學號：B05702095 系級： 會計四 姓名：黃禹翔

1. (0.5%) 請比較你實作的 generative model、logistic regression 的準確率，何者較佳?

就 Kaggle 上的分數來看：

Model	Private Score	Public Score
Generative	0.84350	0.84373
Logistic Regression	0.84903	0.85442

無論在 Private Score 或是 Public Score 上，都是 Logistic Regression 略勝一籌。

2. (0.5%) 請實作特徵標準化(feature normalization)並討論其對於你的模型準確率的影響

就 Logistic Regression 來看，有無標準化的差異是非常大的。

Kaggle 分數：

Model	Private Score	Public Score
Logistic with normalization	0.84903	0.85442
Logistic	0.78798	0.79164

由上表可以發現，有無標準化對於模型的影響很大。我自己的解釋是，標準化把各個特徵的變異限制在同樣的範圍，因此可以使得模型均等的看待各個變數，而不是給予其中某些特徵特別大或特別小的參數，從而得到較一般化的模型與較佳的表現。

3. (1%) 請說明你實作的 best model，其訓練方式和準確率為何?

我的 best model 是利用 scikit-learn 的 Gradient Boosting Classifier 做出來 (使用預設參數)。演算法背後的原理是透過訓練多個比較弱的分類器，並在後一個分類器基礎上針對前一個分類器錯分的樣本加強學習，最後把這些分類器集合起來，最後準確率為：
Private Score：0.86549
Public Score ：0.86977

而在最佳模型之外，我還嘗試了不少模型，訓練過程中我覺得對我幫助比較大的一些地方是：

- 隨機打亂訓練資料集：

不知道理論依據是什麼，但我用打亂過後的資料集訓練的模型表現往往比不打亂得來的好，大概可以提升精確度 0.01 左右。

- Regularization：

在訓練 Logistic Regression 的時候，我發現訓練加大參數懲罰項的係數很有助於 Kaggle 分數的提升，而且 l1-loss 的效果比 l2-loss 的效果更好，這個差距是顯著的，約 0.05 左右，l1-loss 更適合這次的資料集。

- Feature Selection：

由於上面有用 l1-loss，因此我有嘗試利用它來做特徵選擇，但 Kaggle 上的分數不升反降。

- 重新採樣：

考量到要預測的目標有不平衡的情況(Y=0 : Y=1 = 3 : 1)，我有嘗試做一些重新採樣，像是在訓練 logistic regression 的時候，給予正樣本比較大的的權重，但這個幫助不大。或者是用 SMOTE 去模擬正樣本出來幫助訓練，在交叉驗證的時候，打亂後的 SMOTE 模擬資料集表現都比沒有做重新採樣的樣本來得準，落差大概 0.05 左右，有些模型甚至準確率到了九成，但交到 Kaggle 的時候卻跟沒有重新採樣的模型表現差不多或略差，因此我最後使用的模型還是沒有重新採樣的模型。

4. (3%) Refer to math problem

hw2 - handwrite

1. For one point,

$$P(\underline{x}, \underline{t}) = P(\underline{x} | \underline{t}) \cdot P(\underline{t})$$

$$= \prod_{k=1}^{K} \left(P(\underline{x} | C_k)\, \pi_k \right)^{t_k}$$

For the dataset, assuming independent

$$\Rightarrow L = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[P(\underline{x}_n | C_k)\, \pi_k \right]^{t_{nk}}$$

$\log(L)$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left(\log P(\underline{x}_n | C_k) + \log \pi_k \right)$$

the problem is,

$$\max_{\{\pi_k\}} \log(L)$$

$$\text{s.t. } \sum_{k=1}^{K} \pi_k = 1.$$

$\Rightarrow$ use lagrangian :

$$\mathcal{L}(\pi, \lambda) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left[\log P(x_n | C_k) + \log \pi_k \right] + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{1}{\pi_k} \sum_{n=1}^{N} t_{nk} + \lambda = 0.$$

$$\Rightarrow \pi_k = \frac{-1}{\lambda} \sum_{n=1}^{N} t_{nk} = \frac{-N_k}{\lambda} \quad ①$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{n=1}^{N} \pi_k - 1 = 0 \Rightarrow \sum_{n=1}^{N} \pi_k = 1 \quad ②$$

$$\Rightarrow \sum_{n=1}^{N} \frac{-N_k}{\lambda} = 1 = -\frac{N}{\lambda} = 1 \Rightarrow \lambda = -N$$

代入 ①，$\underline{\pi_k = \frac{-N_k}{-N} = \frac{N_k}{N}}$ #

2,

$$\frac{\partial \log(\det \Sigma)}{\partial \sigma_{ij}} \qquad \Sigma = [\,\sigma_{ij}\,]_{m \times m}.$$

$$= \frac{1}{\det(\Sigma)} \cdot \frac{\partial}{\partial \sigma_{ij}} \det(\Sigma)$$ by chain Rule

$$= \frac{1}{\det(\Sigma)} \cdot \frac{\partial}{\partial \sigma_{ij}} \sum_{l} (-1)^{k+l} \sigma_{kl} M_{kl}$$

[M_{kl} 是 Σ 把第k列 第l行 移除的結果]

$$= \frac{1}{\det(\Sigma)} \cdot (-1)^{i+j} M_{ij}$$

$$= e_j \Sigma^{-1} e_i^{T}$$

3.

$$P(\underline{x} \mid C_k) = \mathcal{N}(\underline{x} \mid \mu_k, \underline{\underline{\Sigma}}) \qquad \underline{\mu} = \begin{pmatrix} \vdots \\ \mu_k \\ \vdots \end{pmatrix}_{k \times 1}$$

$$\text{log-likelihood} = \log \prod_{n=1}^{N} \prod_{k=1}^{K} f_{\underline{x}_n}(\underline{x} \mid \mu_k, \underline{\underline{\Sigma}})$$

$(\mathcal{L})$

$$= \log \prod_{n=1}^{N} \prod_{k=1}^{K} \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\underline{x}_n - \mu)^T \Sigma^{-1} (x_n - \mu) \right)$$

$$= -\frac{Nk}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma|$$

$$- \frac{1}{2} \sum_{n=1}^{k} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu)$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = \sum_{n=1}^{N} \Sigma^{-1} (\mu - x_n) = 0$$

$$\Rightarrow \mu = \frac{1}{N} \sum_{n=1}^{N} x_n$$

$$\Rightarrow \mu_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} x_n$$

Double A

$$\frac{\partial \mathcal{L}}{\partial \Sigma^{-1}} = \frac{N}{2}\Sigma - \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)(x_n-\mu)^T$$

$*$

$$0 = N\cdot\Sigma - \sum_{n=1}^{N}(x_n-\mu)(x_n-\mu)^T$$

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_n-\mu)(x_n-\mu)^T$$

$$= \sum_{i=1}^{N}\frac{N_k}{N}\cdot\frac{1}{N_k}\sum_{n=1}^{N}t_{nk}(x_n-\mu_k)(x_n-\mu_k)^T$$

$$= \sum_{k=1}^{K}\frac{N_k}{N}S_k$$

$*$ $C = -\frac{NK}{2}\log(2\pi)$

$$C - \frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)$$

$$= C + \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{m}tr[(x_n-\mu)(x_n-\mu)^T\Sigma^{-1}$$

$\because$ $tr(ABC) = Tr(ACB)$

$|\Sigma^{-1}| = -|\Sigma|$.