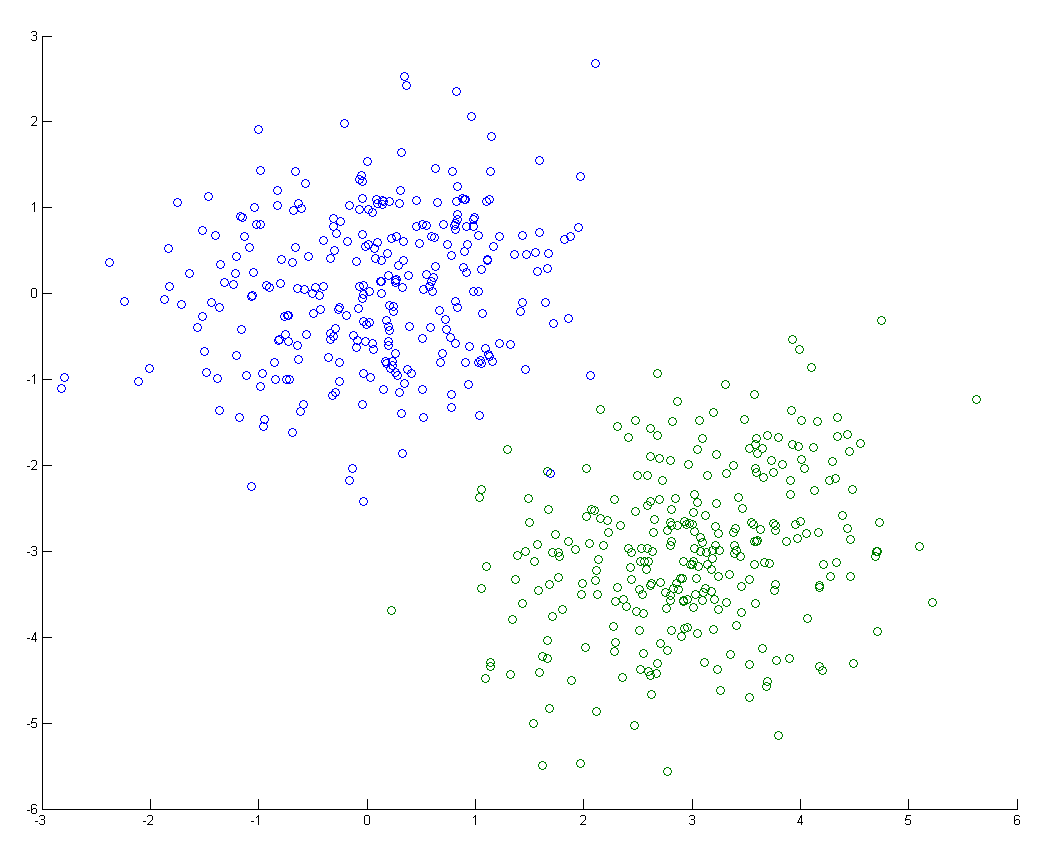
Machine Learning – Assignment 5

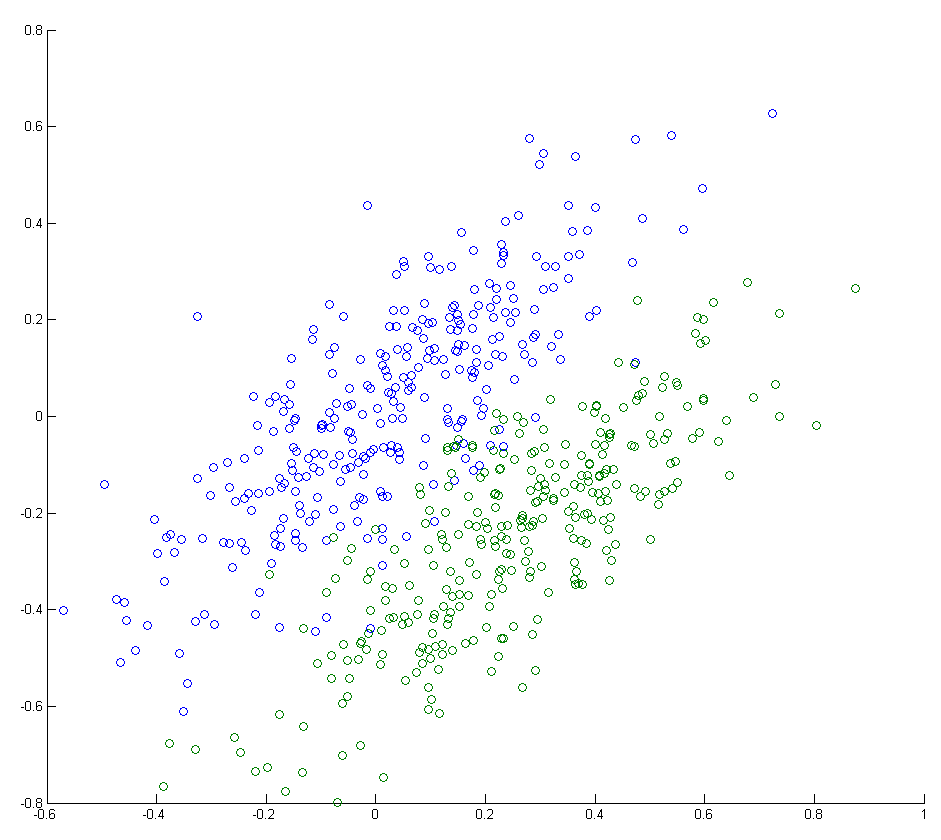
103062528 林玉山

1. **假設d=2，畫出兩組Normal Distribution的Data Set以符合下述兩種狀況**
2. **做PCA得到的vector，與LDA得到的vector具有相同方向**

****

PCA會盡可能地將使資料點投影到新的vector上時能夠區別開來，因此會選擇上圖軸藍軸的方向。LDA則會盡可能地將「不同群」的資料點分開，因此也會趨向於藍軸。最後兩種方式都會獲得藍軸這個vector。

1. **做PCA得到的vector，與LDA得到的vector互相正交**

****

PCA為了盡可能使資料點分開，而這份資料投射到藍軸後比較容易分別，因此會採用藍軸。不過LDA為了能夠使不同群資料分開，也就是盡可能地讓藍色與綠色在新軸上也能夠區別，所以會趨向於使用紅軸。

若將上述資料視為橢圓，PCA會趨向於取長軸。如果資料分群的界線剛好是短軸的話，那就會產生PCA與LDA選的軸正交的狀況。

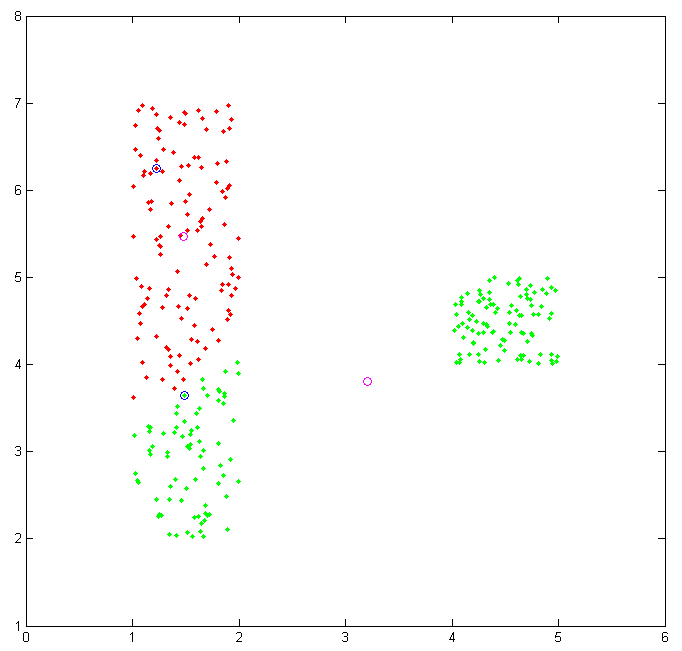
1. **延伸現有的Kernel PCA，使之可以不用center**

(不會寫，來不及複習QQ)

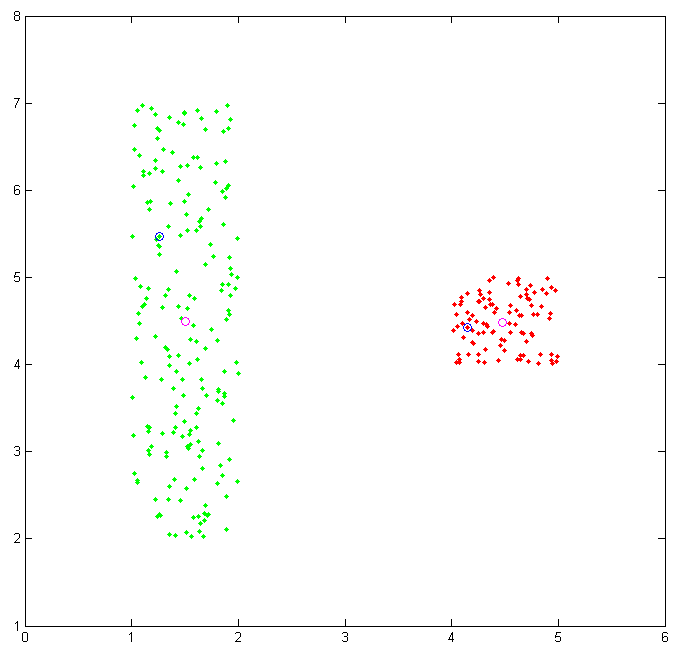
1. **令d=2, k=2，畫出一張一開始的mean選得不好的圖，以及一張選得好的圖**

下兩張圖為同一組data set，分為紅色與綠色兩群。藍色圈圈為一開始的means，紫色為最後收斂時的means。

第一張圖為一開始的兩個means都從左側那群點中取出。從最後收斂的結果可以看出來，左半部的下半段被分進了右邊那群中。很明顯就不是我們要的結果。



第二張圖則一開始就從左右兩群個挑一個點出來做為初始值。可以看見這樣的情況下，左右兩群就被正確地分開了。並且最後得到的means，也正好在兩群的正中央。正是我們期望的結果。由此可以得到「一開始挑選的點會大大影響結果」這個結論。



1. **為什麼在Hierarchical Clustering中，我們可以捨去部分的subtree？**

在Hierarchical Clustering中，當我們選擇了一個distance，h作為斷點，切出了數個clusters之後。那些數量極少的clusters，之所以沒有被併進其他clusters中，是因為他們距離那幾個clusters都有至少h以上的距離。因此就算我們把這些數量極少的clusters都去除掉，也不會影響其他clusters的結果，因為他們的距離都太遠，不足以影響到較大的clusters。

1. **令有一 Graph Laplacian Matrix L，請證明下列性質**

**(a) 證明對於任意向量 ，**

首先，我們先計算L

接著，我們展開

當我們計算綠色部分乘上L時，我們令每一個column的結果為

此時 變為：

我們再將 各自代入，其中對於每一個

因此 為：

在上式中，我們讓左式(綠色)與右式之中，index相反的項相加，可以得到下式

又因為矩陣S是Symmetric的，所以對於

將t換成j即為原式，故得證。

**(b) 試證明**

目前已知 (注意 是純量)：

接著對於第i與第j個data point，一共會有四種情形：

1. i與j同屬於G
2. i屬於G，j不屬於G
3. i不屬於G，j屬於G
4. i與j皆不屬於G

因此我們可以依照四種將原式中的 拆開：

此時代入下式

因為 ，因此

得證。