學號：b04902053 系級： 資工二 姓名：鄭淵仁

1. 請說明你實作的 generative model，其訓練方式和準確率為何？

答：

- 訓練方式：

我把資料標準化，之後再依照上課投影片裡面的公式進行計算，也就是使用了 Gaussian distribution 的機率模型並在 naïve Bayes 的假設下訓練資料。

- 準確率：

我自己在資料上切 validation set 來測試，得到的準確率是 0.841164547632。而在 kaggle 上面的 public score 是 0.84128，private score 是 0.84633。

2. 請說明你實作的 discriminative model，其訓練方式和準確率為何？

答：

- 訓練方式：

我取助教提供的 X_train 裡面所有的 attribute 的 1 維、2 維、3 維以及 sin 函式做為 feature，並把資料標準化，然後跑 5000 次 regression。

- 準確率：

我自己在資料上切 validation set 來測試，得到的準確率是 0.857564031693。而在 kaggle 上面的 public score 是 0.85700，private score 是 0.85874。

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：

- **先將 feature 標準化，再實作 probability generative：**

我自己在資料上切 validation set 來測試，發現如果不先把資料標準化的話，準確是 0.841440943431；如果先把資料標準化的話，準確率會是 0.841256679565。

從上述數據可以看出：有沒有標準化對於 probability generative 的實作結果幾乎沒有影響。

我想原因是因為把資料標準化，不會影響到 feature 和結果的機率分布。如果把 feature 和結果的機率分布想成多維度的圖的話，「把資料標準化」這個動作，只是將這個圖延著不同的維度壓縮、延伸而已，而這個圖上面的點也會跟著壓縮、延伸，結果還會是一樣的。

- **先將 feature 標準化，再實作 logistic regression：**

我自己在資料上切 validation set 來測試，發現如果不先把資料標準化的話，準確率會是 0.776027271052；如果先把資料標準化的話，準確率會是 0.857564031693。

從上述數據可以看出：不使用標準化得出來的結果，準確率明顯較差。

我想原因是因為：大多數 feature 的值都在 0、1 之間，但是有少數幾個 feature 的值比 1 還要大得多（例如：age、fnlwgt、capital_gain、capital_loss、hours_per_week），所以訓練的時候，這些值比較容易影響係數，結果就會讓收斂的方向朝向有偏差的方向進行，使得結果容易卡在 local minimum 而無法到達 global minimum。

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：

我取助教提供的 X_train 裡面所有的 attribute 的 1 維、2 維、3 維以及 sin 函式做為 feature，並把資料標準化，再加上正規化，再開始訓練資料。

跑出來的正規化程度（λ）與準確率的關係如下表所示：

正規化程度（λ）	準確率
0	0.85756
0.01	0.85756
0.1	0.85775
1	0.85793
2	0.85839
3	0.85968
5	0.84918
10	0.80127
100	0.68463

從表格中可以發現：

- 當 λ 很小（≤ 0.01）時，正規化的效用很小。
- 當 λ 大約等於 3 時，正規化的效用最大，然而對整體準確率而言，影響還是微乎其微。
- 而當 λ 太大（≥ 5）時，正規化之後的準確率反而降低了。

5. 請討論你認為哪個 attribute 對結果影響最大？

答：

我先取全部的 attribute，標準化處理後，再刪除其中一個 attribute，再使用 logistic regression 訓練出結果，並用 validation set 計算出準確率，結果如下表（已經依照準確率由小排到大）：

刪去的 attribute	準確率
capital_gain	0.8441
education	0.8514
age	0.8516
occupation	0.8540
capital_loss	0.8541
hours_per_week	0.8554
fnlwgt	0.8556
relationship	0.8558
race	0.8566
workclass	0.8574
sex	0.8580
native_country	0.8582
marital_status	0.8584

從表格中可以發現：刪除 capital_gain 的準確率最小（而且特別小），所以 captial_gain 對結果的準確率影響最大。