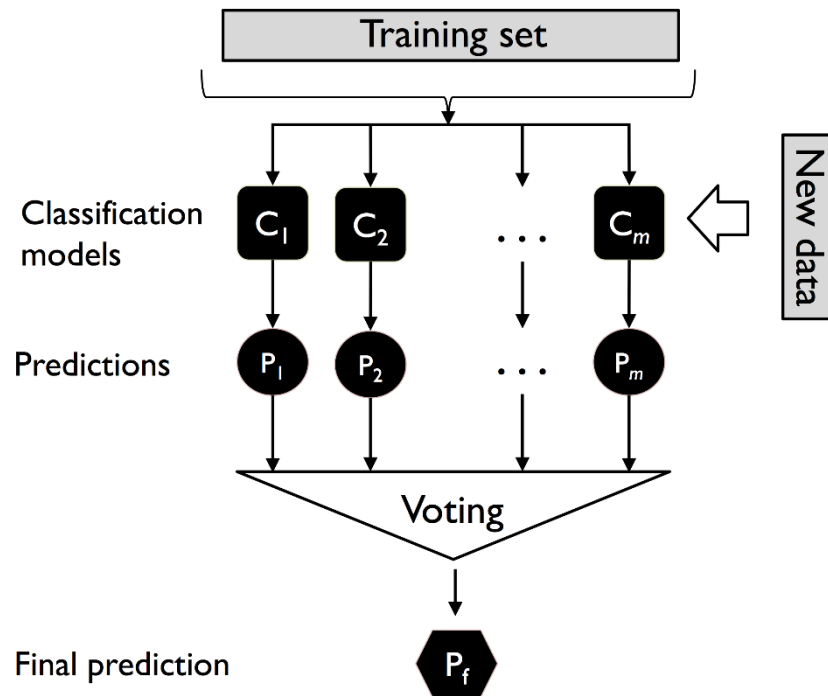
- 多數決 ( Majority Voting )
- 裝袋法 ( Bagging )
- 強化法 ( Boosting )
- AdaBoost

# 整體學習 ( Ensemble Learning )

- ◆ 整體分類器 ( Ensemble Classifier )
  - ◇ 同時使用多個分類法
  - ◇ 使用同一個分類法，但資料拆成很多份，個別訓練
- ◆ 可採用『多數決』 ( Majority Voting )、『最高票』 ( Plurality )



# 流程



# 提升準確率

◆ 假設所有的模型均為相同的 Error Rate

◆ 則整體學習的 Error Rate

$$P(y \geq k) = \sum_k^n \binom{n}{k} \epsilon^k (1 - \epsilon)^{n-k} = \epsilon_{\text{ensemble}}$$

◆ 例如：共有11個模型 Error Rate 均為0.25

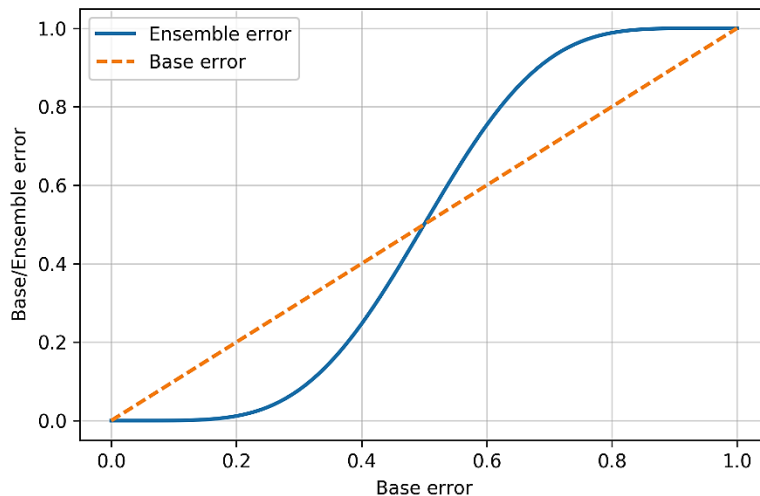
$$P(y \geq k) = \sum_{k=6}^{11} \binom{11}{k} 0.25^k (1 - 0.25)^{11-k} = 0.034$$

◆ 整體學習的 Error Rate < 個別模型



# 觀察

- ◆ 依據上面推論，實作不同的 Error Rate 下，整體學習的表現
- ◆ 程式：ErrorRate.py



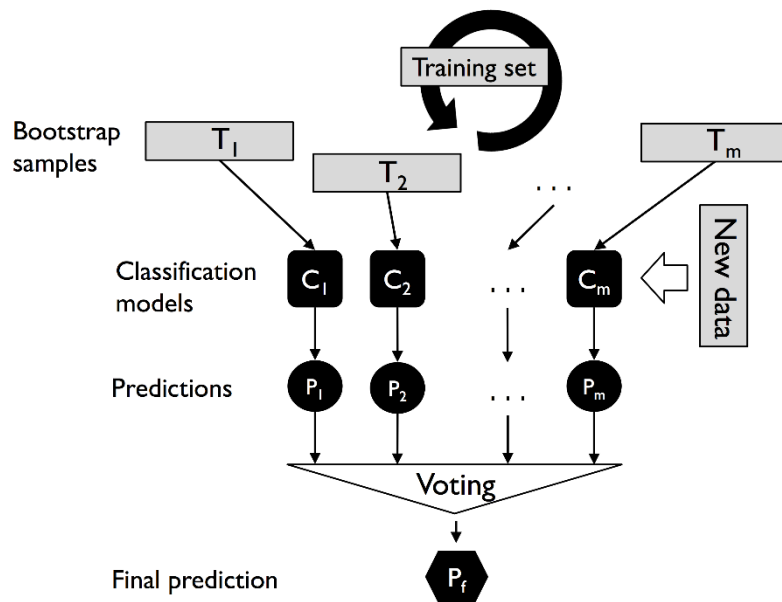
# 簡單範例

- ◆ 使用「羅吉斯迴歸、隨機森林、貝氏分類法」進行整體學習

```
clf1 = LogisticRegression(random_state=1)
clf2 = RandomForestClassifier(random_state=1)
clf3 = GaussianNB()
estimators=[('lr', clf1), ('rf', clf2), ('gnb', clf3)]
ens_clf = VotingClassifier(estimators)
ens_clf = eclf1.fit(X, y)
```

# 裝袋法 ( Bagging )

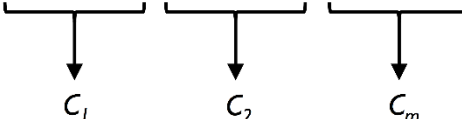
- ◆ 採『放回式抽樣』 ( resampling with replacement )



# 作法

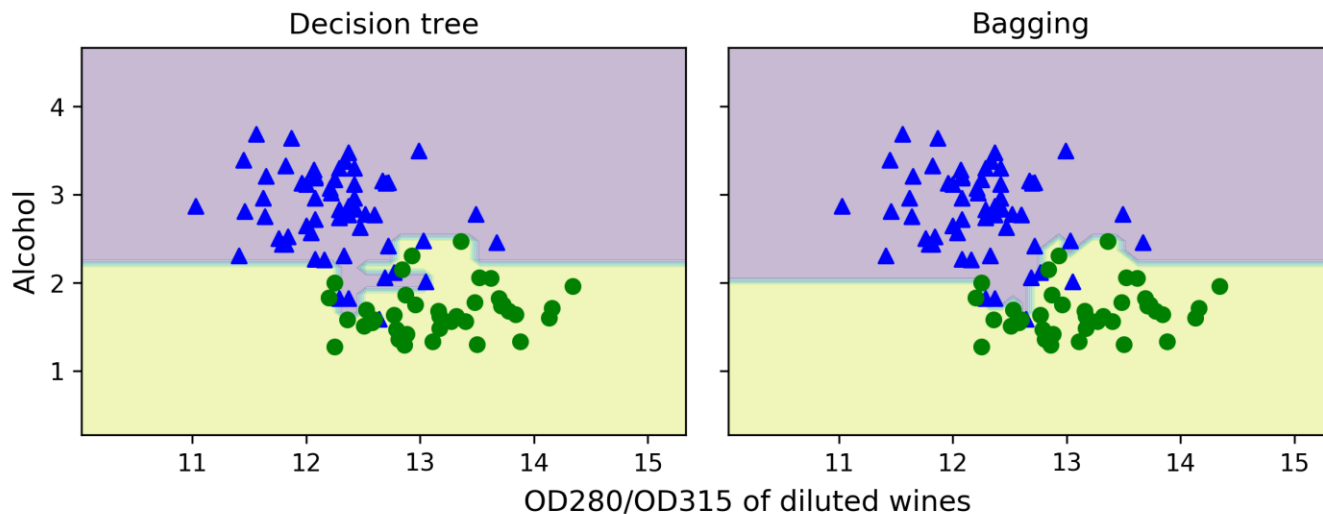
- ◆ 每次抽樣7個樣本裝袋 ( Bag )，由於抽出後會放回，所以，每個袋中的樣本可能重複
- ◆ 隨機森林就是裝袋法的一個特例
- ◆ Scikit-learn 提供 BaggingClassifier 函數

Sample indices	Bagging round 1	Bagging round 2	...
1	2	7	...
2	2	3	...
3	1	2	...
4	3	1	...
5	7	1	...
6	2	7	...
7	4	7	...



# 實作

- ◆ 使用 Wine 資料集，實作整體學習
- ◆ 程式：ch07.ipynb 第29 ~ 33格
- ◆ 請比較準確率



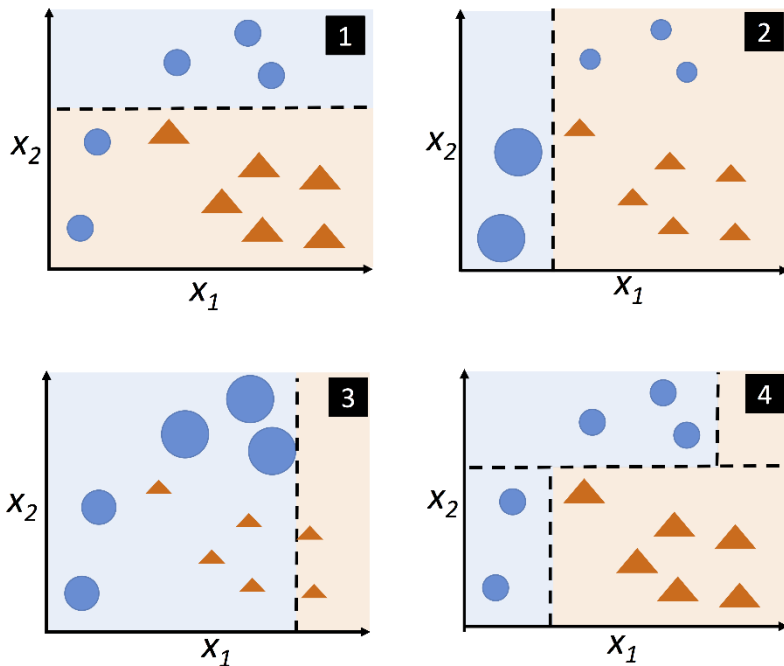
# Quiz (1)

◆ 執行 Bagging.ipynb，並回答下列問題：

1. What is the accuracy of the model with 40 trees?
2. What is the accuracy of the model with reduced feature sets?
3. What is the best value of max\_features?
4. What is the AUC of the model?

# 強化法 ( Boosting )

- ◆ 把重心放在『很難分類的樣本』，從『誤判的訓練樣本』中學習



# 作法

1. 從全部樣本  $D$  中，以不放回抽樣的方式，抽出樣本  $d1$ ，放入弱學習器  $C1$  中訓練
2. 抽出樣本  $d2$ ，並加入50%的  $C1$  分類錯誤的樣本，放入弱學習器  $C2$  中訓練
3. 以  $C1$ 、 $C2$  分類錯誤的樣本為  $d3$ ，放入弱學習器  $C3$  中訓練
4. 結合  $C1$ 、 $C2$ 、 $C3$  弱學習器，以多數決作預測

※ 個別的模型稱為弱學習器 ( Weak Learner )

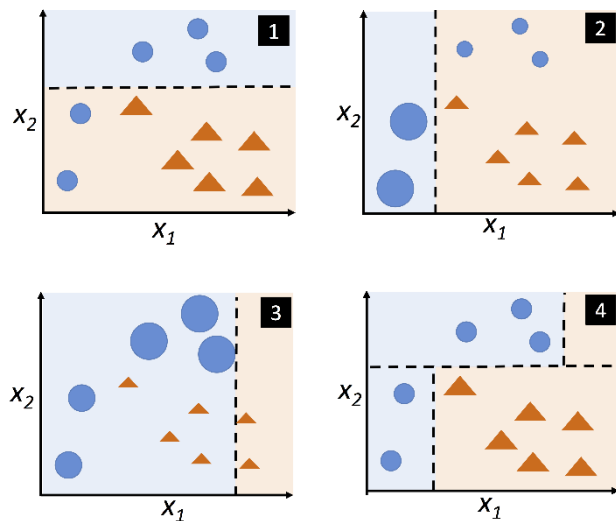


# AdaBoost ( Adaptive Boosting )

◆ 與原始作法的差異：

◆ 訓練採全部樣本

◆ 分類錯誤的樣本給定較大的加權值，重新訓練



# 步驟

1. 將向量  $\mathbf{w}$  設成相等加權值，且  $\sum_i w_i = 1$
  2. 對第  $m$  次強化中的  $j$ ，做以下的步驟：
    - a. 訓練一個弱學習器： $C_j = \text{train}(X, y, \mathbf{w})$
    - b. 預測類別標籤： $\hat{y} = \text{predict}(C_j, X)$
    - c. 計算加權錯誤率： $\varepsilon = \mathbf{w} \cdot (\hat{y} \neq y)$
    - d. 計算係數： $\alpha_j = 0.5 \log \frac{1-\varepsilon}{\varepsilon}$
    - e. 更新加權： $\mathbf{w} := \mathbf{w} \times \exp(-\alpha_j \times \hat{y} \times y)$
    - f. 常態化加權，使得加總為 1： $\mathbf{w} := \mathbf{w} / \sum_i w_i$
  3. 計算最後的預測： $\hat{y} = \left( \sum_{j=1}^m (\alpha_j \times \text{predict}(C_j, X)) > 0 \right)$
1. Set weight vector  $\mathbf{w}$  to uniform weights where  $\sum_i w_i = 1$
  2. For  $j$  in  $m$  boosting rounds, do the following:
  3. Train a weighted weak learner:  $C_j = \text{train}(X, y, \mathbf{w})$ .
  4. Predict class labels:  $\hat{y} = \text{predict}(C_j, X)$ .
  5. Compute weighted error rate:  $\varepsilon = \mathbf{w} \cdot (\hat{y} \neq y)$ .
  6. Compute coefficient:  $\alpha_j = 0.5 \log \frac{1-\varepsilon}{\varepsilon}$ .
  7. Update weights:  $\mathbf{w} := \mathbf{w} \times \exp(-\alpha_j \times \hat{y} \times y)$ .
  8. Normalize weights to sum to 1:  $\mathbf{w} := \mathbf{w} / \sum_i w_i$ .
  9. Compute final prediction:  $\hat{y} = \left( \sum_{j=1}^m (\alpha_j \times \text{predict}(C_j, X)) > 0 \right)$ .

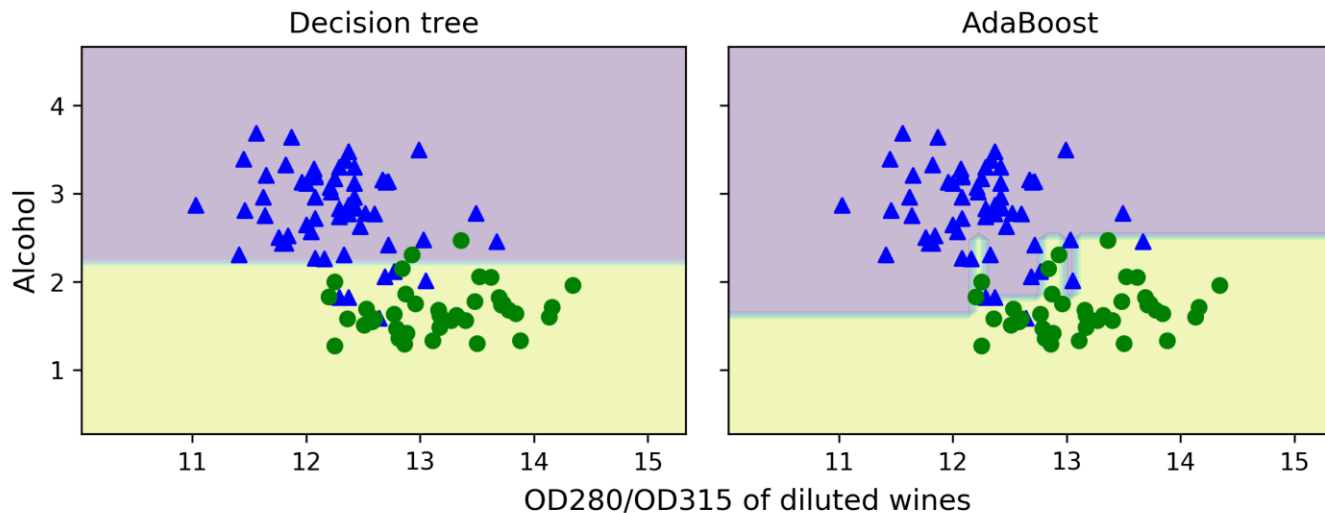
# 範例

Sample indices	x	y	Weights	$\hat{y}(x \leq 3.0)?$	Correct?	Updated weights
1	1.0	1	0.1	1	Yes	0.072
2	2.0	1	0.1	1	Yes	0.072
3	3.0	1	0.1	1	Yes	0.072
4	4.0	-1	0.1	-1	Yes	0.072
5	5.0	-1	0.1	-1	Yes	0.072
6	6.0	-1	0.1	-1	Yes	0.072
7	7.0	1	0.1	-1	No	0.167
8	8.0	1	0.1	-1	No	0.167
9	9.0	1	0.1	-1	No	0.167
10	10.0	-1	0.1	-1	Yes	0.072

※ 分類錯誤的樣本權值加大

# 實作

- ◆ Scikit-learn 提供 AdaBoostClassifier 函數
- ◆ 程式：ch07.ipynb 第 36 ~ 38 格
- ◆ 請比較準確率



# Quiz (2)

◆ 執行 Boosting.ipynb，並回答下列問題：

1. What is the accuracy of the model with reduced feature sets?
2. What is the best value of learning\_rate?
3. What is the best value of learning\_rate?
4. What is the AUC of the model?

# Quiz (3)

◆ 執行 NeuralNetworks.ipynb，並回答下列問題：

1. What is the accuracy of the model with (100,100) hidden\_layer\_size?
2. What is the best value of beta\_1?

# Quiz (4)

◆ 執行 SVM.ipynb，並回答下列問題：

1. What is the accuracy of the model with nonlinear SVM?
2. What is the best value of gamma?
3. What is the AUC of the model?

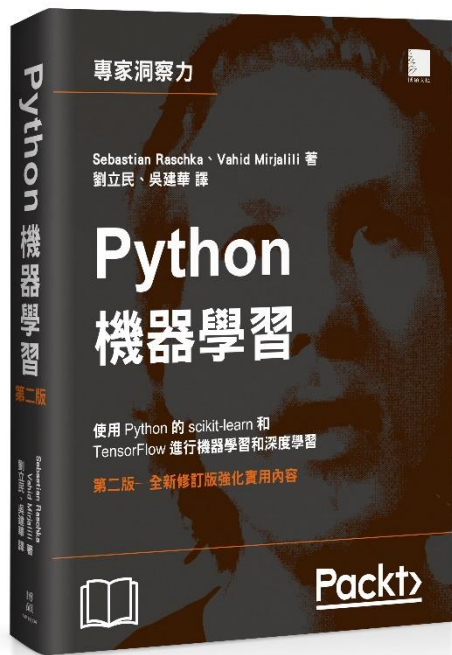
# Quiz (5)

◆ 執行 `NaiveBayes.ipynb`，並回答下列問題：

1. What is the accuracy of the model with Gaussian Naive Bayes?
2. What is the AUC of the Gaussian Naive Bayes model?
3. What is the AUC of the Bernoulli Naive Bayes model?



# 參考用書



- ◆ 書名：Python機器學習（第二版）

<http://www.drmaster.com.tw/bookinfo.asp?BookID=MP11804>

- ◆ 作者：Sebastian Raschka, Vahid Mirjalili ISBN
- ◆ 譯者：劉立民、吳建華
- ◆ 出版社：博碩

# 問卷

<http://www.pcschoolonline.com.tw>

開課查詢

免費體驗專區

課程總覽

專業師

1

學員專區

講師專區



➤ 課程檔案下載：

學員的「上課教材」，下載檔案為壓縮檔 ([解壓縮操作步驟](#))。  
如無法觀看上課教材，請安裝 [PDF閱讀軟體](#)。

公告專區

我的課表

課程劃位

取消劃位

2

課程檔案下載

自107年1月1日起，課程錄影檔由180天改為365天(含)內無限次觀看 (上課隔日18:00起)。

問  
卷

上課日期	課程名稱	課程節次	教材下載		
2017/12/27 2000 ~ 2200	線上真人-ZBrush 3D動畫造型設計	18	<a href="#">上課教材</a>	<a href="#">錄影檔</a>	<a href="#">課堂問卷</a>
2017/12/20 2000 ~ 2200	線上真人-ZBrush 3D動畫造型設計	17	<a href="#">上課教材</a>	<a href="#">錄影檔</a>	
2017/12/18 2000 ~ 2200	線上真人-ZBrush 3D動畫造型設計	16	<a href="#">上課教材</a>	<a href="#">錄影檔</a>	



巨匠線上真人

[www.pcschoolonline.com.tw](http://www.pcschoolonline.com.tw)