Report of Homework 2

姓名: 吳耿暉

學號: 0556171

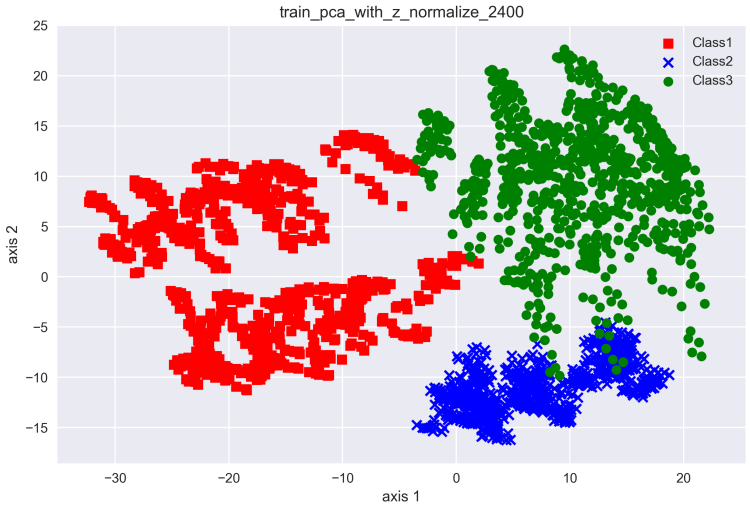
這篇報告會先分為Balanced Data以及Unbalanced Data，接著再往下討論作業程式所需的三個部分：PCA，Generative Model以及Discriminative Model，而每個部分又再分為有無進行z-normalization做比較。z-normalization就是將資料的平均移到0處, 並每個feature除以整體feature的standard deviation。

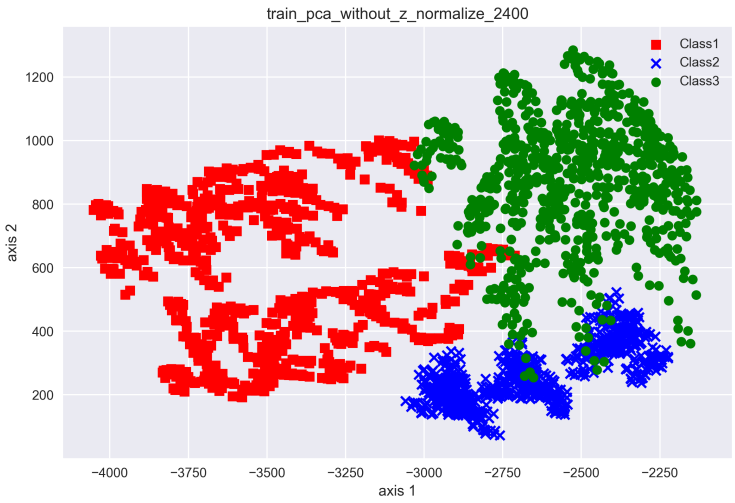
Part I: Balanced Data

這部分使用每個class各800筆資料當作training data, 剩下的200筆則為validation。

1. Principal Component Analysis(PCA)

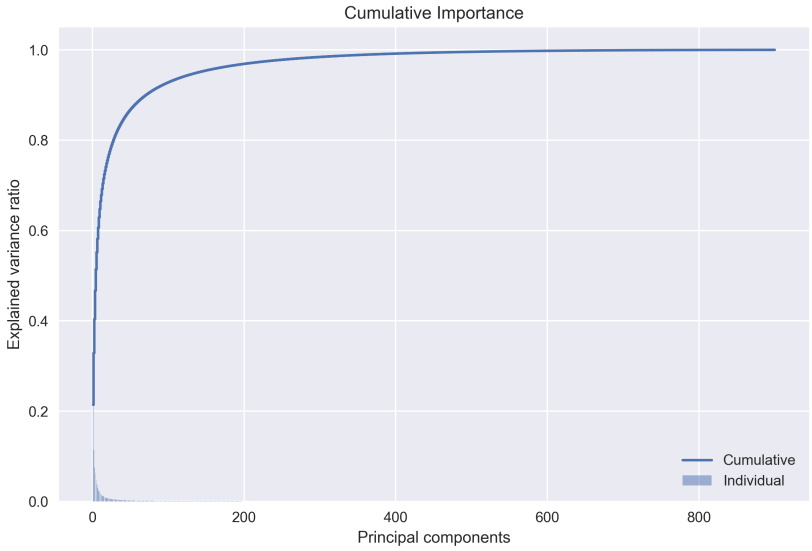
結果如下:





左圖為經過z-normalization的結果, 而右邊則未經過z-normalization, 可以看出在邊緣上經過z-normalization後的資料點經過降維之後分開的距離有稍微比為經過z-normalization的結果遠, 從這邊看起來選擇z-normalization的降維是比較好的。

另外在每個eigenvector的重要性圖累積如右, 這是將每個eigenvalue都除以eigenvalue總合後的以下累積圖。可以看出在前幾個特徵後已經可以對數據有不錯的解釋能力, 重要度前10名累積為[0.214 0.329 0.403 0.467 0.514 0.552 0.582 0.606 0.629 0.647]。



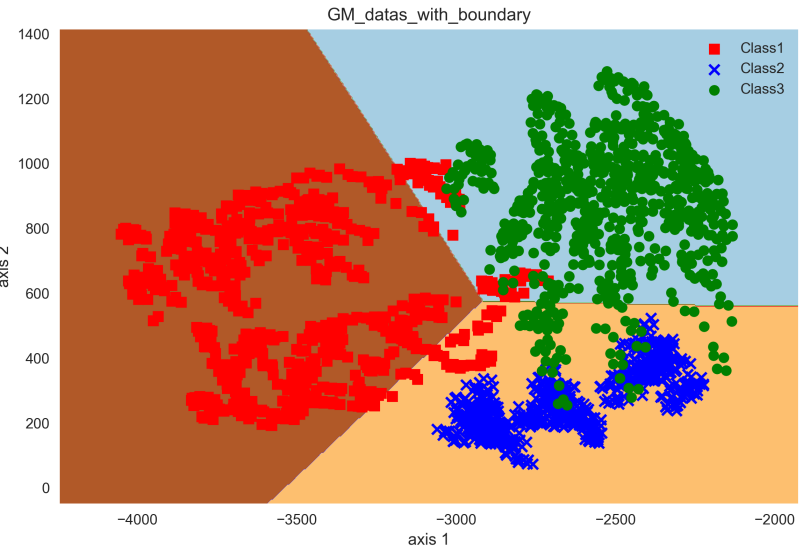
2. Probabilistic Generative Model

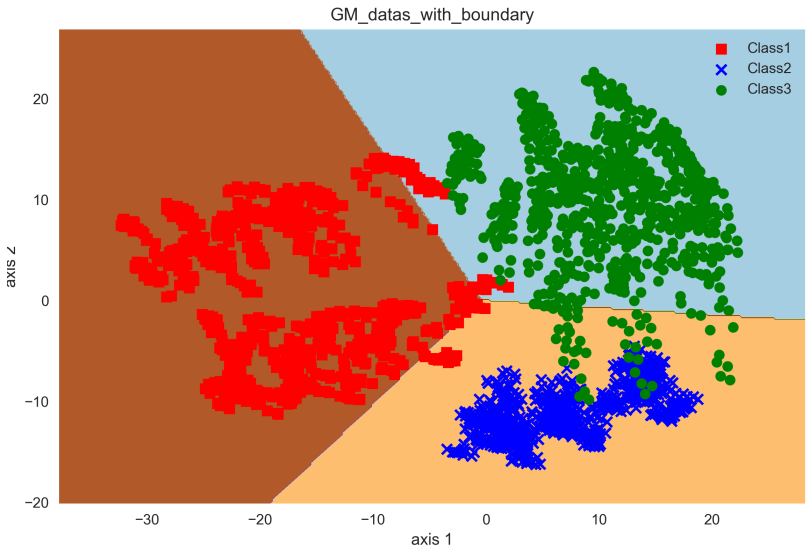
$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{\sum_j p(x|C_j)p(C_j)} \qquad p(x|C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\}$$

在Generative Model中為了求我們需要先計算prior, 這邊簡單的使用training data的資料分布當作每個class出現的機率, 由此來當作prior來進行計算, 而其他部分就只是非常單純的計算了。

下左圖為未經z-normalization的結果, training/validation accuracy為0.928和0.94;

下右則為經z-normalization的結果, training/validation accuracy為0.95和0.97, 雖然, 這邊也可看出資料經過z-normalization再經PCA後得到的的結果基本上是優於不做z-normalization的結果的。





不過比較不常見的就是validation accuracy比training accuracy還要高, 猜測應該是由於training data包含的邊界資料比例較高所造成的。

3. Probabilistic Discriminative Model

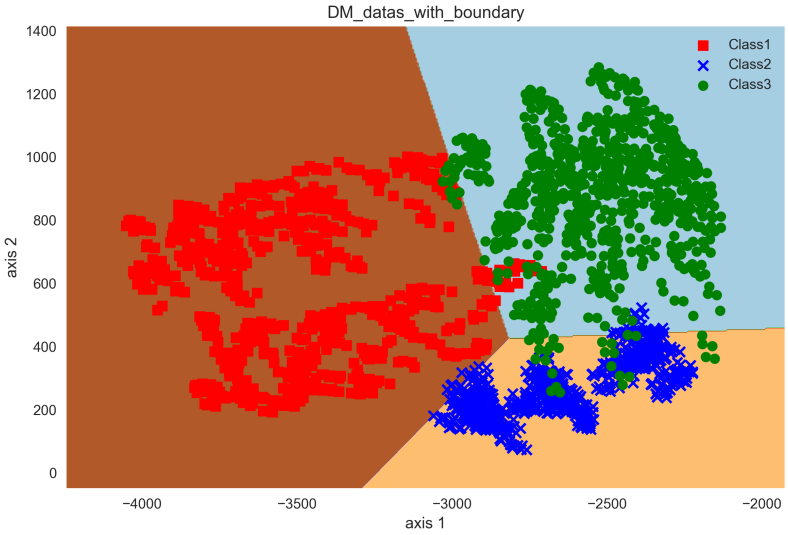
$$p(C_k|x) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \qquad a_k = w_k^T x$$

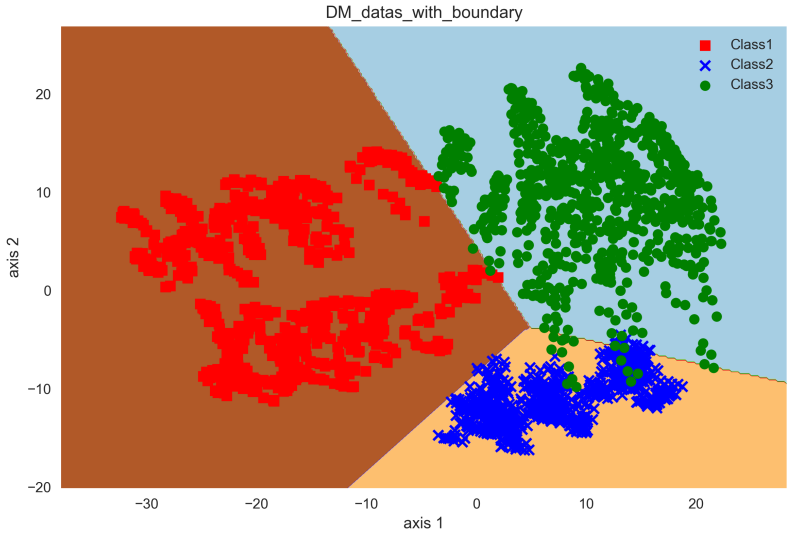
預測的結果就是softmax的輸出結果, 這邊在實作的時候遇到了數值問題, 後來利用所有值都先減去最大值後再去取exponential就可以解決。此處的初始設置如下：

converge條件:0.003, 最大iteration次數: 50, weight初始皆為0。

下左圖為未經z-normalization的結果, training/validation accuracy為0.96和0.945;

下右則為經z-normalization的結果, training/validation accuracy為0.983和0.973。





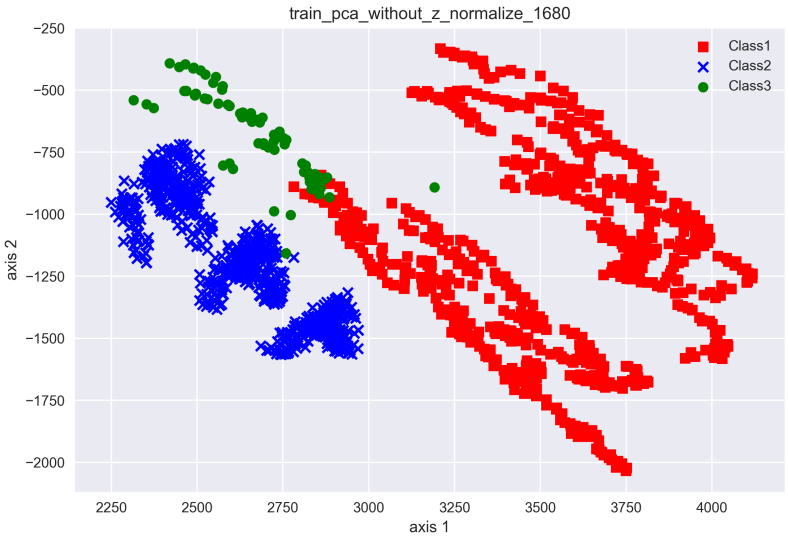
另外有嘗試其另一種initialization方式, 第一個從mean為0, std為1之normal distribution取值後再除掉sqrt(n), 其中n為資料數, 這個方式是由於在輸入data越多時, 一般來說output的variance也會增加, 利用除以數據量的平方根來調整數值範圍可以讓output的variance縮小, idea來自於"http://cs231n.github.io/neural-networks-2/"。用這個方法則得到training/validation accuracy分別為0.982和0.977, 得到了些微的改善。

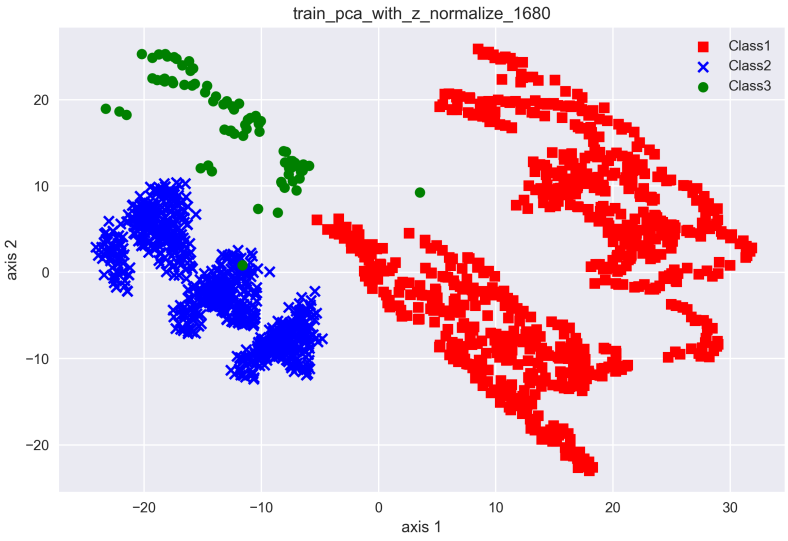
Part II: Unbalanced Data

這部分class1和class2一樣各取800筆資料當作training data, 剩下的200筆則為validation。但class3則只取80筆資料作為training data和20筆資料當作validation data。

1. Principal Component Analysis(PCA)

結果如下:

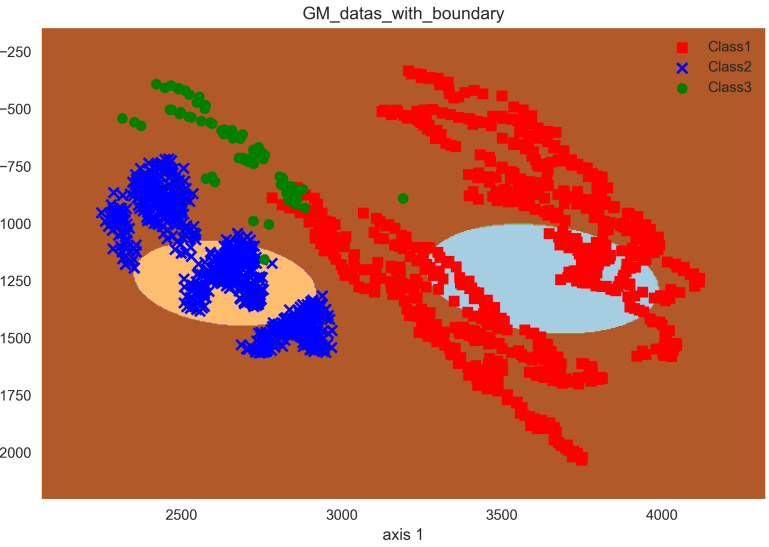


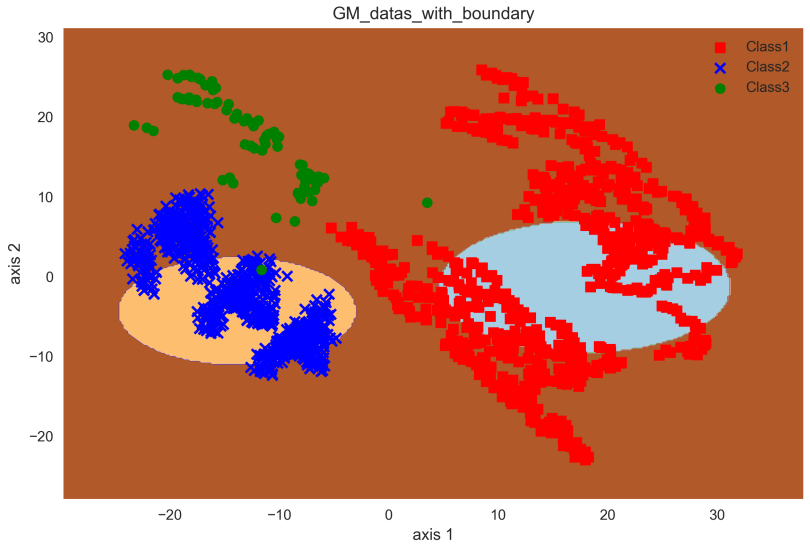


左圖為未經而右圖為經過z-normalization的結果, 從這邊可以更明顯地看到經過z-normalization的資料分開的更好了，但在unbalanced data下讓eigenvalue/eigenvector產生了改變, 導致出來的圖與balanced的非常不同。

2. Probabilistic Generative Model

結果如下，左邊為未經而右邊為經過z-normalization的結果。



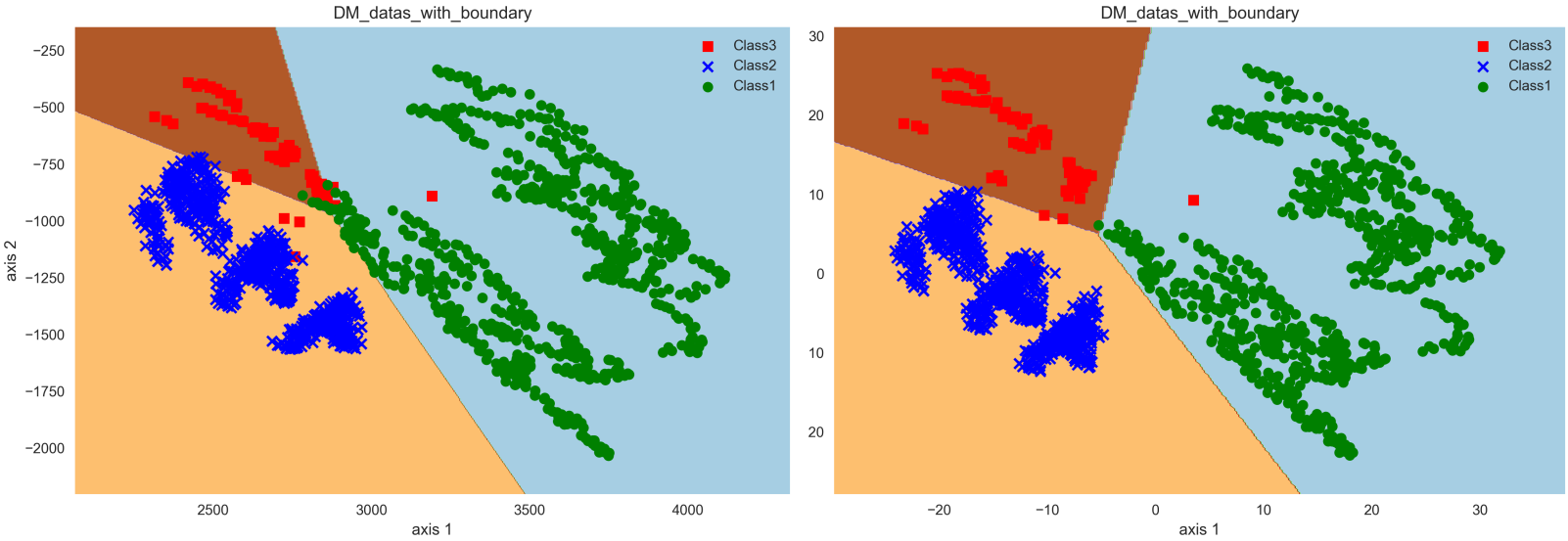


未經z-normalization的training/validation accuracy則為為0.29和0.274; 而經z-normalization得到training/validation accuracy分別為0.495和0.459, 基本上都是非常的差, 由於generative model是從training data來推斷資料的分佈, 可以看出在錯誤的prior下, generative model會得到非常糟糕的結果。

3. Probabilistic Discriminative Model

這邊使用上面所提到的高斯分佈除以sqrt(資料數)的方式對weights進行初始化, 結果如下，左邊為未經而右邊為經過z-normalization的結果。

未經z-normalization的training/validation accuracy則為為0.29和0.274; 而經z-normalization得到training/validation accuracy分別為0.998和1.0(!!!)。



可以看出Discriminative Model相較於Generative Model而言更加的robust, 在un-balanced data的情況下還能夠有很好的表現。

Summary

1. 在使用PCA之前, 對資料進行z-normalization能夠更有效地在低維度把資料分開。
2. Generative Model的prior非常重要, 這代表了我們的training data必須要有能夠反映真實資料的能力, 如果沒有的話會產生糟糕的結果。但是若有好的training data時就能夠非常快速的建立好model並且產生資料。
3. Discriminative Model在預測上較為robust, 在此例上準確度也較高, 但訓練上較為緩慢, 並且沒有辦法產生新的資料。