ML Report 2

學號：B04611015 系級： 資工二 姓名：陳佳佑

1.請說明你實作的generative model，其訓練方式和準確率為何？
答：
雖然除了前六項以外的維度都是binary data，但是我還是將其當成高斯分布，並將全部維度的data代入教授投影片上的公式。此外，我有嘗試過分別使用covariance matrix，很不幸的，有一邊的covariance matrix的det是0，而且還是singular matrix，因此還是採用教授的做法。此外如果直接用高斯分布的公式推導，很容易因為浮點數的underflow造成誤差。最後，我還是採用教授最後一頁的公式，並且沒有遇到任何underflow 與 overflow 的 issue。 Accuracy : 0.84238 (kaggle)

2.請說明你實作的discriminative model，其訓練方式和準確率為何？
答：
首先我單純對data做scaling (除以max)，加入所有的維度，然後透過gradient descent 做full batch。結果是還不錯，至少過了simple baseline (Accuracy 0.85074 Kaggle)。但之後，strong baseline 出來後居然沒過。所以就開始增加model的複雜度，在增加了前六項的0.5, 2, 3, 4次方後，model終於闖過了strong baseline。此外，因為怕造成overfitting，所以我有加regularization。但是之後，分別送了，有regularization的版本與沒有的版本，發現相差不大。 Accuracy : 0.85651 (kaggle)

3.請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。
答：

discriminative model

因為normalize 後，會有值是負的，所以只多加 2, 3, 4次方之資料。
feature normalization : 0.85651 (kaggle best) scaling: 0.85516 (kaggle)

基本上，因為大部分的維度都是binary distribution，所以normalization的意義並沒有很大，所以在此情況下，做normalization的額外意義就是避免overflow還有平衡一點，前幾項的影響力。換而言之，scaling 其實也達成了類似的效果，所以二者的差距才不會很大

generative model

feature normalization : 0.84238 (kaggle) original : 0.84165 (kaggle)

同上discriminative model之論述，因為binary data實在太多，所以影響實在有限，若將後面的資料當成bernoulli distribution，或許會有較大的差別。

4. 請實作logistic regression的正規化(regularization)，並討論其對於你的模型準確率的影響。
答：
如同在題 1.中，為了避免對於logistic regression的model產生overfitting，我嘗試過使用 regularization (lambda值從 1,10,100)，但是幾乎沒有效果，因此放棄使用。我猜想，有可能是因為這次的 training data內的noise很少，造成此一效果。

5.請討論你認為哪個attribute對結果影響最大？
我認為在整個過程中，discriminative model 與 generative model 之選擇，對於結果影響最大。在一開始的discriminative model中，只是用最單純的模型，即可獲得0.85的準確率，並且在增加維度後，很快就獲得顯著的提升。然而對於generative model而言，就算是增加了許多維度，對於結果的影響還是不大。此外，generative model的瓶頸應該是假設的distribution錯誤，如果嘗試看看不同的distribution，應該會有更好進步。