3) The data set “glass.dat” was collected by the department of criminological investigation, it consists of 114 observations with 6 types, where each observation is described by 9 attributes (please refer to “glass\_description.txt” for variable description). At the scene of the crime, the glass left can be used as evidence, if its type is correctly identified.

**Q3:** Construct the decision rules for **classifying the types of glass** using (i) Classification Tree; (ii) LDA; (iii) QDA; (v) Nearest Neighbor; and (vi) Logistic discrimination. Compare all your resulting decision rules and explain which one you will best recommend.

資料簡介

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| refractive | sodium | Magnesium | Aluminum | Silicon | Potassium | Calcium | Barium | Iron | Type |
| 1.52101 | 13.64 | 4.49 | 1.1 | 71.78 | 0.06 | 8.75 | 0 | 0 | 1 |
| 1.51766 | 13.21 | 3.69 | 1.29 | 72.61 | 0.57 | 8.22 | 0 | 0 | 1 |
| 1.51743 | 13.3 | 3.6 | 1.14 | 73.09 | 0.58 | 8.17 | 0 | 0 | 1 |
| 1.51755 | 13 | 3.6 | 1.36 | 72.99 | 0.57 | 8.4 | 0 | 0.11 | 1 |
| 1.51571 | 12.72 | 3.46 | 1.56 | 73.2 | 0.67 | 8.09 | 0 | 0.24 | 1 |
| 1.51748 | 12.86 | 3.56 | 1.27 | 73.21 | 0.54 | 8.38 | 0 | 0.17 | 1 |

（僅顯示前六筆資料）

統計這筆資料各Type數

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type | 1 | 2 | 3 | 4 | 5 | 6 |
| 個數 | 23 | 42 | 11 | 21 | 11 | 6 |

|  |  |  |
| --- | --- | --- |
| 變數 | 簡介 | 型態 |
| refractive | 折射率 | 數值 |
| sodium | 鈉 | 數值 |
| Magnesium | 鎂 | 數值 |
| Aluminum | 鋁 | 數值 |
| Silicon | 矽 | 數值 |
| Potassium | 鉀 | 數值 |
| Calcium | 鈣 | 數值 |
| Barium | 鋇 | 數值 |
| Iron | 鐵 | 數值 |
| Type | 1=building\_windows\_float\_processed  2=building\_windows\_non\_float\_processed  3=vehicle\_windows\_float\_processed  4 = containers  5 = tableware  6 = headlamps | 類別 |

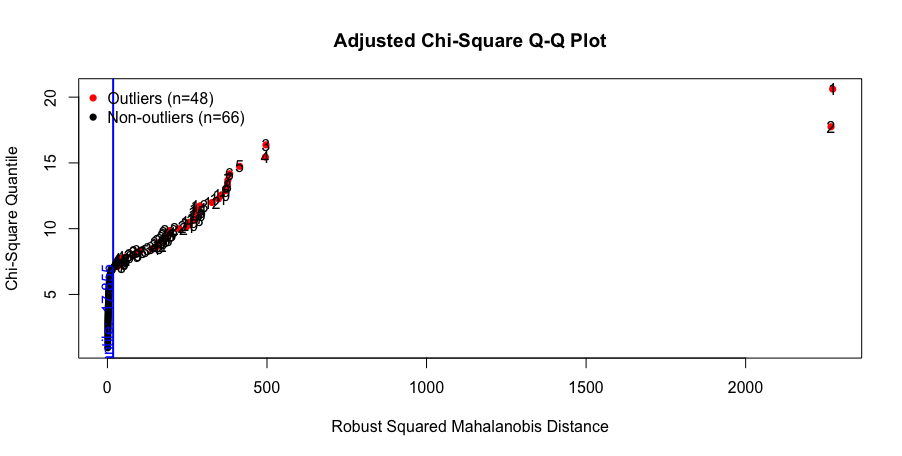
此為利用成分去歸類玻璃用途的資料，總共包含114樣本資料，變數共有10個，其中Type為玻璃的樣式（Type1 ~ Type6），我們要用剩下9個變數的資料去分類分群玻璃用途（Type）。分類分群的方法有 Classification Tree、 LDA、 QDA、Nearest Neighbor與Logistic discrimination，並從中找出分類效果最好的方法(True error rate越小越好)，藉此利用成分去分類分群未知玻璃用途（Type）。

（二）檢測離群值、共線性與資料是否為常態

檢測離群值

由變數Barium, Iron, Type會造成檢測失敗，故我們家這三項變數拔除並檢測是否有離群值存在。

使用Robust Squared Mahalanobis Distance方法檢測是否離群值存在，應盡量避免離群值存在，以避免資料解釋誤差產生。



由塗上很面顯看出有48筆離群值存在，但這筆資料僅有114筆資料，若是我們任意刪除這48筆離群值，則會造成42％的資訊損失，因此我們不刪除任意離群值。

檢測常態

多元常態檢定

|  |  |  |  |
| --- | --- | --- | --- |
| Test | Statistic | p-value | Result |
| Mardia Skewness | 1270.60785 | 3 | NO |
| Mardia Kurtosis | 36.1560802 | 0 | NO |
| MVN | <NA> | <NA> | NO |

個別變數常態檢定

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | Variable | Statistic | p-value | Normality |
| Shapiro-Wilk | refractive | 0.9412 | 0.00001 | NO |
| Shapiro-Wilk | sodium | 0.9228 | <0.001 | NO |
| Shapiro-Wilk | Magnesium | 0.7521 | <0.001 | NO |
| Shapiro-Wilk | Aluminum | 0.9473 | 0.00002 | NO |
| Shapiro-Wilk | Silicon | 0.9213 | <0.001 | NO |
| Shapiro-Wilk | Potassium | 0.4168 | <0.001 | NO |
| Shapiro-Wilk | Calcium | 0.8689 | <0.001 | NO |

從上表發現，檢定多元常態與檢定個別變數常態都呈現拒絕的假設。為了讓資料符合多元常態，我們進行了Boxcox的轉換，但檢定結果依然拒絕的假設。因此我們推測資料可能違反多元常態的假設。

檢測是否有共線性存在

我們觀測相關係數矩陣中，是否有相關係數呈現高度相關( >0.98或 < -0.98 )

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | refractive | sodium | Magnesium | Aluminum | Silicon | Potassium | Calcium | Barium | Iron |
| refractive | 1 | -0.2042 | -0.0243 | -0.385 | -0.3468 | -0.2974 | 0.7675 | -0.2086 | 0.1951 |
| sodium | -0.2042 | 1 | -0.4909 | 0.1508 | 0.1977 | -0.2853 | -0.1684 | 0.5112 | -0.2186 |
| Magnesium | -0.0243 | -0.4909 | 1 | -0.4807 | -0.2258 | 0.0116 | -0.4045 | -0.4854 | 0.1123 |
| Aluminum | -0.385 | 0.1508 | -0.4807 | 1 | -0.1984 | 0.38 | -0.2699 | 0.5692 | -0.1218 |
| Silicon | -0.3468 | 0.1977 | -0.2258 | -0.1984 | 1 | -0.4919 | -0.0046 | 0.0078 | -0.1124 |
| Potassium | -0.2974 | -0.2853 | 0.0116 | 0.38 | -0.4919 | 1 | -0.3359 | -0.0969 | -0.0191 |
| Calcium | 0.7675 | -0.1684 | -0.4045 | -0.2699 | -0.0046 | -0.3359 | 1 | -0.2618 | 0.1509 |
| Barium | -0.2086 | 0.5112 | -0.4854 | 0.5692 | 0.0078 | -0.0969 | -0.2618 | 1 | -0.1624 |
| Iron | 0.1951 | -0.2186 | 0.1123 | -0.1218 | -0.1124 | -0.0191 | 0.1509 | -0.1624 | 1 |

從上表中觀測，別為有兩倆變數之相關係數超過0.98或小於-0.98，因此我們不考慮拿掉任意變數。

（三）運用不同方法去分類玻璃用途（Type）

Apparent error rate： 運用訓練樣本去建立模型，再用建立的模型去預測訓練樣本的資料，並比較預測的資料與真實樣本資料，則Apparent error rate:為此預測錯誤的比率。

True error rate：運用訓練樣本去建立模型，再用建立的模型去預測測試樣本的資料，並比較預的資料與真實樣本資料，則True error rate:為此預測錯誤的比率。

在本題我們運用整筆資料去計算Apparent error rate，並利用資料採取cross validation(leave one out)方式去計算True error rate.

若是Apparent error rate太大，則表示這方法不太用是這筆資料，若不是太大，則可以進一步計算True error rate。

小常識： 交叉驗證 （K-fold Cross Validation）: 此法是將資料拆成K個(一般為5或10)沒交集的群組，接著用K-1群的資料來建模，再用第K群的資料來做預測。重複這步驟K次，直到每一群都被用來做一次預測和被用來建立K-1次模型。交叉驗證可以測量模型的準確度，這對於檢視模型品質會很有幫助。而這題我們使用了K為樣本個數進行分群，也就是所謂的留一驗證 ( leave-out-out cross validation)。

在比較方法好壞時，我們會看True error rate的高低，越低則表示分類準確度越高。

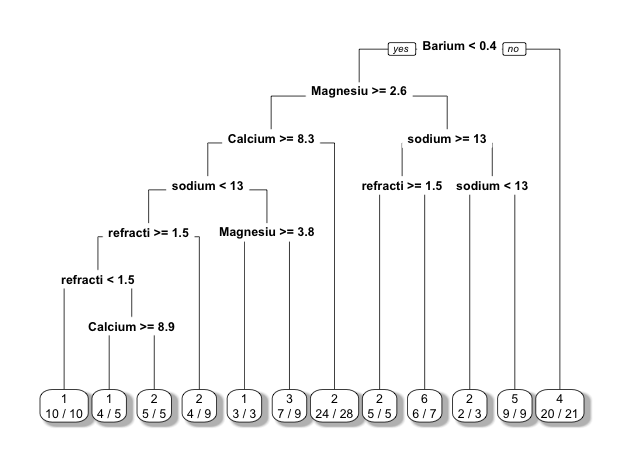
(1) Classification Tree

概念：

我們透過決策樹中的information gaim的方式去選擇分類樣本的變數，並設定分類準則即會產生分支，再對每一個分支重複一樣的動作，依序產生更多的分支，最終會形成一刻類似樹的樣子。

我們利用R的rpart與prp套件將這筆glass的資料使用Classification Tree方式去分類資料，其中R的參數中我們設定minbucket 為3以及minsplit為10即在要分類前，該節點樣本數至少10筆，且末端節點至少包含3筆資料，這樣程式才會繼續執行。

產生以下圖示



計算結果



**Apparent error rate**: 0.13158

**True error rate**: = 0.34211

由於**Apparent error rate與True error rate差距有點大，我們認為**會產生overfitting的問題，會造成模型適應性不好，即帶入其他測試樣本會產生過大誤差。

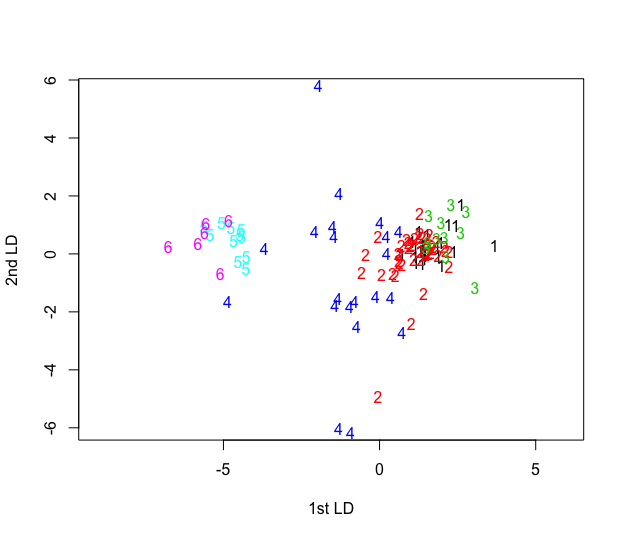
(2) Linear Discriminant Analysis, LDA

概念：找出一條線性組合，讓這條線能夠區分不同類別的資料，並達到同群內的群內變異數小，不同群的群間變異數大的效果。

計算結果：混淆矩陣

|  |  |
| --- | --- |
| 將整筆資料帶入做訓練 | cross validation(leave one out) |
|  |  |
| Apparent error rate : 0.28947 | True error rate : 0.35965 |

（對角線上的數字為分對的個數, 其餘則為分錯的個數，error rate為分錯的比率）



上圖為我們將資料投影到LD1以及LD2的座標軸上，明顯看出type1,2,3混雜在一塊，而type 5,6混雜在一塊，說明了為何在整筆資料建立混肴矩陣時，type1,2,3嚴重會分錯，可能在不同的LD可以清楚分出type5,6。因此在混淆矩陣上type5,6並沒有分錯，但在LD1, LD2上顯示type5, 6混在一起。

(3) Quadratic Discriminant Analysis, QDA

概念：類似於LDA, 但LDA為畫出一直線去區分資料，QDA則是用二次曲線的方式去將資料分類。

但在做QDA前，資料須服從多元常態假設，但在之前的常態假設中，這筆資料並未顯示服從多元常態，且將資料進行Boxcox轉換後，仍不服從多元常態假設。為了練習方便，我們將這筆資料假設服從多元常態。

並且在計算QDA時，Barium幾乎只有在glass為containers時才不為0以及glass為headlamps的個數太少，會造成QDA無法運算，因此我們將這個應因素去除，並計算QDA的Apparent error rate以及 true error rate.

計算結果：混淆矩陣

|  |  |
| --- | --- |
| 將整筆資料帶入做訓練 | cross validation(leave one out) |
|  |  |
| Apparent error rate : 0.18519 | True error rate : 0.38889 |

（對角線上的數字為分對的個數, 其餘則為分錯的個數，error rate為分錯的比率）

(4) Nearest Neighbor, NN

概念：藉由該筆樣本的最接近的k點去分類該筆樣本為哪一類，若該點附近皆為A類，則該筆樣本屬於A類。

在本題之中，我們分別計算k = 1, 2, 3 的Apparent error rate以及True error rate。（樣本的應變數應為類別型態）

當k為1時，我們選擇該筆樣本資料最接近的1個點的類別去定義該筆樣本的類別。

混淆矩陣

|  |  |
| --- | --- |
| 將整筆資料帶入做訓練 | cross validation(leave one out) |
|  |  |
| Apparent error rate : 0 | True error rate : 0.31579 |

（對角線上的數字為分對的個數, 其餘則為分錯的個數，error rate為分錯的比率）

由Apparent error rate為0可知，當k = 1時，會產生嚴重overfitting的問題，會造成模型適應性不好，即帶入其他測試樣本會產生過大誤差。因此k = 1並不是一個好的選擇。

當k為2時，我們選擇該筆樣本資料最接近的1個點的類別去定義該筆樣本的類別。

混淆矩陣

|  |  |
| --- | --- |
| 將整筆資料帶入做訓練 | cross validation(leave one out) |
|  |  |
| Apparent error rate : 0.20175 | True error rate : 0.36842 |

（對角線上的數字為分對的個數, 其餘則為分錯的個數，error rate為分錯的比率）

由於**Apparent error rate與True error rate差距有點大，我們認為可能**會產生overfitting的問題，會造成模型適應性不好，即帶入其他測試樣本會產生過大誤差。

當k為3時，我們選擇該筆樣本資料最接近的1個點的類別去定義該筆樣本的類別。

混淆矩陣

|  |  |
| --- | --- |
| 將整筆資料帶入做訓練 | cross validation(leave one out) |
|  |  |
| Apparent error rate : 0.18421 | True error rate : 0.35088 |

（對角線上的數字為分對的個數, 其餘則為分錯的個數，error rate為分錯的比率）

由於**Apparent error rate與True error rate差距有點大，我們認為可能**會產生overfitting的問題，會造成模型適應性不好，即帶入其他測試樣本會產生過大誤差。

(5) Logistic discrimination

資料的應變數為類別型態，使用全部的因變數去做Logistic regression 模型，進而計算**Apparent error rate與True error rate**

|  |  |
| --- | --- |
| 將整筆資料帶入做訓練 | cross validation(leave one out) |
|  |  |
| Apparent error rate : 0.193 | True error rate : 0.37719 |

由於**Apparent error rate與True error rate差距有點大，我們認為可能**會產生overfitting的問題，會造成模型適應性不好，即帶入其他測試樣本會產生過大誤差。

統整並比較這5個方法的運用於此資料的優劣

|  |  |  |
| --- | --- | --- |
| 方法 | Apparent error rate | True error rate |
| Classification Tree | 0.13158 | 0.34211 |
| Linear Discriminant Analysis | 0.28947 | 0.35965 |
| Quadratic Discriminant Analysis | 0.18519 | 0.38889 |
| Nearest Neighbor (k = 2) | 0.20175 | 0.36842 |
| Nearest Neighbor (k = 3) | 0.18421 | 0.35088 |
| Logistic discrimination | 0.193 | 0.37719 |

根據以上表格中的True error rate比較，由Classification Tree方法做出來的True error rate相對其他方法所做出的True error rate來得低，其次是Linear Discriminant Analysis與Nearest Neighbor( k = 3 )兩個方法，較不建議使用Quadratic Discriminant Analysis方法，因為使用QDA的條件較為嚴厲，資料需服從多元常態，即便不服從多元常態，也是可以將資料進行QDA分類，但誤差可能就會較其他方法來得大。因此，對於此筆資料(glass)，我們建議使用Classification Tree、Nearest Neighbor( k = 3 )與Linear Discriminant Analysis。