

監督式學習之 LDA、QDA 與 KNN 學習器

林哲兆

January 9, 2024

本文將介紹三種監督式學習器，三者的功用一樣都是將樣本進行分群，並檢視分類正確率，與迴歸不同的地方是，LDA、QDA、KNN 的決策邊界是利用後驗分配計算出來的，也就是在知道結果的情況下再對樣本進行分群預測，但 LDA、QDA 兩者與 KNN 計算後驗分配的方式有很大的差異，為此，本文將會利用五組資料檢驗三個學習器的優劣比較。

1 線性判別分析（LDA）

本小節將會介紹線性判別分析（LDA）的理論依據，並透過資料實作，展現在何種資料下，LDA 會有較好的表現；又在什麼情況下，LDA 紿出的結果比較差，以此讓讀者了解該將何種資料使用 LDA 進行分類。

1.1 理論背景

線性判別分析，顧名思義會利用線性的決策邊界進行分群，而讓決策邊界形成線性的原因，主要來自此模型對資料的兩個假設：

- 每個類別的數據都來自於同一個常態分配，但具有不同的均值。
- 每個常態分配的共變異數矩陣相同。

而決策邊界計算的方法是使用後驗分配計算，透過給定 x 計算出 x 屬於某組的機率，將 x 分類進後驗分配較高的組別。

$$Pr(G = k | X = x)$$

因為在計算後驗分配上比較困難，所以常會用貝氏方法，利用先驗分配計算出後驗分配：

$$P(G = k|X) = \frac{P(X|G = k)P(G = k)}{\sum_l P(X|G = l)P(G = l)}$$

在計算出後驗分配後，兩組後驗分配機率相同的地方就會是決策邊界，再透過對數轉換可以得到分界線函數，其數學式如下：

$$\begin{aligned} \ln \frac{Pr(G = k|X = x)}{Pr(G = l|X = x)} &= \ln \frac{f_k(x)}{f_l(x)} + \ln \frac{Pr(G = k)}{Pr(G = l)} \\ &= \ln \frac{Pr(G = k)}{Pr(G = l)} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) \\ &\quad + x^T \Sigma^{-1}(\mu_k - \mu_l) = 0 \end{aligned}$$

另外，由於 LDA 假設共變異數矩陣相同，所以在參數估計方面，僅需要估計一個共變異數矩陣，相對下一小節的 QDA，其模型簡化許多。本文將會利用四筆二元資料與一筆三元資料，來展示 LDA 的決策邊界與誤判率。

1.2 資料實作

第一筆資料是由套件 `make_blob` 所生成的二元分類資料，資料包含兩個特徵與該筆觀察值類別，其中兩群資料離散的程度較大，屬於很好分類的資料，以下是 LDA 模型產生的決策邊界，其錯誤分類率不出意外的是 0%：

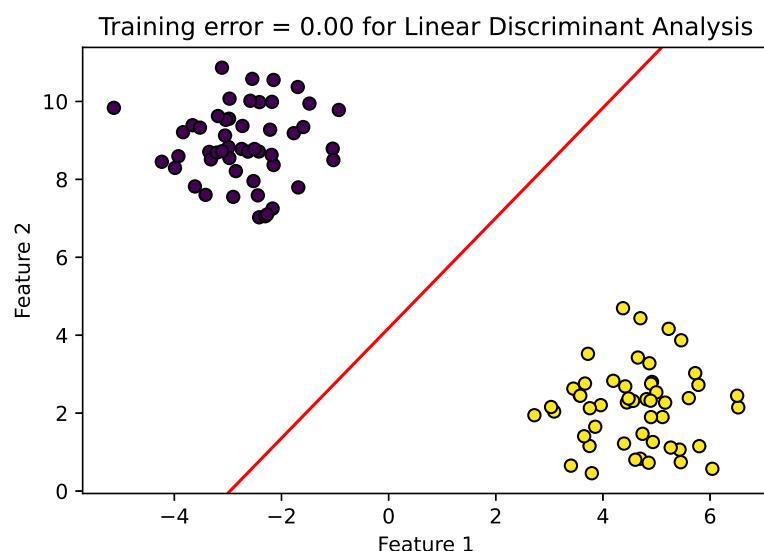


圖 1：資料離散程度大的分類結果

第二筆資料由兩組多變量常態分配亂數產生，其中為了展示在假設成立下，LDA 的分類狀況如何，因此把兩組多變量常態分配的共變異數矩陣設成相等，均值向量設置不同，而樣本數共 400，以下是多變量常態參數的設置與分類結果：

$$\mu_0 = \begin{bmatrix} 7 \\ 3 \end{bmatrix} \quad \mu_1 = \begin{bmatrix} -1 \\ -2 \end{bmatrix} \quad \Sigma_0 = \Sigma_1 = \begin{bmatrix} 15 & 1.2 \\ 1.2 & 20 \end{bmatrix}$$

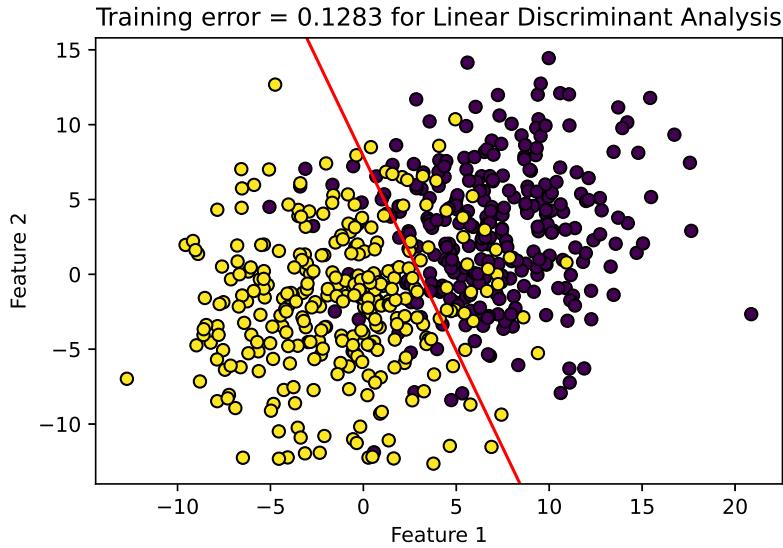


圖 2: 共變異數矩陣相同的資料分類結果

從圖 2 可以看出在共變異數矩陣相同下，兩個組別離散程度相同，因此用一條直線便可以將資料很好的分類，錯誤分類率是 12.68%，因此可以知道在假設成立的情況下，LDA 的分類表現是相當出色的。

第三筆資料同樣由兩組多變量常態分配亂數產生，與第一組資料不同的在共變異數矩陣，為了檢視在假設不成立下 LDA 的分類情況，因此特意把共變異數矩陣設置不同，但均值保持與前一筆資料相同，用以檢視假設成立與不成立時的差距，而樣本數共 400 以下是參數設置與分類結果：

$$\mu_0 = \begin{bmatrix} 7 \\ 3 \end{bmatrix} \quad \mu_1 = \begin{bmatrix} -1 \\ -2 \end{bmatrix} \quad \Sigma_0 = \begin{bmatrix} 15 & 1.2 \\ 1.2 & 20 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 5 & 2 \\ 2 & 60 \end{bmatrix}$$

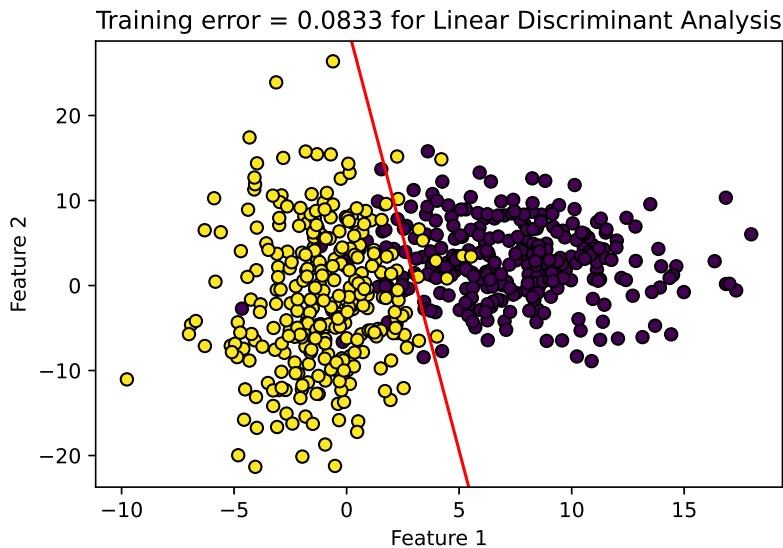


圖 3: 共變異數矩陣不相同的資料分類結果

從圖 3 可知在這筆資料的誤判率是 8.33%，特意生成一個共變異數矩陣不同的資料，為的是與下一小節的 QDA 做比較，因此關於這筆資料的結論會在下一小節說明。

最後一筆資料是來自 kaggle 的一份消費者購買行為資料，其特徵包括兩個：觀測者的年齡與薪資，類別則是其購買與否 (0、1)，樣本數為 400，因為前面都是自行生成的樣本，具教育性但不具實用性，因此特意找來一個比較實用的資料，以下是 LDA 的分類結果：

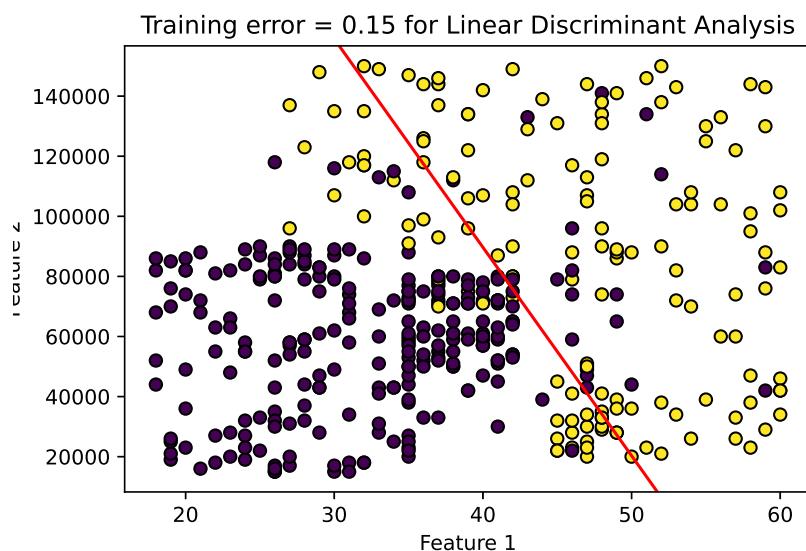


圖 4: 消費者購買行為資料分類結果

從圖 6 可以看出雖然錯誤分類率並不高，但這筆資料的兩個分類的界線屬於不規則型，因此用一條直線將兩者分類可能有點粗糙，LDA 或許不會是這筆資料最佳的分類器。

第五筆資料是一筆三群分類的資料，同樣來自多變量常態亂數，樣本數為 900，具有相同的共變異數矩陣，但均值向量不同，以下是參數設置與分類情況：

$$\mu_0 = \begin{bmatrix} -5 \\ 3 \end{bmatrix}, \mu_1 = \begin{bmatrix} -1 \\ -2 \end{bmatrix}, \mu_2 = \begin{bmatrix} 7 \\ 8 \end{bmatrix}, \Sigma_i = \begin{bmatrix} 5 & 2 \\ 2 & 18 \end{bmatrix}, i = 0 \sim 2$$

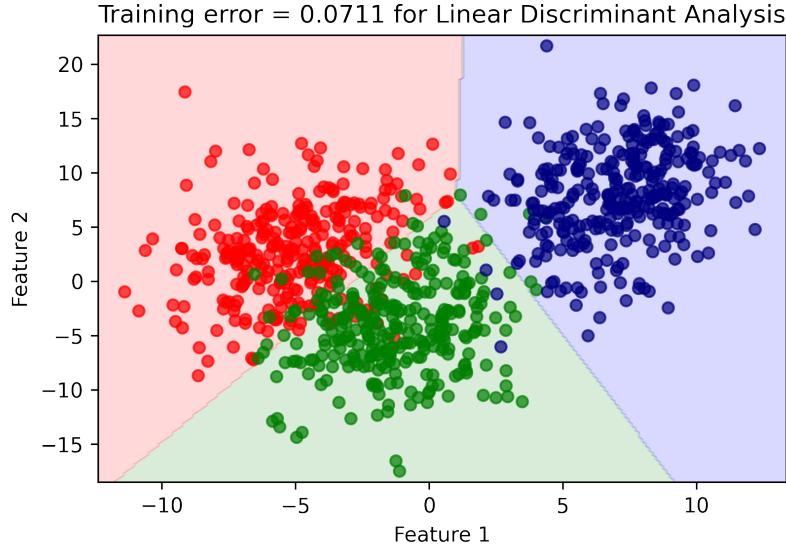


圖 5: 三種且共變異數矩陣相同資料分類結果

從圖 6 可以看出分類錯誤率為 0.711，並不算太高，但為了與假設不成立時的情況對比，所以下一筆資料會把共變異數矩陣設置成不相同，透過兩筆資料來檢視假設成立與否對三分類問題的影響，以下是第六筆資料的參數設置與分類結果：

$$\mu_0 = \begin{bmatrix} -5 \\ 3 \end{bmatrix}, \mu_1 = \begin{bmatrix} -1 \\ -2 \end{bmatrix}, \mu_2 = \begin{bmatrix} 7 \\ 8 \end{bmatrix}$$

$$\Sigma_0 = \begin{bmatrix} 15 & 1.2 \\ 1.2 & 20 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 5 & 2 \\ 2 & 18 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 10 & 0 \\ 0 & 5 \end{bmatrix}$$

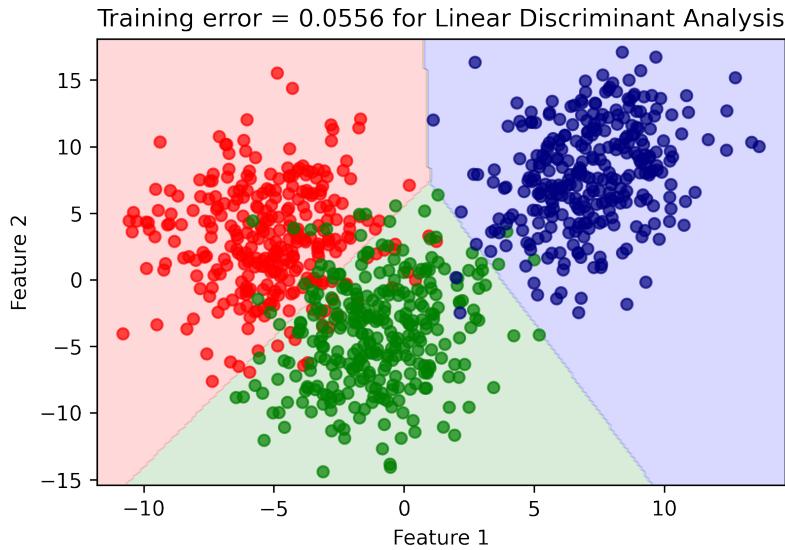


圖 6: 三種且共變異數矩陣不同資料分類結果

在共變異數不同的情況下，LDA 並沒有表現的更差，或許是亂數的資料邊界類似線性，使得即時假設不成立仍然表現的比假設成立時好，因此與第三筆資料相同，下一小節將會與接下來介紹的 QDA 比較，並做出結論。

2 二次判別分析 (QDA)

在介紹完 LDA 後，可以發現 LDA 存在著很強力的假設，並且決策分界較簡易，但當資料變得複雜時，LDA 必定無法很好的分類資料，因此本小節將介紹 LDA 的增廣版：二次判別分析 (QDA)。

2.1 理論背景

前一小節提到 LDA 有一個很強的假設一共變異數矩陣相同，但在二次判別分析中拿掉了這項假設，條件放寬了許多，但需要估計的參數將會增加很多，以下是 QDA 與 LDA 假設不同的地方：

- 每個類別的數據可來自不同多變量常態分配，具有不同的均值。
- 允許各個多變量常態分配的共變異數矩陣不同。

2.2 資料實作

在前一小節有說道，造成線性決策邊界的原因是共變異數矩陣相同的假設，因此當拿掉這項假設時，決策邊界會由線性變成非線性，在二維時會類似一個二次曲線，二次判別分析的名字也由此而來。本小節同樣會使用同樣的六筆資料，來展現 LDA 與 QDA 的比較。

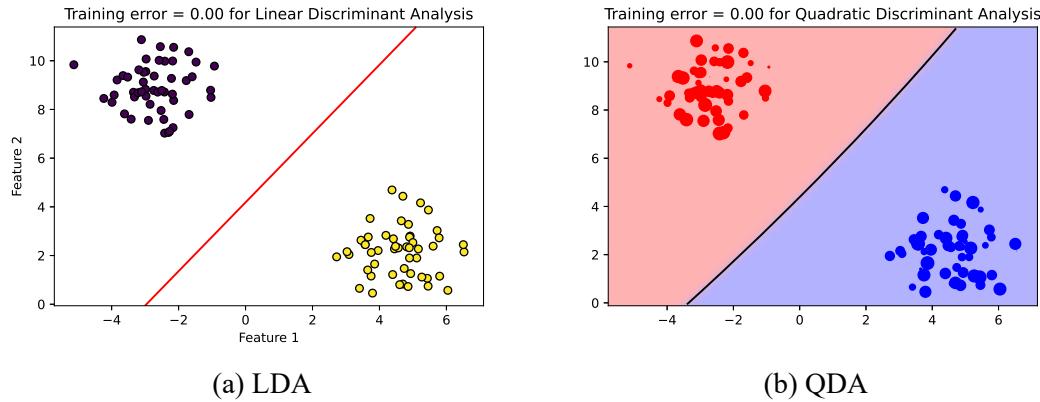


圖 7: LDA 與 QDA 比較

從圖 7 中可知，在資料比較簡單時兩者的決策邊界幾乎看不出差異，因此如果在需要進行分類的資料樣本較小、或是類別離散程度很大的時候，可以使用較簡單的 LDA 模型即可，可以在估計較少參數下達成目的。

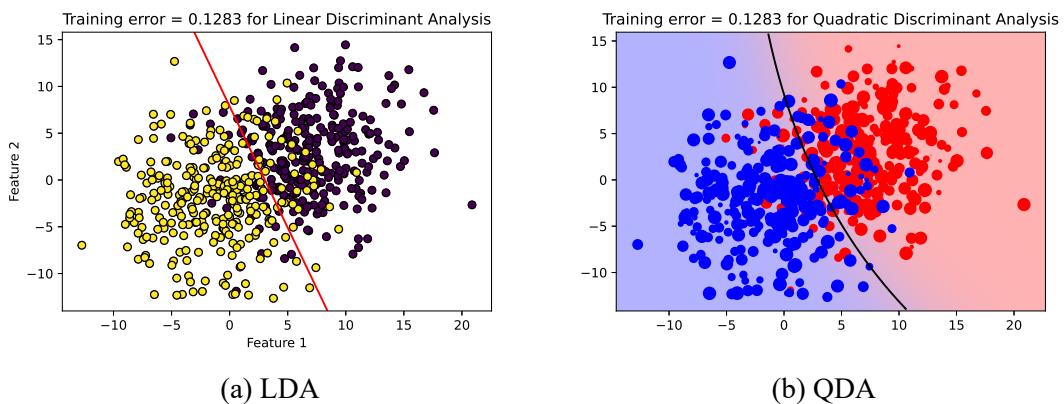


圖 8: LDA 與 QDA 比較 (共變異數矩陣相同)

圖 8 使用的是共變異數矩陣相同的亂數常態資料，可以看出兩者的決策邊界幾乎是一樣的，因為當共變異數假設成立時，二次判別函數會幾乎退化成線性判別函數，導致兩者的界線大致相同，因此可知在共變異數相同時，使用 LDA 即可，兩者的誤判率並不會有太大的差異。

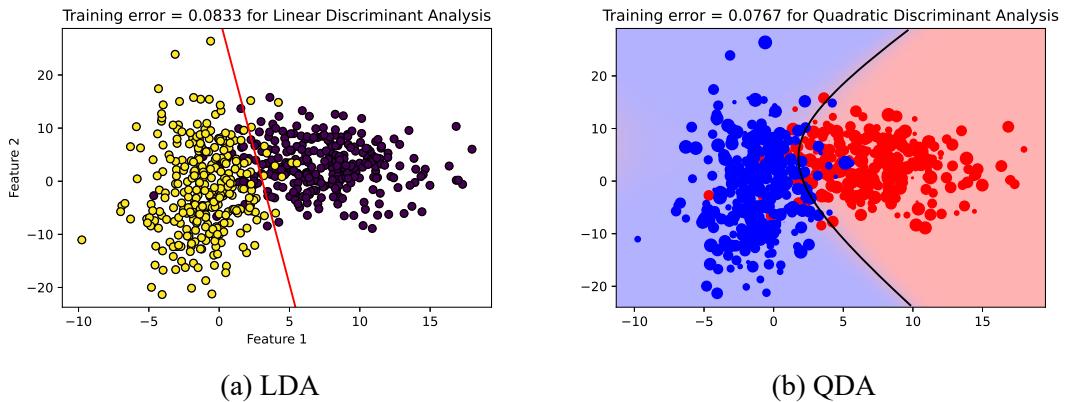


圖 9: LDA 與 QDA 比較 (共變異數矩陣不相同)

在圖 9 中顯然 QDA 誤判率勝過 LDA，可以看出 QDA 與 LDA 的決策邊界相差甚遠，QDA 的邊界更貼合資料，LDA 僅藉由一條直線分界，當資料量增加、複雜度增加時，將無法很好的進行預測，也間接證明了當共變異數矩陣不相同時，QDA 的表現會比 LDA 來的更好。

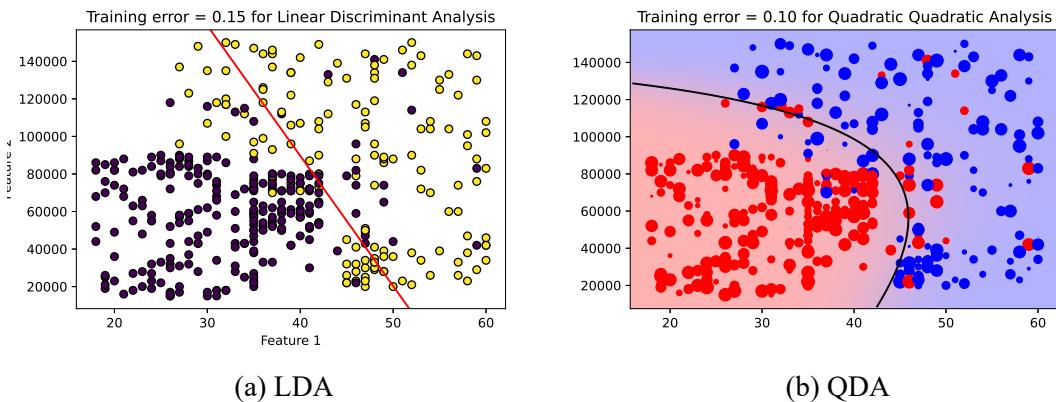


圖 10: LDA 與 QDA 比較 (消費者資料)

圖 10 是使用消費者資料的分類結果，在前一小節可以發現 LDA 並不能很好的分類這筆資料，因為資料分界並非線性，但透過 QDA 分類很好的貼合了資料，誤判率比 LDA 整整少了 0.05，是在所有資料中差異最顯著的，也證實了在複雜資料、分界線非線性時，QDA 相對 LDA 有較好的表現。

二分類的資料結束後，接著將同時展示兩筆三分類的資料，其中一筆是共變異數矩陣相同的常態亂數，另一筆則是不同的共變異數矩陣，分類結果如下：

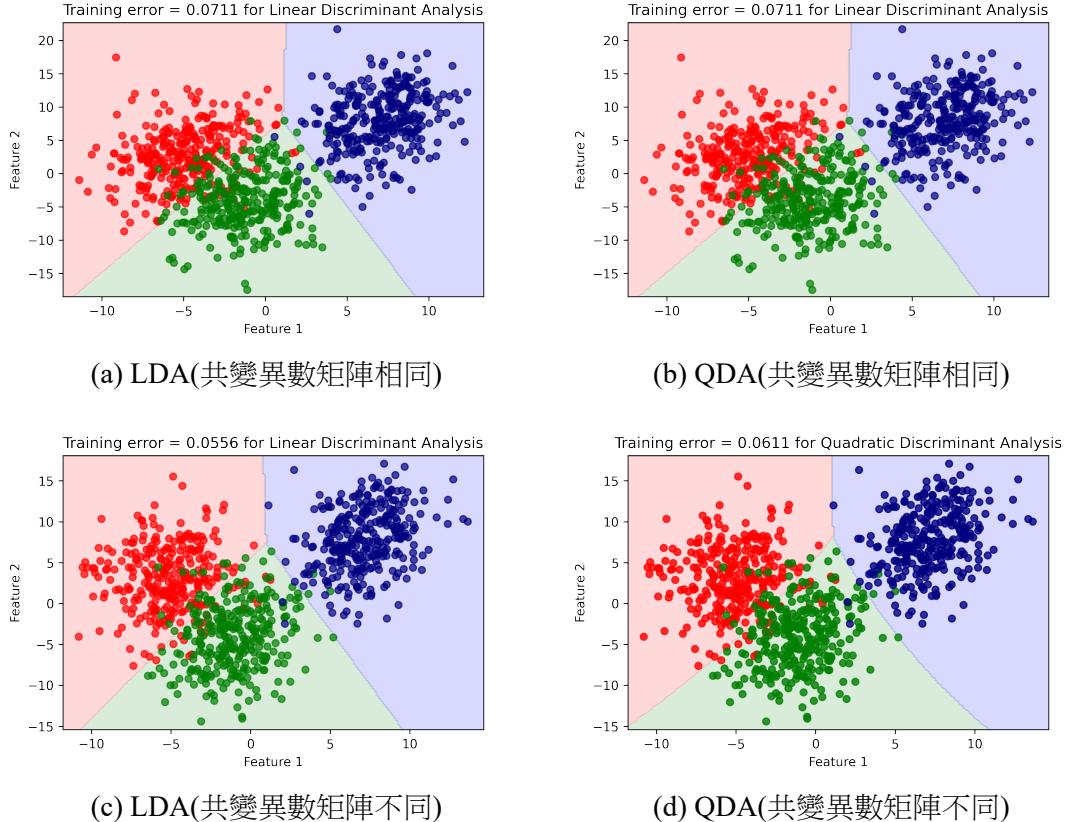


圖 11: 三分類問題

在三分類問題下，四者的差異並不大，推測是資料的邊界較趨近於線性，不太能展現 QDA 的優勢，常態亂數生成的樣本較難產生非線性的邊界，因此改進效果有限，但可以確定的是在共變異數矩陣不同時，QDA 仍是比 LDA 降低了 0.01 的誤判率。另外，可以看到 QDA 並不會出現比 LDA 表現得差的情況，可見 QDA 不管在資料複雜或簡單時都能很好的進行分類，但缺點是需要損失較多的自由度進行參數估計。

3 K-近鄰演算法 (KNN)

KNN 是一個淺顯易懂的分類器，透過設定適當的 K 值即可完成分類且容易解釋，與 LDA、QDA 相同的是，三者都是使用後驗分配計算出決策邊界，但 KNN 計算後驗分配的方式則與前兩者截然不同，因此本小節將會介紹 KNN 的

原理，並使用同樣的六筆資料進行實作，展示其與前兩者的差異性。

3.1 理論

KNN 並沒有對資料進行任何假設，其運作原理是透過捕捉觀察值周圍 K 個資料點的組別數據來計算後驗分配，舉例而言，若某觀察值周圍屬於組別 A 的資料點佔比最高，則會將其歸類為組別 A，而參數 K 就是用來決定要捕捉周圍幾個資料點，此種方法不存在繁雜的觀念，透過直觀的方法將樣本進行分類，以下是 KNN 的優點與缺點：

優點：

- 無需訓練：KNN 是一種基於實例的學習方法，它不需要對數據進行太多的訓練過程，只需儲存訓練數據，因此在訓練階段可以更快。
- 適用於多類別問題：KNN 可以輕鬆處理多類別分類問題。
- 適用於非線性數據：KNN 不對數據做出任何假設，可以處理非線性關係。

缺點：

- 計算複雜度高：在預測時需要計算新數據點與所有訓練數據點之間的距離，這可能導致計算複雜度較高。
- 對數據量和維度敏感：高維數據集或者具有大量特徵的數據會導致“維度災難”，影響演算法的性能。
- 數據不平衡問題：當類別不平衡時，KNN 往往會偏向於具有更多樣本的類別。

3.2 資料實作

在優點與缺點相當的情況下，該在什麼樣的資料下使用 KNN 更是一件重要的事，因此接下來將透過資料展示 KNN 的分群情況，並使用兩個不同的 K 值呈現參數差異 ($K = 5, K = 15$)。

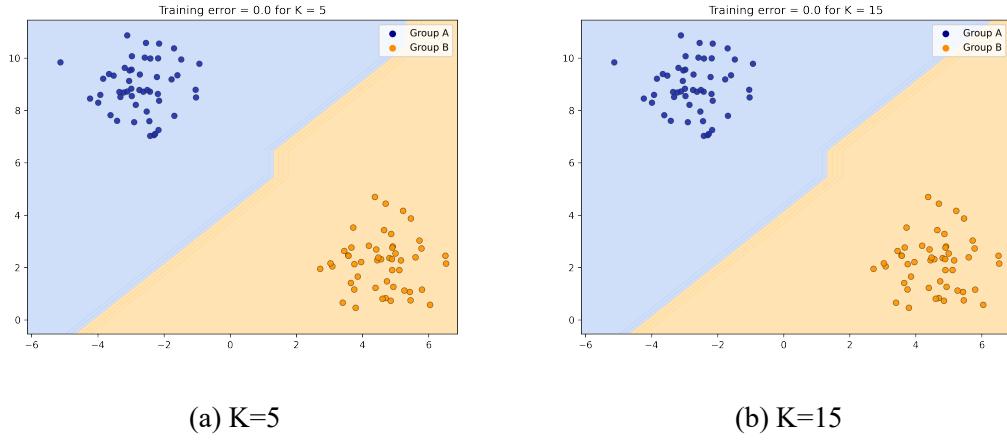


圖 12: 簡單資料分類

第一筆資料與前面的結果相同，都是完全準確的分類，唯一 KNN 與前兩者不同的是，其決策邊界並非線性或是二次曲線，而是不規則的形狀，可以從中間的鋸齒狀的看出來，這即是 KNN 與 LDA、QDA 在邊界上最大的不同。

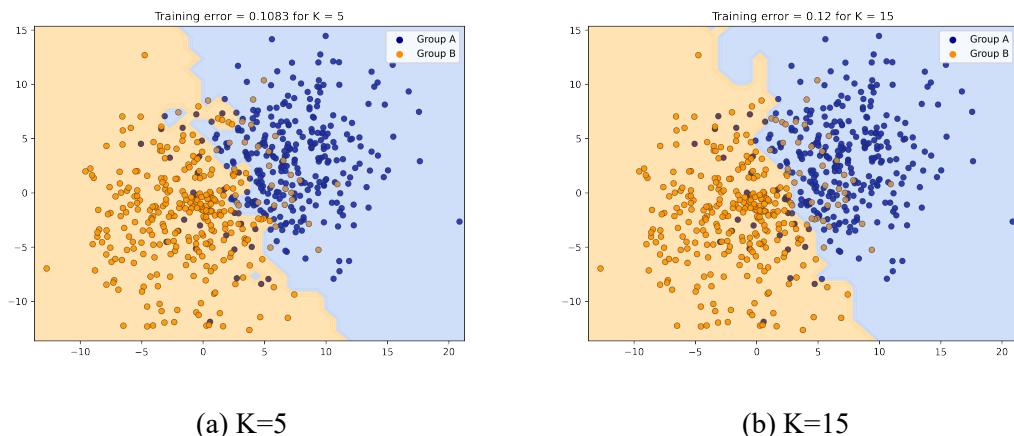


圖 13: 共變異數矩陣相同分類

第二筆資料是共變異數矩陣相同的常態亂數，從圖 13 可以看出雖然邊界類似，但 KNN 在 K=5 時的誤判率是比 K=15 還低的，這表示不代表 K 值取的越大就會越準確，因為取較大的 K 會受到比較多的干擾，如果資料不是相當複雜的話，取適當的 K 值即可。

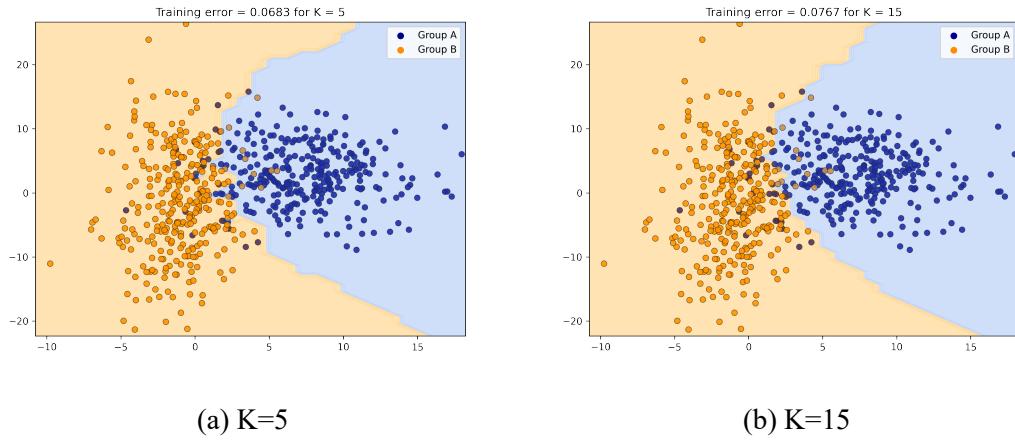


圖 14: 共變異數矩陣不同分類

第三筆資料是共變異數矩陣不同的常態亂數，從圖 14 可以看出 KNN 在這個資料的誤判率是稍微勝過 LDA(0.0867)、QDA(0.0833)，而 K=5 依然比 K=15 表現來的好，特別的是這四種情況的分類界線都不一樣，各有各的決策方法

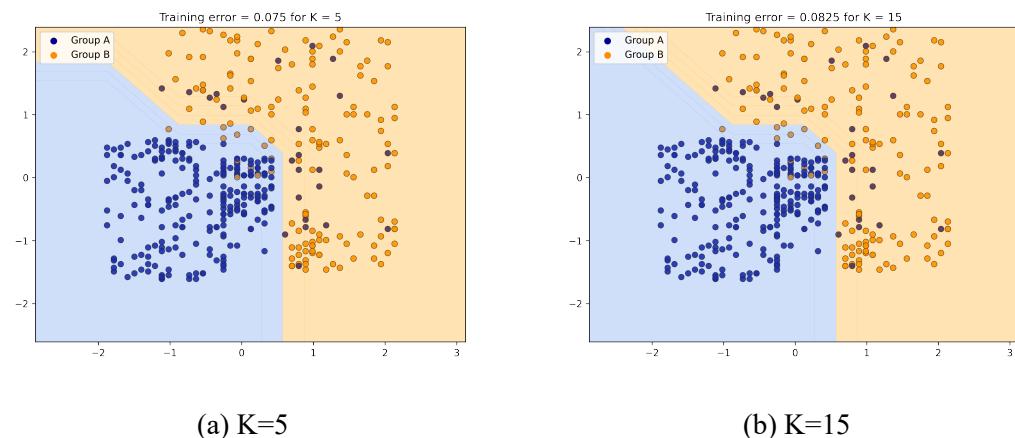


圖 15: 消費者資料分類

第四筆資料是消費者資料，可以看出 KNN 在這份資料中的誤判率遠勝過 LDA(0.15) 與 QDA(0.1)，不管是在 $K=5$ 或是 $K=15$ ，前一小節有提到這份資料是一個不規則形狀類似一個，其邊界類似一個鏡像 Γ 形狀，雖然利用 QDA 已經很好的改善 LDA 的分類結果，但判別分析 (DA) 終究是需要一條直線或平滑曲線將樣本分群，無法完全貼合這筆資料的鏡像 Γ 形狀邊界，因為即使是平滑曲線，其仍舊無法做到在某個點直接 90 度轉向，而 KNN 很好的做到了貼合資料分群邊界，驗證了在不規則邊界下，表現甚至會比能適應不規則邊界的 QDA 好。

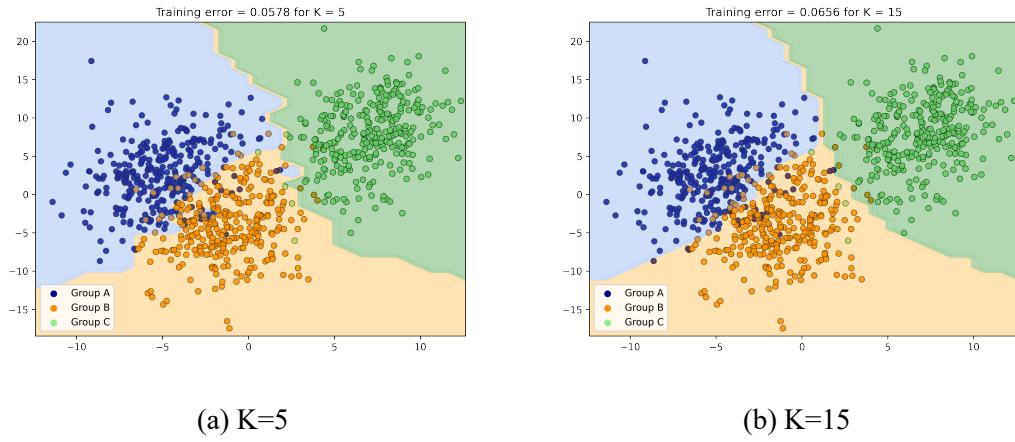


圖 16: 共變異數矩陣相同的三分類資料

圖 17 開始進入三分類的資料，這筆資料是相同共變異數矩陣的三分類樣本，無論是在 $K=5$ 或是 $K=15$ ，誤判率都比 LDA(0.0711)、QDA(0.0711) 還低，三者大致分界都類似一個倒 Y，但 KNN 在資料交會的地方做出較多正確的判斷，因此取得比較低的誤判率。

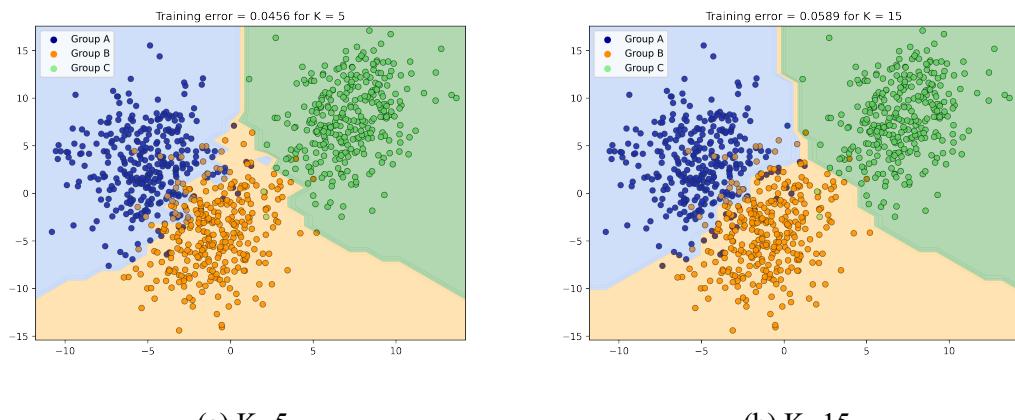


圖 17: 甘謙異數矩陣不同的三分類資料

依舊與第五筆資料的結果相同，誤判率都明顯比 LDA、QDA 還要低，證明了 KNN 確實不需要對資料進行假設，仍舊能有比較低的誤判率，也證明在低維度、資料比較簡單的情況，KNN 是一個很好的分類學習器。

上述六筆資料分類圖展示了 KNN 的特色，也就是非線性的邊界，在大部分的資料中，表現都比兩個判別分析來的出色，其中 $K=5$ 又比 $K=15$ 出色，但有時非線性邊界表現不一定比較好，仍需要視樣本各類別的資料特徵而定。

4 結論

本文介紹三種透過後驗分配計算決策邊界的學習器，三者特色皆有不同，以下利用表格展示三種模型在二分類與三分類資料的誤判率：

表 1: 三種模型在四筆二分類資料的誤判率

模型	共變異數矩陣相同	共變異數矩陣不同	消費者資料
LDA	0.1268	0.0867	0.15
QDA	0.1268	0.0833	0.10
KNN(K=5)	0.1083*	0.0717*	0.075*
KNN(K=15)	0.12	0.0883	0.0825

在二分類資料中，KNN 很顯然的贏過 LDA 與 QDA，而 QDA 又略勝於 LDA，但前三筆資料中差異並不大，可以說三者分類能力在分界明顯、資料離散程度大的時候，正確率是差不多的，因此在正確率差不多的情況下，使用比較簡易、易解釋的 LDA 或許會更方便。但當資料為非線性分類邊界時，KNN 的正確率肯定是會比 LDA、QDA 還要優秀的，如同表 1 消費者資料該欄呈現，KNN 的誤判率甚至只有 LDA 的一半，顯然 KNN 對資料的限制並不大。在 K 值選擇上，由於資料並不複雜，因此 K 選擇 5 即可，選擇過大的 K 在這四份資料中反而造成誤判率更高，所以在實際分析資料時，需反覆嘗試出適合的 K 值，K 值的大小並不保證正確率。

表 2: 三種模型在兩筆二分類資料的誤判率

模型	共變異數矩陣相同	共變異數矩陣不同
LDA	0.0711	0.0566
QDA	0.0711	0.0611
KNN(K=5)	0.0578	0.0456
KNN(K=15)	0.0656	0.0589

在三分類資料中，結果大致上與二分類結果相同，都是 KNN 表現最出色，但在共變異數矩陣不同時，QDA 的表現竟然是比 LDA 差的，推論應該是資料產生

時邊界就已呈現線性，因此使用 QDA 不會比 LDA 更有效率。KNN 則是與前面的結果相同，選擇 K=5 即可。

藉由三者的比較希望能讓讀者更了解 LDA、QDA 與 KNN 的不同與相同之處，或是在什麼情況該使用何者，畢竟分類問題沒有準確的答案，需要的是一次又一次的嘗試，才能找出屬於該筆資料最適合的分類學習器。