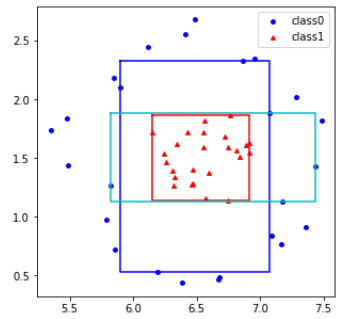
Machine Learning Homework 2

1. (a)

 一張含有 圖表, 螢幕擷取畫面, Rectangle, 行 的圖片

自動產生的描述一張含有 文字, 字型, 螢幕擷取畫面, 白色 的圖片

自動產生的描述

左圖：找到內層紅框(S)之後，往外延伸找外層藍框(G1、G2) 顏色較深的藍框(G1)是先將紅框(S)的左右邊界固定，往上下找第一個會碰到的點，再將這兩個點當作新的上下邊界，往左右再找第一個會碰到的點；而顏色較淺的青藍色框(G2)是將上下跟左右的順序調換來找。接下來較G1跟G2的範圍大小，選擇比範圍較大的那個(即G1)，因為在找most general hypothesis時會想要找到比較大的範圍能包含所有可能的class1極盡可能沒有class2。

右圖：將G和S相加取平均就得出青藍色框，取平均的目的是希望這個hypothesis能夠比較有彈性，而平均是在specific和general之間，因此錯誤的機率會相比單純採用specific或general來的小，因此選用這個方法。

(b)

一張含有 螢幕擷取畫面, 圖表, 圓形, 文字 的圖片

自動產生的描述

一張含有 文字, 字型, 螢幕擷取畫面, 白色 的圖片

自動產生的描述

用(a)小題中紅框的中心點作為圓心，對角線長度作為直徑畫圓。因為橢圓形較難控制涵蓋點的範圍，此法較為簡易，而根據結果來看，此分類方法還算有效，故選擇此方法作為分類依據。

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |

3.

(a)



從這些數據我們能看到每個feature的mean、variance等資訊，其中比較能夠從這些數據來分析結果的資訊大概只有mean，比如申請人數和錄取人數間的差異：平均而言，有3001人申請，但只有2018人被接受。這可能表示整體來看，大學的錄取率約為2/3；或者是高中成績排名前10%和前25%的學生百分比：平均而言，27.56%的學生屬於前10%，55.8%的學生屬於前25%，這可能表明這些學校在學術上具有一定的選擇性。但只根據數據也很難去理解他的分佈，因此接下來才會希望透過圖表方便我們觀察。

(b)

一張含有 文字, 螢幕擷取畫面, 圖表, 地圖 的圖片

自動產生的描述

從這些scatter plot中我們可以清楚看到兩個feature之間的分佈情形與關聯性，也可以觀察到極端值得存在。而這些feature的對應關係我們可以觀察到比如申請人數和註冊人數的關係：隨著申請人數的增加，註冊人數也有增加的趨勢，因為申請人數越多，理論上有更多的學生有機會註冊。而和Top10%有關的集中在左下，因為Top10%的學生也只佔全體10%，因此它的比例集中在比較低的地方也很合理。但從scatter plot我們只觀察到feature之間的關係。

(c)

一張含有 文字, 圖表, 行, 螢幕擷取畫面 的圖片

自動產生的描述

從box plot我們可以觀察到data的中位數、四分位數及極端值。而以下幾點是我的觀察：Elite學校的外州學費在整體上似乎比非Elite學校要高。這可以通過兩個box的中位數來看出，其中Elite學校的中位數高於非Elite學校。還有在兩組中都有極端值，但在非Elite學校中，這些極端值似乎特別高，遠高於普通範圍。Box plot展示了Elite和非Elite學校在學費上的顯著差異，並且即使在同一個分組內，學校之間的學費也可能有很大的差異。

(d)

一張含有 文字, 圖表, 螢幕擷取畫面, 繪圖 的圖片

自動產生的描述一張含有 圖表, 螢幕擷取畫面 的圖片

自動產生的描述

接下來我透過histogram去觀察data的分佈情形，也透過histogram去選定我想要用於model的feature。在這些圖表中我想要挑選分佈比較像normal distribution的feature，因為接下來使用的model是基於normal distribution，還有能從mean分出差異的圖，以及variance比較小的圖。首先像是Top25%、Outstate、PhD這些我認為就是屬於比較好的feature，因為以上都大致符合我上面所說的條件，而舉個反例perc.alumni我認為就不算好的feature，因為它的variance比較大，因此不適合做為分辨Elite的feature。

最後我選擇Top25%、Outstate、PhD以及Books作為我的feature，後面會再討論Books這個feature的影響。

(e)

一張含有 文字, 螢幕擷取畫面, 圖表, 繪圖 的圖片

自動產生的描述

(f)

一張含有 文字, 字型, 螢幕擷取畫面, 印刷術 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 行 的圖片

自動產生的描述

首先我使用Top25%、Outstate、PhD以及Books這四個feature套進Naïve Bayes Classifier之後，得出的準確率約為94%，從confusion matrix也可以看到False Positives，也就是模型錯誤預測為Elite類別的數量小於False Negatives，也就是模型錯誤預測為非Elite類別的數量。代表被誤判成非Elite的情況比較多。

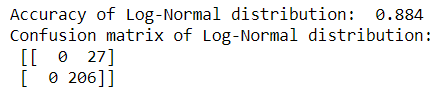
一張含有 文字, 字型, 螢幕擷取畫面, 印刷術 的圖片

自動產生的描述一張含有 文字, 字型, 行, 螢幕擷取畫面 的圖片

自動產生的描述

接下來我想討論Books這個feature對結果的影響。我將Books這個feature移除之後發現最後得出來的結果和它存在時是相同的，代表說Books這個feature存在與否其實對結果沒有影響，因此我們也能看出其實當feature選的足夠好的話其實model就能很準確地預測了，feature的多寡並不是直接和準確度做關聯。

(g)



因為在上題中normal distribution的準確度滿高的，因此在這題我使用了與normal distribution有關的lognormal distribution去做，但結果卻跟normal distribution相差甚遠。

判斷結果顯示全部的學校都是class 0，原因可能是因為lognormal比較適用於極端事件較多的數據集。因為lognormal distribution是由多個隨機變量相乘的結果，所以相乘後會產生出很多接近0或是非常大的數字。因此，可能是因為我選擇的feature中有很多偏小的隨機變量，造成相乘後的結果都接近0。