

# 應用類神經網路於 紅酒品質預測

R06H41019 黃俊凱 R06H41018 黃富纖  
R08H41005 張和家 R07H41012 羅翊仁

# Outline

- **Red Wine** Dataset
- Experiment Flows
- Supervised learning: Classification problem
  - BPNN
  - RBFNN
  - SVM
- Unsupervised learning: Clustering analysis
  - K-Means
  - Fuzzy C-Means

# Red Wine Dataset

- Source: Kaggle
- Sample: 1,599
- 11個特徵值(attribute)
- 類別標籤為紅酒品質等級

# Red Wine Dataset

## 變數與紅酒品質

## 內容與解釋

---

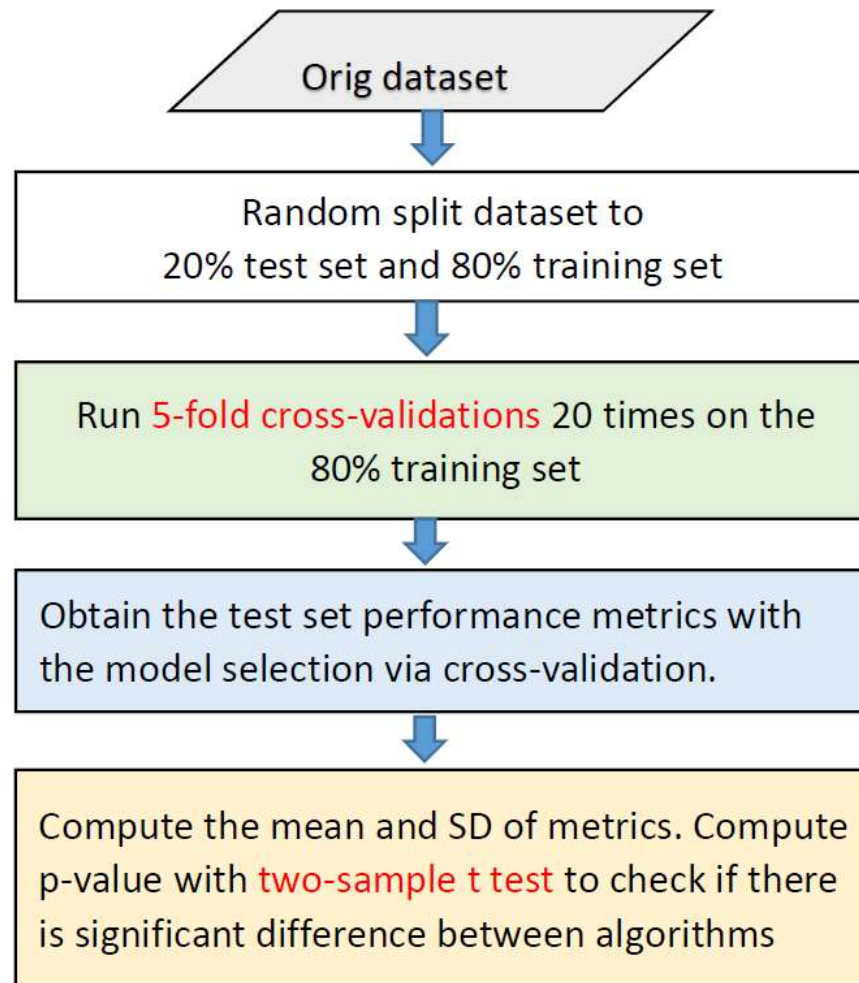
fixed acidity	固定酸度或非揮發性酸度
volatile acidity	酒中的醋酸含量，太高的話會有不好聞的醋味
citric acid	酒中的檸檬酸含量，可以增加酒的風味
residual sugar	發酵過後剩餘的糖分，通常大於45g/l算是甜的
chlorides	酒中的鹽含量
free sulfur dioxide	酒中的硫元素離子，抗氧化、抑菌並促進有益微生物生長
total sulfur dioxide	二氧化硫含量，當濃度超過50 ppm時用可以聞得到味道
density	水分的濃度，取決於酒精含量及糖分而定
pH	pH值用以斷定葡萄酒的整體酸度，pH值0到7屬於酸，pH值7到14屬於鹼。多數的葡萄酒pH在3之4之間
sulphates	硫酸鹽，是葡萄酒的添加劑，可作為抗氧化劑
alcohol	酒精含量的占比
quality	紅酒品質等級分數，0分表示最差，10分表示最好，但資料中沒有真正很差的0~2分也沒有最好的10分的酒

---

# Red Wine Dataset

- **Classification:**
  - **2 classes**
    - Positive:  $\text{quality} \geq 6.5$
    - Negative:  $\text{quality} < 6.5$
    - Performance metrics: Accuracy, F1 score
    - Positive : Negative = 217 : 1382
  - **3 classes**
    - Poor:  $\text{quality} \leq 4$
    - Fair:  $5 \leq \text{quality} < 7$
    - Good:  $\text{quality} \geq 7$
    - Performance metrics: Accuracy
    - Poor : Fair : Good = 63 : 1319 : 217
- **Clustering:**
  - By K-Means and Fuzzy C-Means

# Experiment Flow



# Unpaired Two-Sample T-test

- Suppose that  $\overline{X}_1$  is the sample mean of data 1 and  $\overline{X}_2$  is the sample mean of data 2,  $s_1^2$  is the sample variance of data 1,  $s_2^2$  is the sample variance of data 1, the number of data 1 is  $N_1$ , the number of data 2 is  $N_2$

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 > \mu_2$$

- $T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \sim t$  distribution with degree of freedom  $v$ ,

where

- $$v = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2 - 1}},$$

- Reject  $H_0$  if  $T > t_{1-\alpha, v}$

# Experiment Results

- BPNN
- RBF
- SVM
- Clustering



# BPNN

- 特徵集選取
  - 我們使用線性回歸為基的特徵選取方法。虛無假設為線性回歸的係數  $\beta_i = 0$ ，即
$$H_0: \beta_i = 0$$
  - 當p-value愈小表示有愈大的證據拒絕虛無假設，即該特徵愈重要。當取 p-value<0.01 在紅酒資料集會得到4個特徵值的資料集。對於BPNN我們分別試驗了分類問題為2類及3類，同時特徵值分別為11及4的資料集。
- BPNN 隱藏層神經元個數及隱藏層層數參數調校
  - 使用 api GridSearchCV 找出隱藏層神經元個數及隱藏層層數的最佳參數值，
  - 隱藏層層數分別為2,3,4，隱藏層神經元個數為 16 到 32 間隔為1，即16, 17, 18, 19, 20, 21, 22, ..., 31, 32。

# 隱藏層神經元個數及 隱藏層層數參數調校

	隱藏層層數	隱藏層神經元個數
<b>2 class</b>		
11 attributes	2	18
4 attributes	2	20
<b>4 class</b>		
11 attributes	2	29
4 attributes	2	19

# BPNN (1/2)

- 二類別分類問題
  - Accuracy: 11 attributes( $80.078\% \pm 2.675\%$ ) 統計上顯著優於4 attributes ( $74.984\% \pm 5.745\%$ ) (p-value= $0.00064 < 0.001$ )
  - F1: 11 attributes( $0.558 \pm 0.034$ ) 統計上顯著優於4 attributes( $0.506 \pm 0.055$ ) (p-value= $0.00054 < 0.001$ )
- 三類別分類問題
  - Accuracy:
    - 11 attribute( $82.984\% \pm 1.300\%$ )與 4 attribute( $82.234\% \pm 1.276\%$ )的正確率在統計上有些微顯著的差異 ( p-value =  $0.036 < 0.05$ )

# BPNN (2/2)

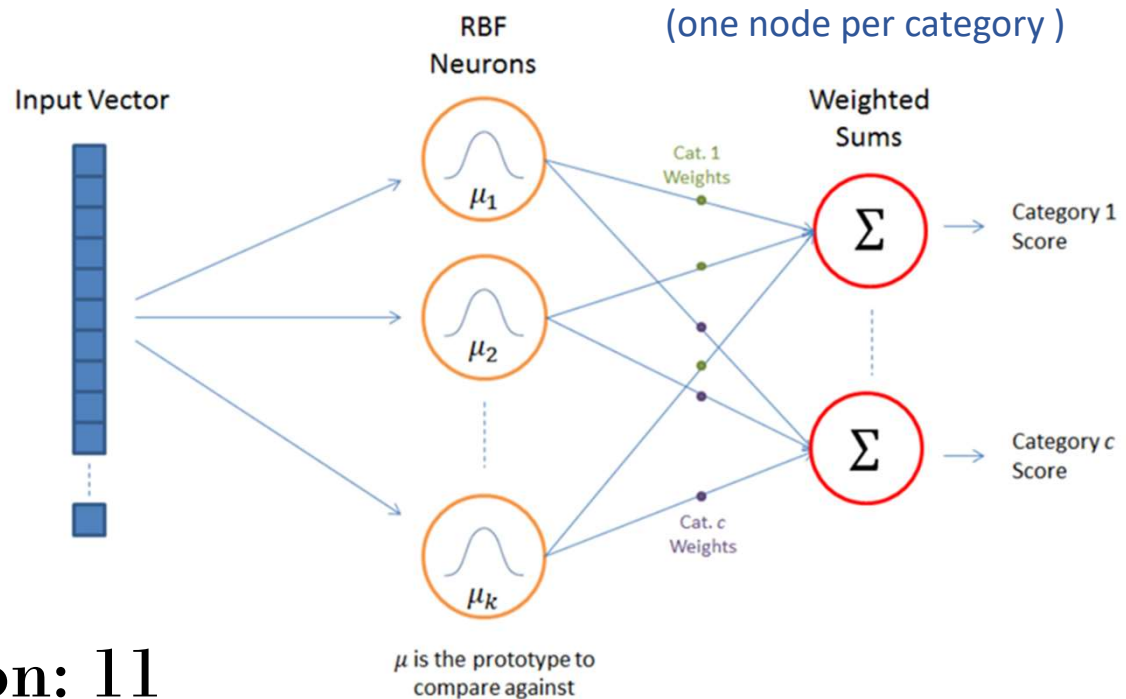
表 3.3.2.1 Three classes result for BPNN

Model Metrics	11 attributes	4 attributes
Accuracy		
Max	84.688%	84.688%
95%CI	(82.415%,83.554%)	(81.675%,82.793%)
Mean±SD	82.984%±1.300%	82.234%±1.276%

表 3.3.1.1 Two classes result for BPNN

Model Metrics	11 attributes	4 attributes
Accuracy		
Max	84.375%	81.875%
95%CI	(78.906%,81.250%)	(72.467%,77.502%)
Mean±SD	80.078%±2.675%	74.984%±5.745%
Sensitivity		
Max	87.234%	85.106%
95%CI	(84.679%,85.534%)	(85.106%,85.106%)
Mean±SD	85.106%±0.976%	85.106%±0.000%
PPV		
Max	48.193%	43.956%
95%CI	(39.991%,43.321%)	(33.829%,38.599%)
Mean±SD	41.656%±3.799%	36.214%±5.443%
NPV		
Max	97.248%	96.943%
95%CI	(96.748%,96.971%)	(96.446%,96.734%)
Mean±SD	96.860%±0.255%	96.590%±0.328%
Specificity		
Max	84.249%	81.319%
95%CI	(77.851%,80.574%)	(70.291%,76.193%)
Mean±SD	79.212%±3.106%	73.242%±6.734%
Odds_ratio		
Max	30.565	24.874
95%CI	(20.456,24.597)	(14.580,19.111)
Mean±SD	22.527±4.724	16.845±5.169
F1		
Max	0.615	0.580
95%CI	(0.543,0.573)	(0.482,0.530)
Mean±SD	0.558±0.034	0.506±0.055

# RBFNN



- input dimension: 11
- the number of neurons in the hidden layer: 10
- initialization of centers of RBFNN:  
randomly select from the given data set
- output dimension: 兩類別分類問題為2，  
三類別分類問題為3

# RBFNN

- 二類別分類問題
  - 在訓練集上執行20次5-fold交叉驗證之MSE與Accuracy
  - 準確度大多在80%以上，顯示模型有不錯的訓練結果

次數	1	2	3	4	5	6	7	8	9	10
MSE	0.1155	0.1827	0.1165	0.1371	0.1725	0.1234	0.1279	0.1121	0.1081	0.1062
Accuracy	0.8538	0.8233	0.8538	0.8538	0.8257	0.8530	0.8522	0.8530	0.8538	0.8538
次數	11	12	13	14	15	16	17	18	19	20
MSE	0.1201	0.1243	0.1330	0.1474	0.1012	0.1989	0.1171	0.1122	0.1104	0.1081
Accuracy	0.8553	0.8546	0.8522	0.8436	0.8577	0.7475	0.8538	0.8538	0.8554	0.8538

# RBFNN

- 二類別分類問題
  - 最佳模型在測試集上的預測結果
  - 20次預測之平均Accuracy為0.9043，平均F1 score為0.8636

次數	1	2	3	4	5	6	7	8	9	10
Accuracy	0.9063	0.8625	0.9063	0.9063	0.9063	0.9094	0.9063	0.9063	0.9063	0.9063
次數	11	12	13	14	15	16	17	18	19	20
Accuracy	0.9063	0.9063	0.9063	0.9063	0.9063	0.9063	0.9063	0.9063	0.9063	0.9063
次數	1	2	3	4	5	6	7	8	9	10
F1	0.8617	0.8738	0.8617	0.8617	0.8617	0.8691	0.8617	0.8617	0.8617	0.8617
次數	11	12	13	14	15	16	17	18	19	20
F1	0.8617	0.8617	0.8808	0.8617	0.8617	0.8617	0.8617	0.8617	0.8617	0.8617

**RBFNN在此資料集上有良好的預測能力！**

# RBFNN

- 三類別分類問題
  - 在訓練集上執行20次5-fold交叉驗證之MSE與Accuracy
  - 準確度大多在80%以上，顯示模型有不錯的訓練結果

次數	1	2	3	4	5	6	7	8	9	10
MSE	0.1043	0.1038	0.1277	0.1506	0.1019	0.1018	0.1141	0.0964	0.1111	0.1138
Accuracy	0.8178	0.8147	0.7802	0.6975	0.8147	0.8155	0.8147	0.8170	0.8170	0.8069
次數	11	12	13	14	15	16	17	18	19	20
MSE	0.1182	0.1122	0.1057	0.0970	0.1120	0.1006	0.1347	0.1144	0.1055	0.1164
Accuracy	0.8084	0.8147	0.8163	0.8131	0.7881	0.8155	0.7865	0.8108	0.8147	0.8123



# RBFNN

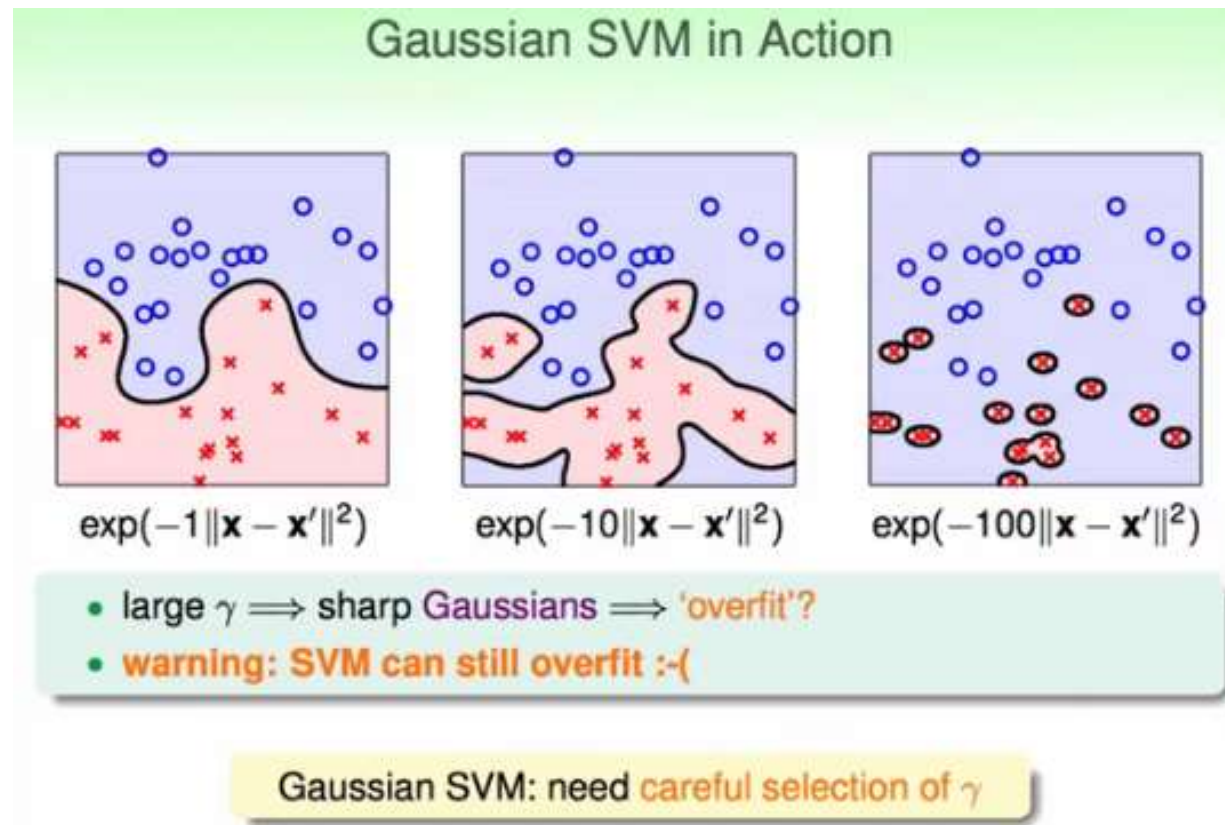
- 三類別分類問題
  - 最佳模型在測試集上的預測結果
  - 20次預測之平均Accuracy為0.8637

次數	1	2	3	4	5	6	7	8	9	10
Accuracy	0.8656	0.8656	0.8656	0.8656	0.8656	0.8656	0.8656	0.8656	0.8656	0.8656
次數	11	12	13	14	15	16	17	18	19	20
Accuracy	0.8656	0.8406	0.8656	0.8656	0.8562	0.8656	0.8656	0.8656	0.8625	0.8656

雖低於兩類別分類，但仍顯示RBFNN有不錯的預測能力！

# SVM

- 使用kaggle上grid search過最佳的c與gamma來訓練 (c=1,gamma=0.9)



# SVM (1/2)

- 二類問題
  - Accuracy:
    - 11 attributes( $77.141\% \pm 5.649\%$ )
- 三類問題
  - Accuracy:
    - 11 attribute( $85.188\% \pm 0.839\%$ )

# SVM (2/2)

Model Metrics	11 attributes
Accuracy	
Max	86.250%
95%CI	(84.820%,85.555%)
Mean±SD	85.188%±0.839%

Model Metrics	11 attributes
Accuracy	
Max	87.500%
95%CI	(74.665%,79.616%)
Mean±SD	77.141%±5.649%
Sensitivity	
Max	87.234%
95%CI	(85.004%,85.421%)
Mean±SD	85.213%±0.476%
PPV	
Max	54.795%
95%CI	(35.749%,41.842%)
Mean±SD	38.795%±6.952%
NPV	
Max	97.183%
95%CI	(96.593%,96.859%)
Mean±SD	96.726%±0.304%
Specificity	
Max	87.912%
95%CI	(72.849%,78.653%)
Mean±SD	75.751%±,6.621%
Odds_ratio	
Max	41.558
95%CI	(16.446,23.603)
Mean±SD	20.025±8.165
F1	
Max	0.667
95%CI	(0.502,0.558)
Mean±SD	0.530±0.064

# 模型比較

- 二類別分類問題

	BPNN	RBNFF	SVM
平均 Accuracy	80.08%	90.43%	77.14%
平均F1 Score	0.558	0.864	0.530

- 三類別分類問題

	BPNN	RBNFF	SVM
平均 Accuracy	82.98%	86.37%	85.19%

# 模型比較

- The accuracy of the three models has significance difference, and the order of the accuracy is different in 2 and 3 class.
- The order of the accuracy and f1 is consistent in 2 class.

x	y	performance	class	pvalue
rbf	bpnn	accuracy	2	9.07E-15
bpnn	svm	accuracy	2	0.022496
rbf	svm	accuracy	2	8.00E-10
2 class accuracy: rbf > bpnn > svm				
rbf	bpnn	f1	2	1.38E-20
bpnn	svm	f1	2	0.043931
rbf	svm	f1	2	7.24E-16
2 class f1: rbf > bpnn > svm				
rbf	bpnn	accuracy	3	2.09E-11
svm	bpnn	accuracy	3	1.76E-07
rbf	svm	accuracy	3	2.36E-06
3 class accuracy: rbf > svm > bpnn				

# Clustering

## Clustering



fuzzy C-means: center 1

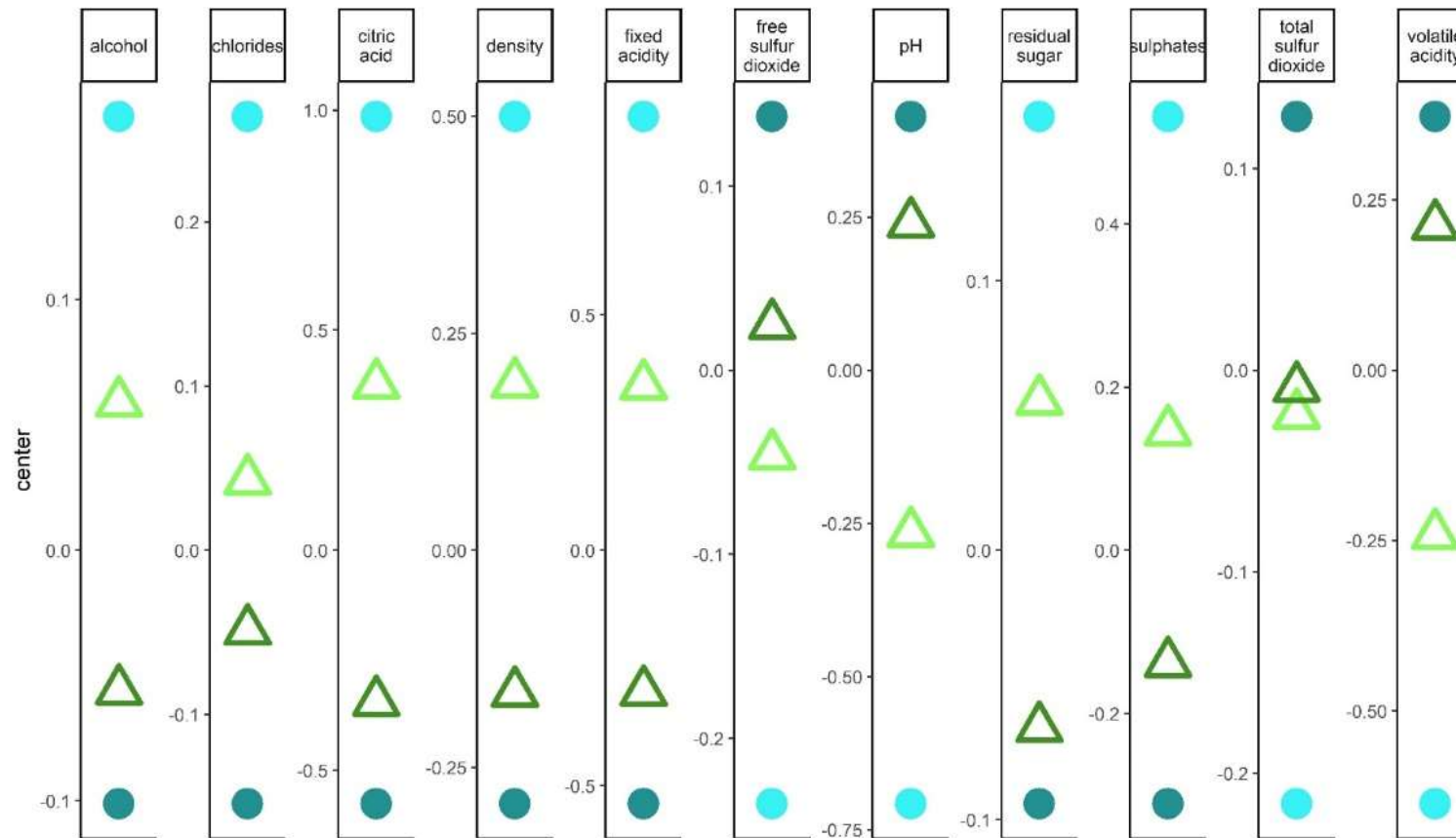
fuzzy C-means: center 2



K-means: center 1

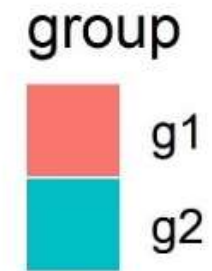
K-means: center 2

- K-Means 與 fuzzy C-Means 分成兩類

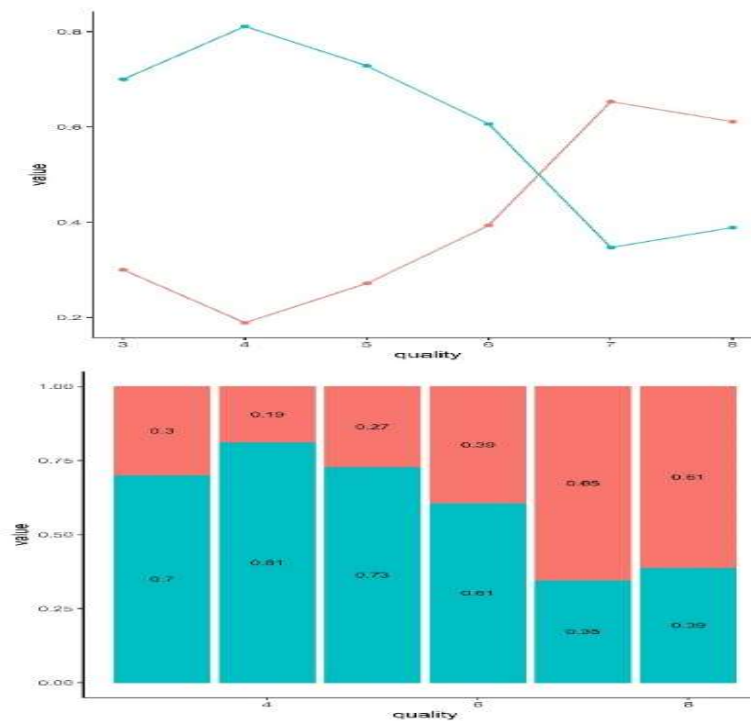


# Clustering

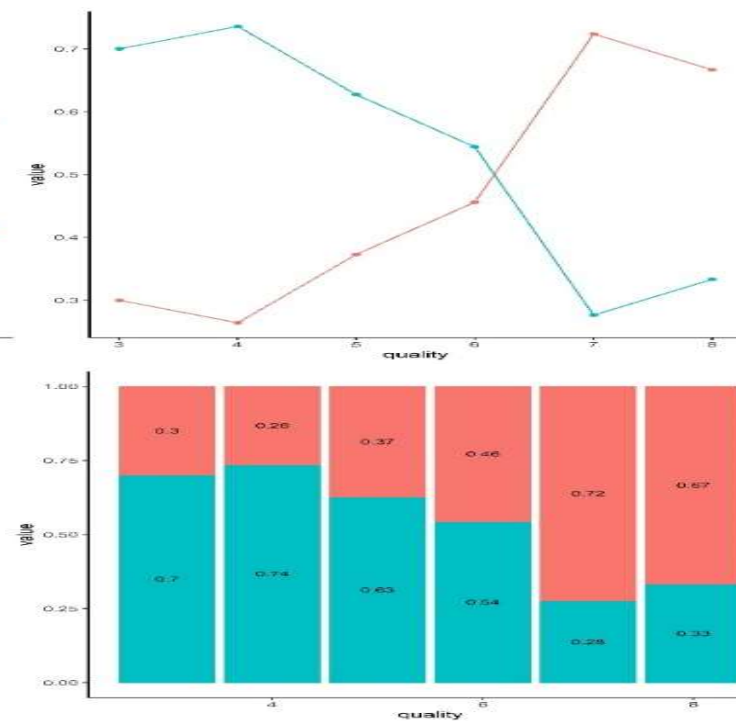
- K-Means 與 fuzzy C-Means 分成兩類



## K-Means






## fuzzy C-Means








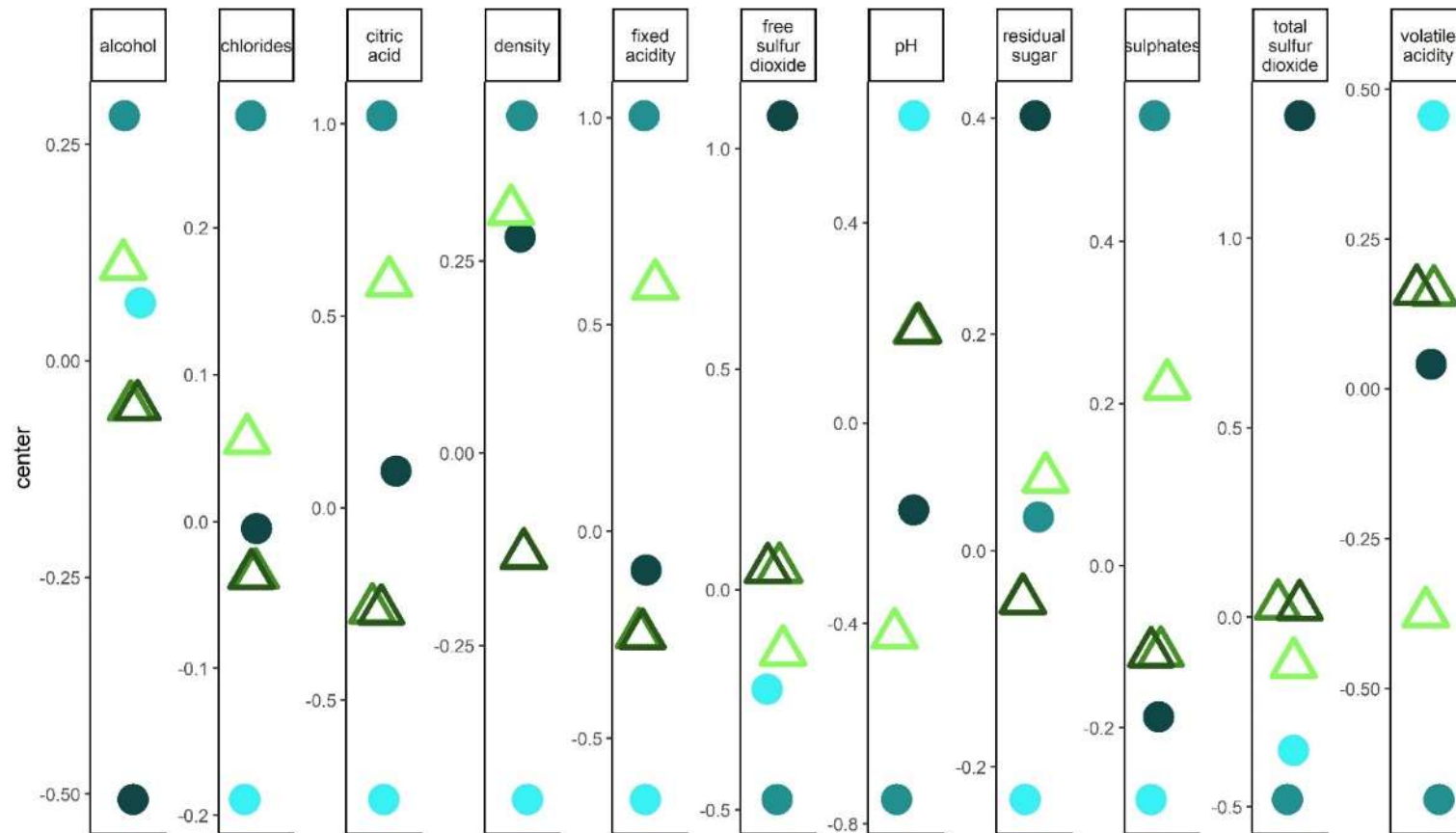
# Clustering

## Clustering

 fuzzy C-means: center 1  
 fuzzy C-means: center 2  
 fuzzy C-means: center 3

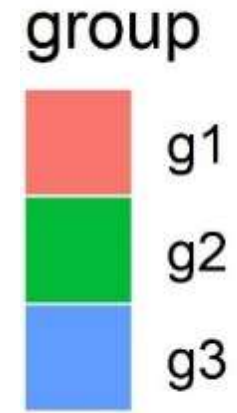
 K-means: center 1  
 K-means: center 2  
 K-means: center 3

- K-Means 與 fuzzy C-Means 分成三類

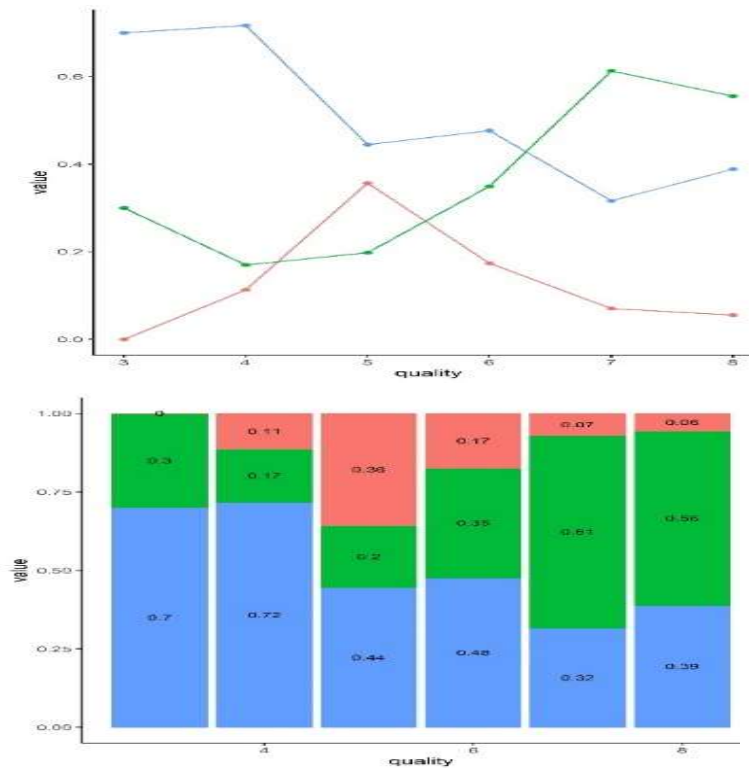


# Clustering

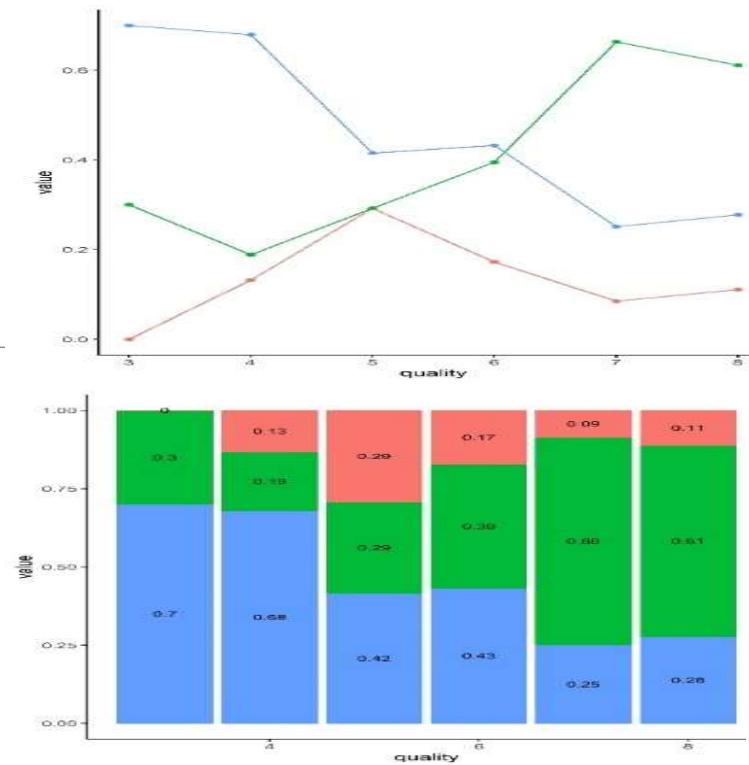
- K-Means 與 fuzzy C-Means 分成三類



K-Means



fuzzy C-Means



# Conclusions

- 類神經網路非常強大，透過調整參數等方法，總能得到良好的結果。如果沒有得到良好的結果，我們可能輕易地歸咎於參數調整不當，而忽略了其他的因素。
- 雖然類神經網路適合應用於作用機制複雜且難以描述的問題，但我們仍應對資料本身做深入的討論，幫助我們訓練出更好的模型。