Machine Learning _ Hw#2　R04942131_蔡銘穎

(a) Approach and Discussion

本次作業要求我們辨識垃圾郵件，我以上課所教的兩個方法實作，分別是 Logistic regression 和 Probabilistic Generative model，在做 Logistic regression 之前，我先把資料做了 feature scaling，目的是讓我的 model 在 training 的時候，減少我的收斂時間，而這個 preprocessing 的動作確實是非常有效的，最後僅做了 6 次 iteration 就能收斂。而 Logistic regression 與上次實作的 Linear regression 最大的不同就是其 input 必須先經過 sigmoid function，這是一種增強我們 input data 的方式，第二個不同點則是 Loss function，比較特別的是我們必須先判斷預測結果為何，再代回對應的 Loss function 部分，否則會造成計算 log0 的問題。其中有遇到準確率一直無法上升的問題，後來發現在 training data 上做了 feature scaling 之後，必須把 mean 及 standard deviation 記錄下來，並在 testing data 做一樣的 feature scaling 才對，最後在 Kaggle 上 testing 的準確率為 0.92。

Probabilistic Generative model 的部分較為簡單，需要先算出事前機率 P(0)以及 P(1)，在使用機率模型預測其結果即可，在這邊我使用的模型為上課所用的 Gaussian distribution，好處是它的變數 mean 及 covariance matrix 並不難處理，而且已推得其最佳解。

這個方法我使用了每一組 data 的所有資料當作我的 feature，也就是說 mean vector 和 covariance matrix 維度分別是 1x47 和 47x47，Probabilistic Generative model 實作起來並沒有特別困難，在過程中遇到的問題只有一開始還沒共用 covariance matrix 前，class1 的 covariance matrix 是 singular 的，這會造成我們無法得到 Gaussian distribution，因為其必須求 Σ^{-1}，但最後共用的時候問題就解決了。

最後在 Kaggle 上 testing 的準確率為 0.79。