Machine Learning - 2021Fall - HW2

學號：R10921098 系級：電機所碩一 姓名：楊仁傑

1. (1%) 請比較說明generative model、logistic regression兩者的異同為何？再分別列出本次使用的資料中五個分得正確/不正確的sample，並說明為什麼如此？

ANS:

1.(a)

logistic regression：

1. 找出最佳的w, b，以預測結果。
2. 不做任何參數假設，因此無法通過假定的probability distribution 得到P(x|Ci)。
3. 由於需要大量迭代計算找出最佳的w, b，因此在效率上比起generative model會略差。
4. 優點：在Data足夠多的情況下，訓練出來得模型準確率比起generative model來得高。

generative model：

1. 找出u^*, Σ^*，以預測結果。
2. 假設一個帶參數的probability contribute，結合maximun likelihood 估計法最終得到最樣的參數。
3. 由於僅需要計算一次u^*, Σ^*，因此在效率上比起logistic regression會較好。
4. 優點：對Data依賴性較低，在Data數量不足時，受到的影響較小。

小結：

1. 若輸入特徵為獨立事件，使用generative model較好；反之，使用logistic regression較好。
2. 若使用資料集數量較少，使用generative model較好；反之，使用logistic regression較好。

1.(b)

```
List first five prediction errors by generative model
Index : [[1 4 5 7 8]]
y_train: [[0 0 0 1 1]]
y_pred : [[1. 1. 1. 0. 0.]]
y_pred(un np.around()):
 [[0.55336337 0.69453725 0.92619742 0.43917172 0.37245237]]
```

Figure 1 . 列出使用generative model預測錯誤的前五筆資料

```
List first five prediction errors by logistic regression
Index : [[ 4  5  7  8 11]]
y_train: [[0 0 1 1 1]]
y_pred : [[1. 1. 0. 0. 0.]]
y_pred(un np.around()):
 [[0.58299422 0.70916128 0.48797276 0.46694069 0.47559779]]
```

Figure2. 列出使用logistic regression預測錯誤的前五筆資料

從Figure1.和Figure2.中可知，大多數的模型預測結果錯誤是因為落於0.5上下，屬於給定的特徵不足，使得模型無法正確的辨識出該對象收入是否大於50K。少部分(e.g. index=5)才是模型完全預測錯誤，導致輸出結果錯誤，個人認為這應該屬於資料中的特例，因此模型才無法利用現有的特徵去進行判斷。

index=5：
37 Private 284582 Masters 14 Married-civ-spouse Exec-managerial Wife White Female 0 0 40 United-States <=50K

我們將index=5的train data叫出來檢查，發現此人在私營企業上班，工作為行政管理階級，同時有碩士學位，並且已婚，美國籍，就現有的特徵而言很難以想像此人的收入<=50K，故承上所述，我認為這應該是屬於資料中的特例，故模型無法正常的預測。(註：此處也有可能是標籤錯誤，但不在討論範圍內)

2. (1%) 請實作兩種feature scaling的方法 (feature normalization, feature standardization)，並說明哪種方法適合用在本次作業？

ANS:

normalization:

$$X_{nor} = \frac{X - X_{min}}{X_{Max} - X_{min}} \in [0, 1]$$

standardization:

$$X_{std} = \frac{X - u}{\sigma} \sim N(0, 1)$$

feature scaling	Private Score	Public Score
normalization	0.82925	0.83144
standardization	0.83920	0.83968

Table1. logistic regression with two difference feature scaling

從Table1.中可以看到使用standardization標準化會比使用normalization正規化效果來得更好，我認為原因如下：

1. 由於normalization正規化是取決於最大值跟最小值進行計算，若最大值或最小值為離群點的話，那normalization正規化的效果就會較差。進而影響到模型的表現。

2. 使用standardization標準化時，會計算資料的標準差以及平均值，若資料中有離群點，也較難以影響到standardization標準化的結果，所以模型會有好一點的表現(此處表現提高約為1%左右)

3. (1%) 請說明你實作的best model及其背後「原理」為何？你覺得這次作業的dataset比較適合哪個model？為什麼？

ANS:

在本次的Best model中，我所使用的是gradientboostingclassifier，它是基於Gradient Boosting Decison Tree(以下稱GBDT)的分類算法。

GBDT在數據分析和預測中的效果很好的原因是因為它是一種基於決策樹的集成算法。其中Gradient Boosting 是集成方法boosting中的一種算法，通過梯度下降來對新的學習器進行迭代。在GBDT的迭代中，假設前一輪得到的搶學習器為ft-1(x)，對應的損失函數則為L(y,ft-1(x))。因此新一輪迭代的目的就是找到一個弱分類器ht(x)，使得損失函數L(y,ft-1(x)+ht(x))達到最小。

那我認為若不使用sklearn中的套件的話，logistic regression的模型會比generative model的模型來的好，原因如第一題所示，由於本次作業的train data為32562筆資料，是資料量夠大的情況，所以使用logistic regression會有更好的效果。

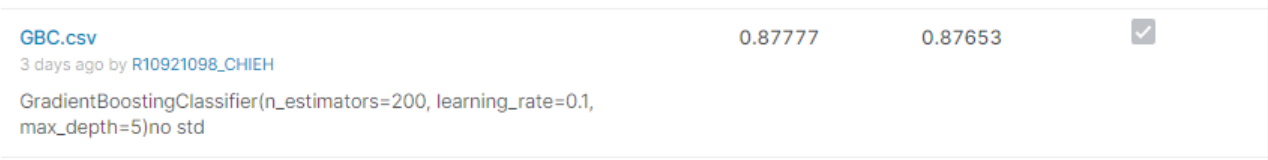

Figure3. 使用gradientboostingclassifier的預測準確率

4. (3%) Refer to math problem

(https://hackmd.io/@GfOkB4kgS66YhhM7j6TJew/BJ-wGv8HY)

ANS:

1.

Likelihood Function:

$$L(\theta) = \prod_{n=1}^{N}\prod_{k=1}^{K}(P(x_n|c_n)\pi_k)^{t_{nk}}$$

Take the ln of Likelihood Function:

$$ln(L(\theta)) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}[ln(P(x_n|c_n) + ln(\pi_k)]$$

If $ln(L(\theta))$ exists maximum, let $\sum_{k=1}^{K}\pi_k = 1$, so we can find the $L(\theta)$ maximum.

$$L(\pi, \lambda) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}[ln(P(x_n|c_n) + ln(\pi_k)] + \underbrace{\lambda(\sum_{k=1}^{K}\pi_{k-1})}_{\text{By lagrange multiplier}}$$

now, we know

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

$$\therefore \pi_k = \frac{-N_k}{\lambda} - (1)$$

$$\because \frac{\partial L}{\partial \lambda} = \sum_{k=1}^{K}\pi_{k-1} - 1 = 0$$

$$\therefore \sum_{k=1}^{K}\pi_k = 1 - (2)$$

using (1) and (2), we can get

$$\sum_{k=1}^{K}\pi_k = \sum_{k=1}^{K}\frac{-N_k}{\lambda} = \frac{-N}{\lambda} = 1$$

$$\Rightarrow \lambda = -N$$

$$\therefore \pi_k = \frac{N_k}{N}$$

2.

First, use

$$\frac{\partial log|A|}{\partial a_{ij}} = \frac{(-1)^{i+j}|A_{ij}|}{|A|}$$

then, we can know

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

Second,

$$\begin{aligned} e_j\Sigma^{-1}e_i^T &= e_j\frac{1}{det(\Sigma)}\Sigma e_i^T \\ &= \frac{1}{det(\Sigma)}(-1)^{i+j}M_j\Sigma e_i^T \\ &= \frac{1}{det(\Sigma)}(-1)^{i+j}M_{ij} - (2) \end{aligned}$$

Final, becasuse (1) = (2), so we prove

$$\frac{\partial log(det\Sigma)}{\partial \sigma_{ij}} = e_j \Sigma^{-1} e_i^T$$

3.(a)

By Gaussian distributions,

$$P(x|C_k) = N(x|u_k, \Sigma)$$

Likelihood Function:

$$L(\theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} exp(\frac{-1}{2}(x_n - u_k)^T \Sigma^{-1}(x_n - u_k))$$

Take the ln of Likelihood Function:

$$ln(L(\theta)) = \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{-N}{2} ln(2\pi) - \frac{1}{2} log|\Sigma| - \frac{1}{2}[(x_n - u_k)^T \Sigma^{-1}(x_n - u_k)]$$

$$\begin{aligned} L &= \sum_{n=1}^{N} \sum_{k=1}^{K} N(x|u_k, \Sigma) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk}(x_n - u_k)^T \Sigma^{-1}(x_n - u_k) + \lambda \end{aligned}$$

$$\begin{aligned} \because \frac{\partial L}{\partial u_k} &= \frac{\sum_{n=1}^{N} t_{nk}(x_n - u_k)^T \Sigma^{-1}(x_n - u_k)}{\partial u} \\ &= \sum_{n=1}^{N} t_{nk} \Sigma^{-1}(x_n - u_k) \text{ (Σ is postive definite)} \\ &= N_k u_k - \sum_{n=1}^{N} t_{nk} x_n = 0 \end{aligned}$$

$$\therefore u_k = \frac{\sum_{n=1}^{N} t_{nk} x_n}{N_k}$$

3.(b)

$$\begin{aligned} L(u, \Sigma|x^n) &= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{-N}{2} ln(2\pi) - \frac{1}{2} log|\Sigma| - \frac{1}{2}[(x_n - u_k)^T \Sigma^{-1}(x_n - u_k)] \\ &= \frac{-N}{2} log|\Sigma^{-1}| - \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} [t_{nk}(x_n - u_k)^T \Sigma^{-1}(x_n - u_k)] + \lambda \end{aligned}$$

$$\because \frac{\partial L}{\partial \Sigma^{-1}} = -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} [t_{nk}(x_n - u_k)(x_n - u_k)^T] + \frac{N}{2}\Sigma = 0$$

$$\begin{aligned} \therefore \Sigma &= \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} [t_{nk}(x_n - u_k)(x_n - u_k)^T]}{N} \\ &= \frac{N_k}{N} \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} [t_{nk}(x_n - u_k)(x_n - u_k)^T]}{N_k} \\ &= \frac{N_k}{N} S_k \end{aligned}$$