專題：誰該被當掉：關於機率的應用

目的:

本專題以常見的分數分佈的問題, 說明機率觀念在解決問題上的能力。透過直覺的想法, 佐以機率的理論與電腦程式的寫作, 可以從學生分數的表現中分離出不同的程度的群組。想法、理論與實作結合的練習, 是本單元主要的目的。其中, 演算法 (Algorithm) 的程式寫作是主要的練習目標。

為什麼:

機率的概念給人的感覺是理論的、抽象的, 通常是用來做理論證明或推論的。其實, 機率的觀念也是提供解決問題的一種手段, 只不過學過機率論並不能保證學會這個方法。因爲從機率的觀點切入問題的核心需要不斷的練習, 譬如對樣本的機率分配假設, 寫出概似函數或後驗機率, 再配合實作的演練, 方能逐漸駕輕就熟, 成爲未來解決問題的方法之一。

做什麼:

本單元想從機率的觀點出發, 將隱藏在一組樣本裡的群組資訊顯現出來, 譬如依學生的考試分數將學生分成「有學習意願」與「無學習意願」兩群。[1] 這牽涉到對樣本資料的假設(譬如服從常態分配), 資料的模型(譬如雙峰常態) 及分群的方法。學生的分數(樣本) 透露多少學生該屬於哪一群的資訊? 某生有多大可能性屬於某一群? 這些都是機率式的思考模式, 需要練習以機率函數的方式表達, 其中概似函數、後驗機率與先驗機率是最常使用的機率函數。這些機率函數的意義、使用與表達是單元練習的重點。而依據這些機率函數的分群方法, 則需要演算法的幫法才能眞正達到分群的目的(看到分群的結果。) 演算法的落實便是程式寫作的發揮。從觀念的切入、資料與模型假設與程

[1] 一般而言, 人類的學習能力常被假設具常態分配, 所以諸如考試分數之類的能力評量資料常呈現鐘形的分佈情況。但這個結果根植於所有參與者都具學習意願的假設, 當有一部份人拒絕學習時, 情況會變得複雜。於是在某些情況下, 當資料呈現出「不正常」的分佈時, 透過分群的方式, 常可以發現問題的原因, 譬如城鄉差距。

式的撰寫, 這一整套的過程是本單元練習的重點。

怎麼做:

本實驗以學生的學期分數爲例, 考驗老師當人的抉擇; 是在 60 分畫一條界線, 低於 60 分者被當, 還是另有其他選擇? 是否該考慮分數的分配情況? 機率的觀念是否可以幫忙做決定? 圖 1 展示四種學期成績的分配情況, 如果妳是老師, 最想見到哪一種分佈呢?

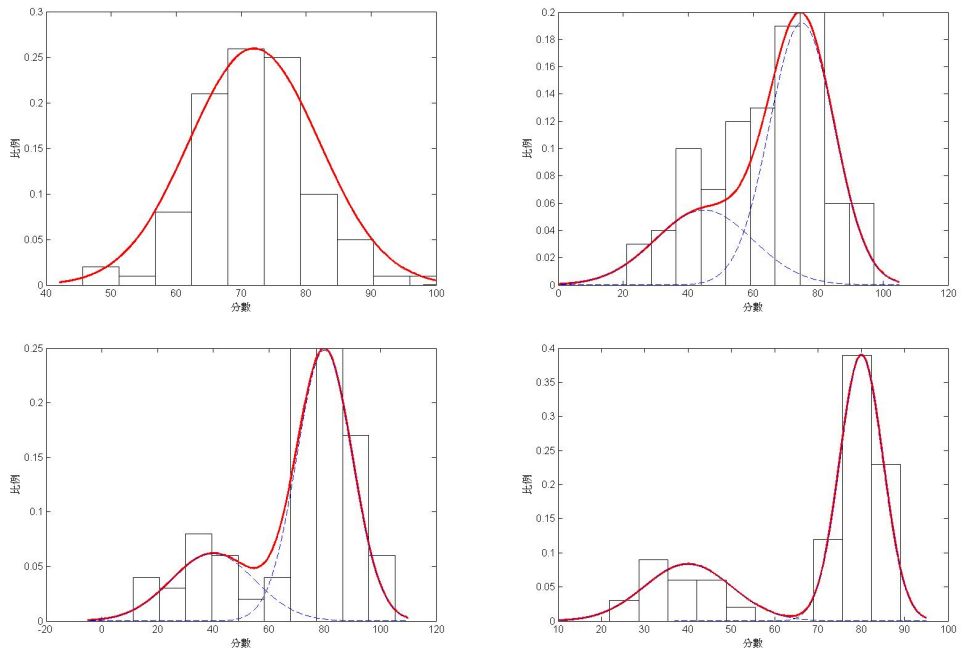

圖 1: 四種學期成績的分佈情況

圖 1 除左上圖外, 其餘三張圖都呈現出程度不一的雙峰分佈, 教育單位或專家可以舉出各種原因來說明這個「非常態」分佈的現象, 不過這不是本實驗的目的, 這個實驗想透過機率的觀念去估計這個雙峰的分配。理論很簡單, 方法很單純, 就是需要寫程式來完成這個工作。

問題描述: 假設全班的學期分數$y_1, y_2, \cdots, y_N$ 來自兩個常態分配的母體,[2]即 y_i 可能來自 $N(\mu_1, \sigma_1^2)$ 或 $N(\mu_2, \sigma_2^2)$ 。其機率密度函數寫成

$$p(\mathbf{y}) = \pi_1 p(\mathbf{y}|\mu_1, \sigma_1^2) + \pi_2 p(\mathbf{y}|\mu_2, \sigma_2^2) \tag{1}$$

其中 π_1, π_2 分別代表兩個常態分配的組合比例。我們想從已知的分數樣本去估計式 (1) 中的未知參數: $\pi_1, \mu_1, \mu_2, \sigma_1, \sigma_2$。[3]

實驗1: 實驗步驟

1. 先猜測兩常態群組的參數:$\mu_1, \sigma_1, \mu_2, \sigma_2$ 做爲第一次的估計。

2. 判斷每一筆分數樣本最可能來自哪一個的常態群組, 即

   $y_i \in$ 群組1 , 如果 $p(y_i|\mu_1, \sigma_1) > p(y_i|\mu_2, \sigma_2)$, 反之,$y_i \in$ 群組2

   $i = 1, 2, \cdots, N$

3. 根據步驟2的分群結果, 估計參數

$$\begin{aligned}
\pi_1 &= \frac{N_1}{N}, \quad N_1 : y_i\text{屬於群組1 的個數} \\
\mu_1 &= \frac{1}{N_1} \sum_{y_i \in \text{群組1}} y_i \\
\mu_2 &= \frac{1}{N_2} \sum_{y_i \in \text{群組2}} y_i, \quad N_2 : y_i\text{屬於群組2 的個數} \\
\sigma_1^2 &= \frac{1}{N_1 - 1} \sum_{y_i \in \text{群組1}} (y_i - \mu_1)^2 \\
\sigma_2^2 &= \frac{1}{N_2 - 1} \sum_{y_i \in \text{群組1}} (y_i - \mu_2)^2
\end{aligned}$$

重複步驟 2,3 直到所有估計值不再改變爲止。每次的重複都會更新這些參數, 但變化幅度可能會越來越小, 最後幾乎不再變化, 稱爲「收斂。」寫程式時可以設定一個較大

[2]一般而言, 一個母體由多個常態的母體所組成, 其分配稱爲混合常態分配(Normal Mixture)。

[3]$\pi_2 = 1 - \pi_1$

迴圈數, 列印每次迴圈最後的更新結果, 以肉眼設定收斂的標準, 記錄最後的結果。或是在程式迴圈中, 設定一個微小的差距值 ϵ , 若前後兩次的估計值的差小於 ϵ, 則程式停止。譬如以本實驗爲例, 共有5個估計值, 其整體的大小以下式表示。

$$\text{norm}([\pi_1 \;\; \mu_1 \;\; \mu_2 \;\; \sigma_1 \;\; \sigma_2])$$

不過當估計值不只一個時, 必須小心使用這個方式, 特別是不同估計值間的規模(scale)差距太大時, 適當的調整是必要的。

步驟2以機率密度函數值來判別群組的方式, 並未考慮組群的大小, 直覺上不夠好。如果換成 $\pi_1 p(y_i|\mu_1, \sigma_1) > \pi_2 p(y_i|\mu_2, \sigma_2)$, 是否表現比較好?

實驗2: 實驗步驟同實驗1, 不過修改步驟 2 中對於樣本 y_i 的群組屬性的判斷

1. 先猜測兩常態群組的參數:$\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2$ 做爲第一次的估計。

2. 判斷每一筆分數樣本最可能來自哪一個的常態群組, 即

$$y_i \text{ 的群組屬性 } \Delta_i \overset{\text{抽樣}}{\sim} Bino(1, \gamma_{i1}), \;\; \text{其中}$$
$$\gamma_{i1} = \frac{\pi_1 p(y_i|\mu_1, \sigma_1)}{\pi_1 p(y_i|\mu_1, \sigma_1) + (1-\pi_1) p(y_i|\mu_2, \sigma_2)}$$

3. 估計參數

$$\begin{aligned}
\pi_1 &= \frac{1}{N}\sum_{i=1}^{N} \Delta_i \\
\mu_1 &= \frac{\sum_{i=1}^{N} \Delta_i y_i}{\sum_{i=1}^{N} \Delta_i} = \frac{1}{N\pi_1}\sum_{i=1}^{N} \Delta_i y_i \\
\mu_2 &= \frac{\sum_{i=1}^{N} (1-\Delta_i) y_i}{\sum_{i=1}^{N} (1-\Delta_i)} = \frac{1}{N(1-\pi_1)}\sum_{i=1}^{N} (1-\Delta_i) y_i \\
\sigma_1^2 &= \frac{\sum_{i=1}^{N} \Delta_i (y_i-\mu_1)^2}{\sum_{i=1}^{N} \Delta_i} = \frac{1}{N\pi_1}\sum_{i=1}^{N} \Delta_i (y_i-\mu_1)^2 \\
\sigma_2^2 &= \frac{\sum_{i=1}^{N} (1-\Delta_i)(y_i-\mu_2)^2}{\sum_{i=1}^{N} (1-\Delta_i)} = \frac{1}{N(1-\pi_1)}\sum_{i=1}^{N} (1-\Delta_i)(y_i-\mu_2)^2
\end{aligned}$$

重複步驟 2,3 直到所有估計值不再改變爲止。這裡抽樣得到的 Δ_i 非 1 即 0, 當 $\Delta_i = 1$ 代表 y_i 屬於群組 1, 當 $\Delta_i = 0$ 代表 y_i 屬於群組 2。這是另類判斷群組別的方式, 其實參數估計的方式只是分別處理屬於不同群組的樣本而已, 與實驗 1 的估計方式相同。

由於群組的屬性來自抽樣的結果, 參數的估計並不會如實驗1產生收斂的效果, 也就是參數估計的結果會持續變動, 只是變動幅度會趨於穩定。這時候不能採用實驗 1 的方式停止程式的遞迴。而是讓程式持續執行相當數量的迴圈, 並記錄每次估計的結果, 迴圈結束後, 最終的估計值採最後 m 次估計值的平均數。至於總迴圈數與 m 值視情況而定, 沒有一定的數值可供參考。

實驗3: 實驗步驟同實驗1, 不過不再判斷樣本 y_i 的群組屬性

1. 先猜測兩常態群組的參數:$\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2$ 做爲第一次的估計。

2. 計算樣本 y_i 屬於群組 1 的機率

$$\gamma_{i1} = \frac{\pi_1 p(y_i|\mu_1, \sigma_1)}{\pi_1 p(y_i|\mu_1, \sigma_1) + (1-\pi_1) p(y_i|\mu_2, \sigma_2)}$$

3. 估計參數[4]

$$
\begin{aligned}
\pi_1 &= \frac{1}{N}\sum_{i=1}^{N}\gamma_{i1} \\
\mu_1 &= \frac{1}{N\pi_1}\sum_{i=1}^{N}\gamma_{i1}y_i \\
\mu_2 &= \frac{1}{N\pi_2}\sum_{i=1}^{N}\gamma_{i2}y_i = \frac{1}{N(1-\pi_1)}\sum_{i=1}^{N}(1-\gamma_{i1})y_i \\
\sigma_1^2 &= \frac{1}{N\pi_1}\sum_{i=1}^{N}\gamma_{i1}(y_i-\mu_1)^2 \\
\sigma_2^2 &= \frac{1}{N\pi_2}\sum_{i=1}^{N}\gamma_{i2}(y_i-\mu_2)^2
\end{aligned}
$$

重複步驟 2,3 直到所有估計值不再改變爲止。

三種方式的結果孰優孰劣, 從給定的樣本分數中其實是分辨不出來的, 因爲我們並不曉得眞正的參數是什麼。如果要瞭解不同方法間的優劣, 只好從模擬資料中去判別。資料是根據自己決定的參數創造出來的, 標準答案已經在那裡了, 估計的結果自然可以根據標準答案去衡量。試著自己產生模擬資料, 做出下列幾種比較

- 當兩個常態成分的比例比較懸殊時。
- 當兩個常態成分比較接近時, 即 μ_1 與 μ_2 較接近時。
- 當兩個常態成分的變異數相同或不同時。

做實驗前, 得先擬好實驗內容, 準備好數據, 才開始動手。

實驗前的練習:

[4]這裡的平均數可以稱爲「加權平均數。」

觀察混合常態的長相; 假設由兩個常態分配合成, 其合成密度函數爲

$$f_Y(y) = \pi_1 f_1(y; \mu_1, \sigma_1^2) + \pi_2 f_2(y; \mu_2, \sigma_2^2)$$

畫出下列混合常態的密度函數(如圖2所示), 可以利用 MATLAB subplot 的功能在一張圖上畫 6 個圖。(本題摘自參考文獻[1],p.113)

1. $\pi_1 = 0.5, \mu_1 = 0, \sigma_1 = 1, \pi_2 = 0.5, \mu_2 = 2, \sigma_2 = 1$
2. $\pi_1 = 0.25, \mu_1 = 0, \sigma_1 = 1, \pi_2 = 0.75, \mu_2 = 3, \sigma_2 = 1$
3. $\pi_1 = 0.8, \mu_1 = 1, \sigma_1 = 1, \pi_2 = 0.2, \mu_2 = 1, \sigma_2 = 4$
4. $\pi_1 = 0.6, \mu_1 = 0, \sigma_1 = 1, \pi_2 = 0.4, \mu_2 = 2, \sigma_2 = 2$
5. $\pi_1 = 0.9, \mu_1 = 0, \sigma_1 = 1, \pi_2 = 0.1, \mu_2 = 2.5, \sigma_2 = 0.2$
6. $\pi_1 = 0.6, \mu_1 = 0, \sigma_1 = 1, \pi_2 = 0.4, \mu_2 = 2.5, \sigma_2 = 1$

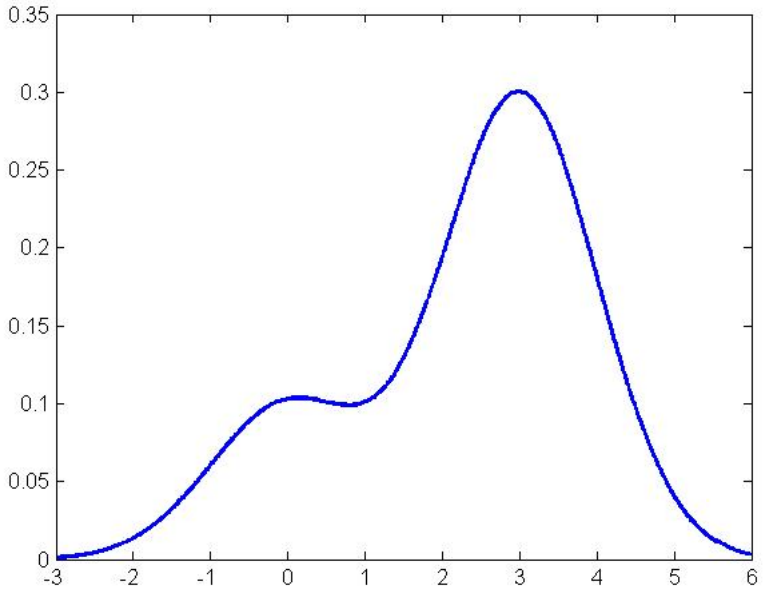

圖 2: 兩個常態混合的密度函數

参考文献

[1] B. Flury, "A First course in Multivariate Statistics," Springer.

[2] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning:Data Mining, Inference, and Prediction,"Springer