



人工智慧概論

Introduction to AI

第4章 機器學習

蘇維宗 (Wei-Tsung Su)
suwt@au.edu.tw
564D



歷史版本

版本	說明	日期	負責人
v1.0	初版	2020/02/14	蘇維宗
v1.1	加入監督式學習(迴歸、分類與幾個演算法)	2020/05/15	蘇維宗



課程目標

- 為何需要機器學習?
- 監督式學習概要
 - 迴歸演算法
 - 分類演算法
- 非監督式學習概要

為何需要機器學習?



機器學習

如果代理人可以根據**觀察環境**來**改善未來效能**即為學習(learning)。

為何需要讓代理人能夠學習？

- 無法給予所有的狀況(例如, 可以走**任何迷宮**的機器人)
- 無法預測可能的改變(例如, 可以預測明天股票市場的程式)
- 不知道如何撰寫程式(例如, 可以辨識人臉的系統)



在使用機器學習之前...

要改善甚麼? (**待解的問題**)

已經知道甚麼? (**已知的資料**) / 機器學習是一種**資料分析技術**

如何表示已知道的資料與要改善的目標?

機器學習可以得到的回饋?



機器學習可以得到的回饋

非監督式學習(unsupervised learning)

透過輸入來學習

監督式學習(supervised learning)

透過觀察輸入/輸出來學習一個函數 $f(\text{輸入}) = \text{輸出}$

強化式學習(reinforcement learning)

透過一連串的獎勵/懲罰來進行學習

監督式學習概要



監督式學習的步驟

訓練資料集(training set)

準備多個經過標籤(labeled)的資料

$(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$ 其中 $y = f(x)$

假說函數(hypothesis) / 模型(model)

透過機器學習找到一個假說函數 h 其結果近似於真實函數 f

測試資料集(test set)

透過與訓練資料集不同的測試資料集來驗證假說函數的正確性

假說函數 v.s. 真實函數

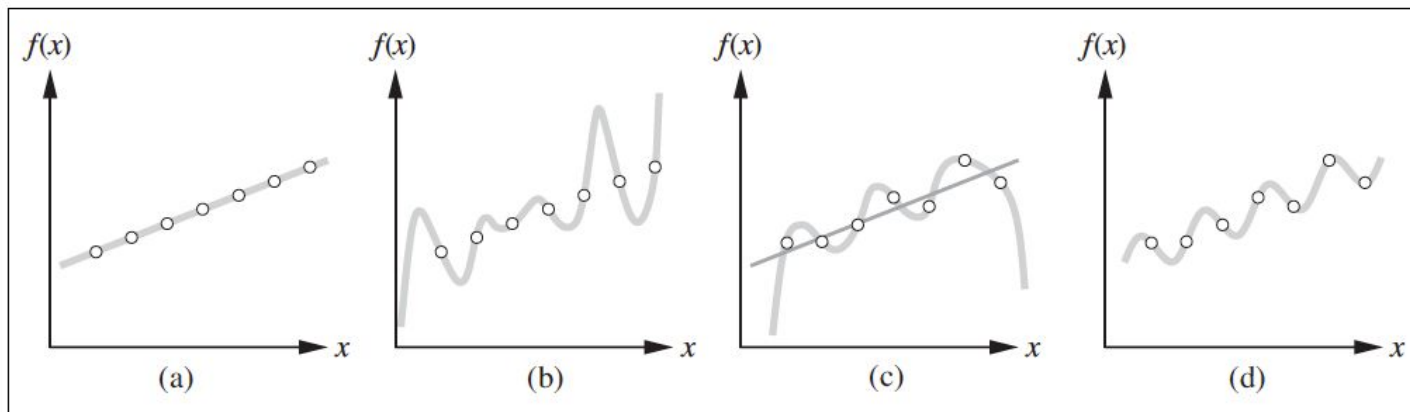


Figure 18.1 (a) Example $(x, f(x))$ pairs and a consistent, linear hypothesis. (b) A consistent, degree-7 polynomial hypothesis for the same data set. (c) A different data set, which admits an exact degree-6 polynomial fit or an approximate linear fit. (d) A simple, exact sinusoidal fit to the same data set.



經典的監督式學習演算法

迴歸(Regression)

主要用來預測連續數值(薪水預測、房價預測等)。例如

- 簡單線性迴歸
- 多項式迴歸
- ...

分類(Classification)

主要用來預測離散數值(預測人種、預測水果種類等)。例如

- 決策樹
- 隨機森林
- ...

監督式學習

迴歸(Regression)



迴歸

根據此資料集找到一個假說函數 h ，使得

$$h(\text{年資}) = \text{薪資}$$

自變數(Explained Variables)

應變數(Dependable Variables)

這樣就可以根據年資來預測薪資！

備註：因為薪資分佈是連續的，所以可以考慮使用迴歸演算法來解決這個問題。

序號	年資 (Feature)	薪資 (Label)
1	1	30,000
2	2	25,000
3	3	35,000
4	3.5	32,000
5	4	40,000
6	4.5	32,000
7	6	35,000
8	7	40,000
9	8	50,000

迴歸(續)

訓練資料集(training set)

例如, 取第1, 2, 4, 5, 7, 8等6筆資料

假說函數(hypothesis) / 模型(Model)

運用迴歸演算法訓練 $h(\text{年資}) = \text{薪資}$

測試資料集(test set)

例如, 取第3, 6, 9等3筆資料

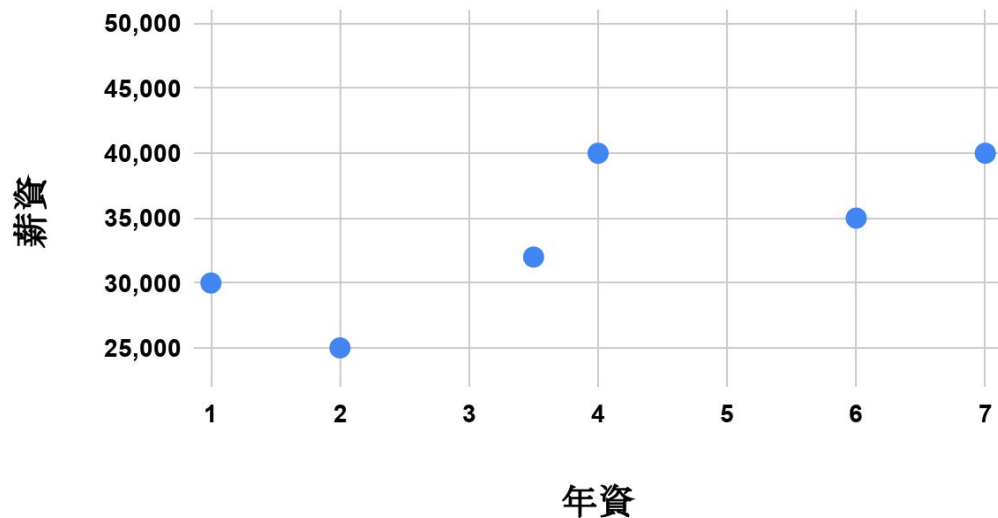
序號	年資 (Feature)	薪資 (Label)
1	1	30,000
2	2	25,000
3	3	35,000
4	3.5	32,000
5	4	40,000
6	4.5	32,000
7	6	35,000
8	7	40,000
9	8	50,000

迴歸(續)

還有那些問題需要解決

- 迴歸模型長怎樣?
 - 迴歸演算法?
- 迴歸模型預測的準嗎?
- ...

薪資年薪分布圖

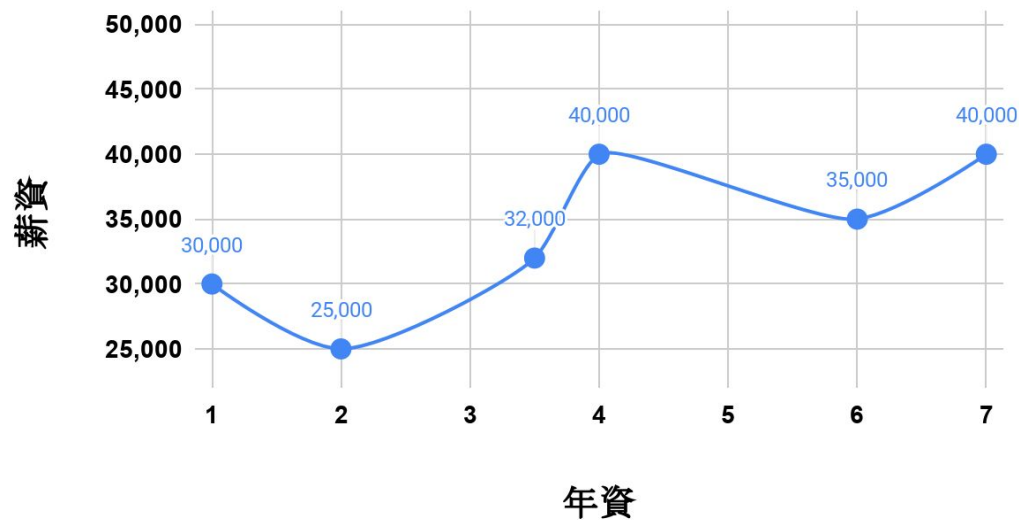


迴歸模型

平滑曲線可能造成過度適配
(Overfitting)問題。

過度適配問題的原因是模型
過度依賴訓練資料集。

薪資年薪分布圖

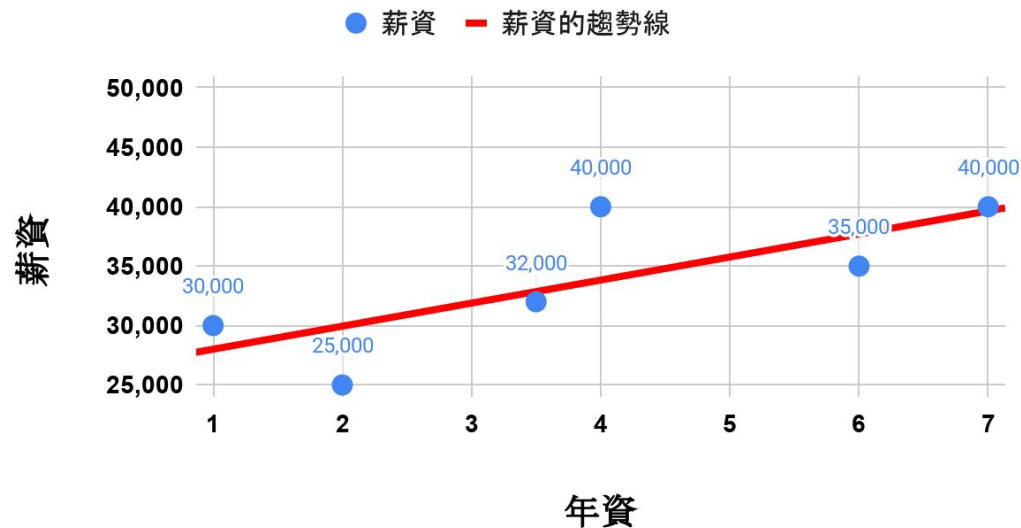


迴歸模型(續)

趨勢線可能造成缺乏適配
(Underfitting)問題。

缺乏適配問題的原因是模型
參數太少或過於簡單以致無
法捕捉到規律。

薪資年薪分布圖

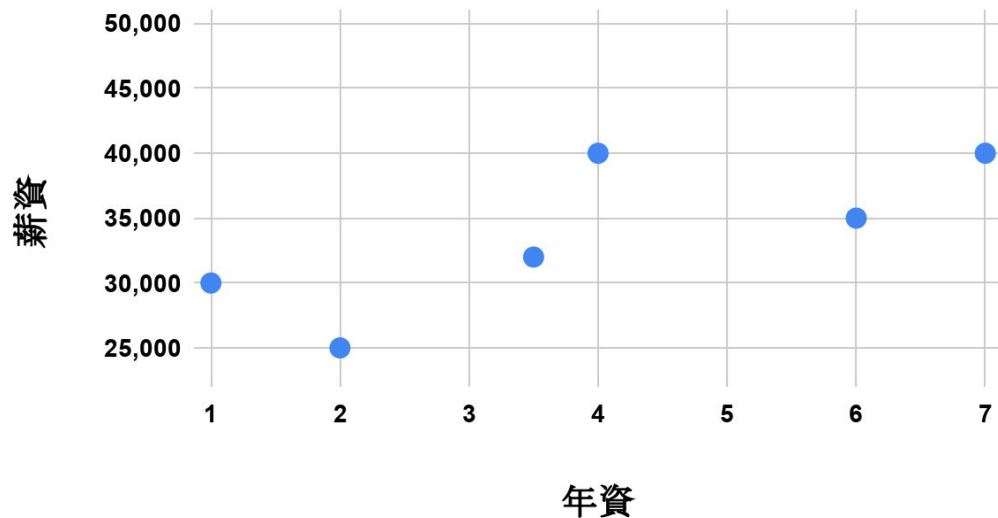


迴歸(續)

還有那些問題待解決

- 迴歸模型長怎樣?
 - 迴歸演算法?
- 迴歸模型預測的準嗎?
- ...


薪資年薪分布圖



如何評估迴歸模型?

Mean Square Error (MSE)

實際結果 預測結果


$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)$$

優點: 容易計算與理解

缺點: 沒有正規化

如何評估迴歸模型? (續)

Mean Square Error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)$$

請問此預測結果的MSE為多少?

序號	年資	薪資	預測薪資
1	1	30,000	
2	2	25,000	
3	3	35,000	29,000
4	3.5	32,000	
5	4	40,000	
6	4.5	32,000	39,000
7	6	35,000	
8	7	40,000	
9	8	50,000	45,000

如何評估迴歸模型? (續)

R² Score

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \tilde{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

實際結果 預測結果

實際結果的平均值

優點: 正規化(0 ~ 1)

缺點: 計算複雜

如何評估迴歸模型? (續)

R² Score

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \tilde{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

請問此預測結果的R²為多少?

厂序號	年資	薪資	預測薪資
1	1	30,000	
2	2	25,000	
3	3	35,000	29,000
4	3.5	32,000	
5	4	40,000	
6	4.5	32,000	39,000
7	6	35,000	
8	7	40,000	
9	8	50,000	45,000



課堂練習4-1

撰寫程式計算此訓練資料集的

1. MSE
2. R^2 Score

序號	年資 (Feature)	薪資 (Label)
1	1	30,000
2	2	25,000
3	3	35,000
4	3.5	32,000
5	4	40,000
6	4.5	32,000
7	6	35,000
8	7	40,000
9	8	50,000



評估迴歸模型(scikit-learn)

```
1. from sklearn.metrics import mean_square_error
2. from sklearn.metrics import r2_score
3. mean_square_error(y_true, y_predict)
4. r2_score(y_true, y_predict)
```




課堂練習4-2

使用scikit-learn內建的套件撰寫計算
此訓練資料集的

1. MSE
2. R^2 Score

序號	年資 (Feature)	薪資 (Label)
1	1	30,000
2	2	25,000
3	3	35,000
4	3.5	32,000
5	4	40,000
6	4.5	32,000
7	6	35,000
8	7	40,000
9	8	50,000

監督式學習

迴歸(Regression) / 簡單線性迴歸(Simple Linear Regression)

簡單來說，就是找到一個**線性模型**讓預測結果的誤差最小。

薪資 年資 係數 截距

coefficient intercept

薪資



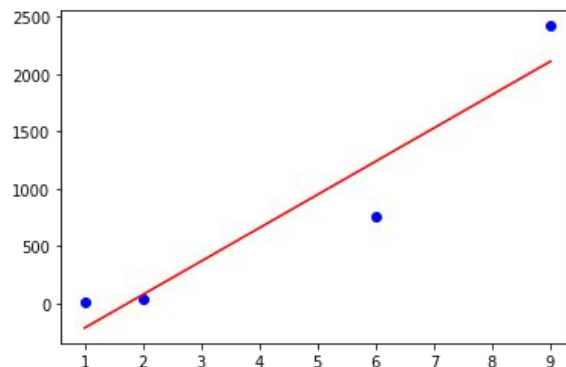
簡單線性迴歸(scikit-learn)

- 使用模組

[sklearn.linear_model.LinearRegression](#)

- 範例程式

```
1. from sklearn.linear_model import LinearRegression
2. X = [[1], [2], [6], [9]]
3. y = [5, 30, 750, 2421] #  $y \approx 3 * x_0^3 + 1$ 
4. reg = LinearRegression().fit(X, y)
5. reg.score(X, y) # R2 score
```





課堂練習4-3

以簡單線性迴歸訓練預測模型

1. 訓練資料集(1,2,4,5,7,8)
2. 測試資料集(3,6,9)
3. 繪製真實資料與預測模型圖

序號	年資 (Feature)	薪資 (Label)
1	1	30,000
2	2	25,000
3	3	35,000
4	3.5	32,000
5	4	40,000
6	4.5	32,000
7	6	35,000
8	7	40,000
9	8	50,000

監督式學習

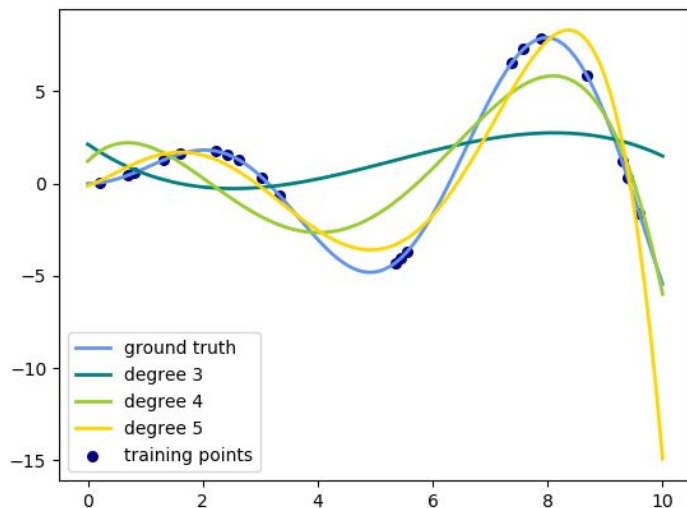
迴歸(Regression) / 多項式迴歸(Polynomial Regression)

多項式迴歸

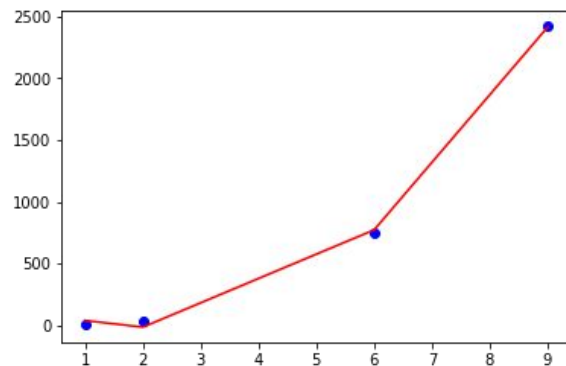
找到一個**多項式模型**來預測結果，可透過不同的次方(degree)來產生不同模型。

例如，一元 m 次多項式的模型為

$$h(x) = c_0 + c_1 \cdot x + c_2 \cdot x^2 + \dots + c_m \cdot x^m$$



多項式迴歸(scikit-learn)



- 使用模組

[sklearn.preprocessing.PolynomialFeatures](#)

- 範例程式

```
1. from sklearn.linear_model import LinearRegression
2. from sklearn.preprocessing import PolynomialFeatures
3. X = [[1], [2], [6], [9]]
4. y = [5, 30, 750, 2421] #  $y \approx 3 * x_0^3 + 1$ 
5. poly = PolynomialFeatures(degree=2)
6. TX = poly.fit_transform(X)
7. preg = LinearRegression().fit(TX, y)
8. preg.score(TX, y) # R2 score
```



課堂練習4-4

以多項式迴歸訓練預測模型

1. 訓練資料集(1,2,4,5,7,8)
2. 測試資料集(3,6,9)
3. 繪製真實資料與預測模型圖

序號	年資 (Feature)	薪資 (Label)
1	1	30,000
2	2	25,000
3	3	35,000
4	3.5	32,000
5	4	40,000
6	4.5	32,000
7	6	35,000
8	7	40,000
9	8	50,000



課堂練習4-5

以提供的資料集完成下列工作

1. 簡單線性迴歸模型
2. 多項式迴歸模型(可測試不同次方)
3. 比較各種方法的準確度

資料集

```
x = [[1], [3], [5], [7],  
      [9], [10], [12], [13],  
      [15], [17], [20], [22]]
```

```
y = [4, 9, 13, 19, 35, 50,  
      70, 100, 150, 200, 301,  
      496]
```



作業4-1

請利用台電的資料集進行下列分析

1. 使用線性迴歸模型分析單一屬性與用電量的關係
2. 使用多項式迴歸模型(不同次方)分析單一屬性與用電量的關係
3. 找出預測用電量最準確的迴歸分析方法與單一屬性
4. 請列出所有使用的迴歸分析方法、屬性與對應的R2 Score。

繳交方式: 報告(電子檔) / 格式不拘, 但內容完整度會影響分數

繳交期限: 6/1/ 2020

繳交方式: 上傳至ILMS

Tmax: 最高日溫平均

Tmin: 最低日溫平均

Wndspd: 風速(0:低於 6 節, 1:高於 6 節)

Cldcvr: 雲層(0~3: 晴朗到完全覆蓋)

Kwh: 用電量

監督式學習

分類(Classification)

分類

根據此資料集找到一個假說函數 h ，使得

$$h([\text{烤箱溫度}, \text{烤箱濕度}]) = \text{顧客評價}$$

自變數(Explained Variables)

應變數(Dependable Variables)

這樣就可以根據烤箱溫度與濕度預測滿意度！

備註：因為顧客評價(滿意、不滿意)是離散的，所以可以考慮使用分類演算法來解決這個問題。

烤箱溫度 (Feature)	烤箱濕度 (Feature)	顧客評價 (Label)
123	23	滿意
126	23	不滿意
124	25	不滿意
122	23	滿意
124	26	不滿意
124	22	滿意
127	23	?

分類(續)

訓練資料集(training set)

例如, 取第1, 2, 4, 5, 7, 8等6筆資料

假說函數(hypothesis) / 模型(Model)

運用分類演算法訓練 $h([溫度, 濕度]) = 滿意度$

測試資料集(test set)

例如, 取第3, 6, 9等3筆資料

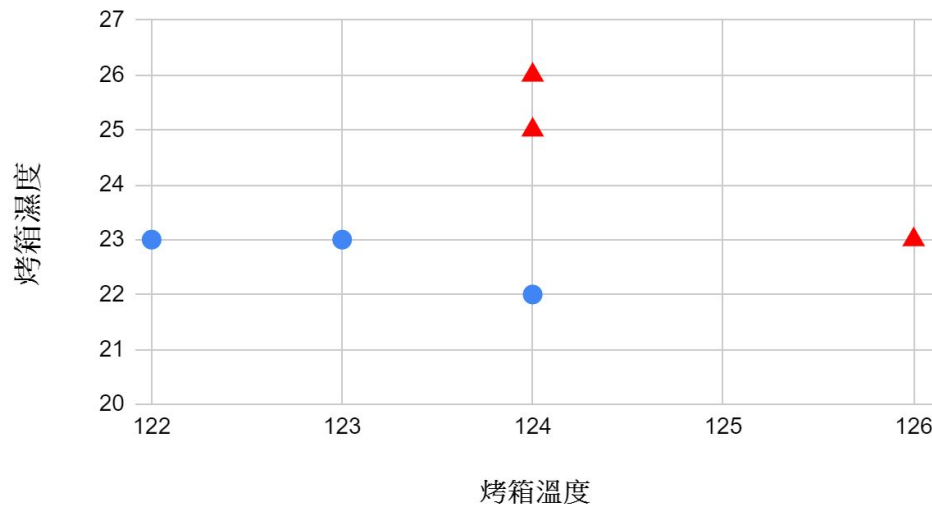
烤箱溫度 (Feature)	烤箱濕度 (Feature)	顧客評價 (Label)
123	23	滿意
126	23	不滿意
124	25	不滿意
122	23	滿意
124	26	不滿意
124	22	滿意
127	23	?

分類(續)

如何找出從烤箱溫度與烤箱濕度預測顧客滿意度的模型？

因為是預測離散數值，可考慮使用分類

烤箱濕度/溫度與顧客評價關係圖





如何評估分類模型?

TP: True Positive (判斷正確)

TN: True Negative (判斷正確)

FP: False Positive (把不對的判斷成對的)

FN: False Negative (把對的判斷成不對的)

正確分類率 = $(TP + TN) / (TP + TN + FP + FN)$

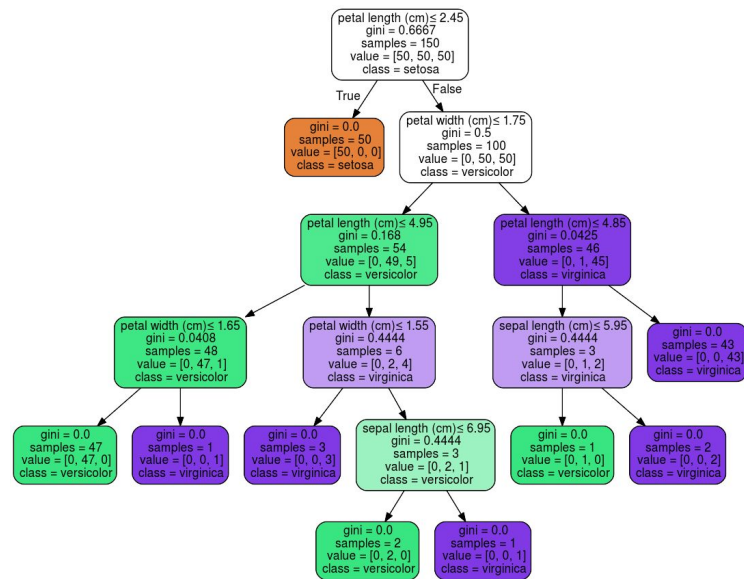
監督式學習

分類(Classification) / 決策樹(Decision Tree)

決策樹(Decision Tree)

決策樹的主要功能，是藉由分類已知的資料集來建立一個樹狀結構，並從中歸納出資料集裡、類別欄位與其它欄位間的隱藏規則。

所產生出來的決策樹，也能利用來做樣本的預測。



監督式學習

分類(Classification) / 隨機森林(Random Forest)

隨機森林(Random Forest)

隨機森林是一種集成學習(ensemble learning)方法。

簡單來說, 透過多組不同的決策樹(可能是演算法參數不同或是訓練資料集不同)來找決定最後的預測模型。

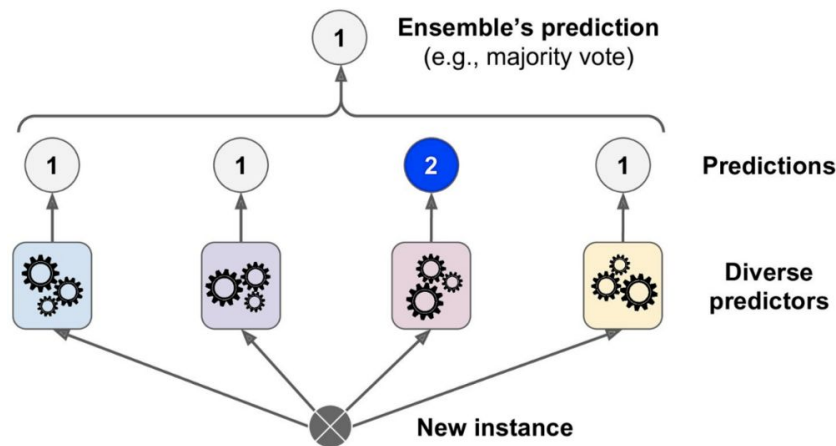


Figure 7-2. Hard voting classifier predictions

Q & A



Computer History Museum, Mt. View, CA