

# Machine Learning - Homework 2

資工四 B05902023 李澤諺

October 26, 2019

## Part 1. Programming Problem

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

下表為我實作中 performance 最好的 generative model 和 logistic regression 分別在 training 和 testing 時所得到的 accuracy：

	Train	Validation	Test	
			Public	Private
Generative Model	0.84842	0.83100	0.85442	0.85112
Logistic Regression	0.85856	0.84000	0.85995	0.85444

其中，我隨機選出 1000 筆 data 作為 validation data，剩下的則作為 training data，而我實作中 performance 最好的 generative model 為：先將 age、fmlwt、capital\_gain、capital\_loss、hours\_per\_week 這 5 個 feature 的值提升至 2 到 3 次方，並將所有 feature 進行 normalization 後，再建立 generative model，而我實作中 performance 最好的 logistic regression 為：先將 age、fmlwt、capital\_gain、capital\_loss、hours\_per\_week 這 5 個 feature 的值提升至 2 到 10 次方，並將所有 feature 進行 normalization 後，再使用助教提供的 logistic regression 函式 (除了 epoch 改為 300 以外，其餘 hyper-parameter 皆不變) 而得。

由上表可以看出我實作的 model 中，logistic regression 的 performance 較 generative model 好。

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。

下表為我實作的 generative model (方法如第 1 題所述)，有使用 normalization 和沒有使用 normalization 分別在 training 和 testing 時所得到的 accuracy：

	Train	Validation	Test	
			Public	Private
Normalization	0.84842	0.83100	0.85442	0.85112
Without Normalization	0.84730	0.83000	0.85429	0.85038

下表為我實作的 logistic regression(方法如第 1 題所述)，有使用 normalization 和沒有使用 normalization 分別在 training 和 testing 時所得到的 accuracy：

	Train	Validation	Test	
			Public	Private
Normalization	0.85856	0.84000	0.85995	0.85444
Without Normalization	0.76321	0.75200	0.76928	0.76722

下表為我實作的 best model(方法如第 3 題所述)，有使用 normalization 和沒有使用 normalization 分別在 training 和 testing 時所得到的 accuracy：

	Train	Validation	Test	
			Public	Private
Normalization	0.86746	0.86800	0.87137	0.86561
Without Normalization	0.86724	0.86400	0.87100	0.86524

由此可以大致看出，有使用 normalization 可以使得 model 的 performance 較好。

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

我將 data 進行 normalization 之後，隨機選出 1000 筆作為 validation data，剩下的則作為 training data，以此訓練 sklearn 的 GradientBoostingClassifier，其中，model 的 parameter 如下：

loss	'deviance'
learning_rate	0.1
n_estimators	100
validation_fraction	0.1
n_iter_no_change	10
tol	0.0001

最後所得到的 accuracy 如下表所示：

Train	Validation	Test	
		Public	Private
0.86746	0.86800	0.87137	0.86561

## Part 2. Math Problem

### 1.

Consider a generative classification model for  $K$  classes defined by prior class probabilities  $p(C_k) = \pi_k$  and general class-conditional densities  $p(x|C_k)$ , where  $x$  is the input feature vector. Suppose we are given a training data set  $\{x_n, t_n\}$  where  $n = 1, \dots, N$ , and  $t_n$  is a binary target vector of length  $K$  that uses

the 1-of-K coding scheme, so that it has components  $t_{nk} = 1$  if pattern  $n$  is from class  $C_k$ , otherwise  $t_{nk} = 0$ . Assuming that the data points are drawn independently from this model, show that the maximum-likelihood solution for the prior probabilities is given by

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

where  $N_k$  is the number of data points assigned to class  $C_k$ .

### solution

令  $x_n$  所屬的 class 為  $C_{x_n}$ 。  
因為 likelihood function 為

$$P(x_1, x_2, \dots, x_N) = \prod_{n=1}^N P(x_n) = \prod_{n=1}^N P(C_{x_n})P(x_n|C_{x_n})$$

將上式取 log，可得 log likelihood function 為

$$\begin{aligned} \log P(x_1, x_2, \dots, x_N) &= \sum_{n=1}^N \log P(C_{x_n}) + \sum_{n=1}^N \log P(x_n|C_{x_n}) \\ &= \sum_{k=1}^K N_k \log P(C_k) + \sum_{n=1}^N \log P(x_n|C_{x_n}) \\ &= \sum_{k=1}^K N_k \log \pi_k + \sum_{n=1}^N \log P(x_n|C_{x_n}) \end{aligned}$$

只要 log likelihood function 有最大值，即可使得 likelihood function 亦有最大值，此外，注意有限制條件  $\sum_{k=1}^K \pi_k = 1$ 。

因此，可以試著使用 Lagrange multiplier，在限制條件  $\sum_{k=1}^K \pi_k = 1$  下，求出 log likelihood function 的最大值。

令  $f = \log P(x_1, x_2, \dots, x_N)$ ，以及  $g = \sum_{k=1}^K \pi_k = 1$ 。  
因為

$$\begin{aligned} \frac{\partial}{\partial \pi_i} f &= \frac{\partial}{\partial \pi_i} \left( \sum_{k=1}^K N_k \log \pi_k + \sum_{n=1}^N \log P(x_n|C_{x_n}) \right) \\ &= \frac{\partial}{\partial \pi_i} \sum_{k=1}^K N_k \log \pi_k + 0 \\ &= \frac{\partial}{\partial \pi_i} N_i \log \pi_i = \frac{N_i}{\pi_i} \end{aligned}$$

所以

$$\nabla f = \begin{pmatrix} \frac{\partial}{\partial \pi_1} f \\ \frac{\partial}{\partial \pi_2} f \\ \vdots \\ \frac{\partial}{\partial \pi_K} f \end{pmatrix} = \begin{pmatrix} \frac{N_1}{\pi_1} \\ \frac{N_2}{\pi_2} \\ \vdots \\ \frac{N_K}{\pi_K} \end{pmatrix}$$

而

$$\frac{\partial}{\partial \pi_i} g = \frac{\partial}{\partial \pi_i} \sum_{k=1}^K \pi_k = \frac{\partial}{\partial \pi_i} \pi_i = 1$$

所以

$$\nabla g = \begin{pmatrix} \frac{\partial}{\partial \pi_1} g \\ \frac{\partial}{\partial \pi_2} g \\ \vdots \\ \frac{\partial}{\partial \pi_K} g \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

因此，若令  $\nabla f = \lambda \nabla g$ ，则有

$$\begin{pmatrix} \frac{N_1}{\pi_1} \\ \frac{N_2}{\pi_2} \\ \vdots \\ \frac{N_K}{\pi_K} \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \lambda \\ \lambda \\ \vdots \\ \lambda \end{pmatrix}$$

所以  $\pi_i = \frac{N_i}{\lambda}$ ，又

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

因此可得  $\lambda = N$ ，故

$$\pi_i = \frac{N_i}{\lambda} = \frac{N_i}{N} \quad \square$$

**2.**

Show that

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

where  $\Sigma \in \mathbb{R}^{m \times m}$  is a (non-singular) covariance matrix and  $e_j$  is a row vector (ex:  $e_3 = [0, 0, 1, 0, \dots, 0]$ ).

### solution

令  $\Sigma$  的  $(p, q)$  cofactor 為  $C_{pq}$ 。

$$\begin{aligned}
 \frac{\partial}{\partial \sigma_{ij}} \log \det \Sigma &= \frac{1}{\det \Sigma} \frac{\partial}{\partial \sigma_{ij}} \det \Sigma \\
 &= \frac{1}{\det \Sigma} \frac{\partial}{\partial \sigma_{ij}} (\sigma_{i1} C_{i1} + \sigma_{i2} C_{i2} + \cdots + \sigma_{im} C_{im}) \\
 &= \frac{1}{\det \Sigma} C_{ij} = \frac{1}{\det \Sigma} (\text{adj} \Sigma)_{ji} = \left( \frac{1}{\det \Sigma} \text{adj} \Sigma \right)_{ji} \\
 &= (\Sigma^{-1})_{ji} = e_j \Sigma^{-1} e_i^T \quad \square
 \end{aligned}$$

### 3.

Consider the classification model of problem 1 and result of problem 2 and now suppose that the class-condition densities are given by Gaussian distributions with a shared covariance matrix, so that

$$p(x|C_k) = \mathcal{N}(x|\mu_k, \Sigma)$$

Show that the maximum likelihood solution for the mean of the Gaussian distribution for class  $C_k$  is given by

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

which represents the mean of those feature vectors assigned to class  $C_k$ .

Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

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

where

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

Thus  $\Sigma$  is given by a weighted average of the covariance of the data associated with each class, in which the weighting coefficients are given by the prior probabilities of the classes.

### solution

由第 1 題可知，只要 log likelihood function 有最大值，即可使得 likelihood function 亦有最大值，而 log likelihood function 為

$$\begin{aligned}
 \log P(x_1, x_2, \dots, x_N) &= \sum_{k=1}^K N_k \log \pi_k + \sum_{n=1}^N \log P(x_n | C_{x_n}) \\
 &= \sum_{k=1}^K N_k \log \pi_k + \sum_{k=1}^K \sum_{x \in C_k} \log P(x_n | C_k) \\
 &= \sum_{k=1}^K N_k \log \pi_k + \sum_{k=1}^K \sum_{n=1}^N t_{nk} \log P(x_n | C_k) \\
 &= \sum_{k=1}^K N_k \log \pi_k + \sum_{k=1}^K \sum_{n=1}^N t_{nk} \log \mathcal{N}(x_n | \mu_k, \Sigma)
 \end{aligned}$$

其中，僅有  $\sum_{k=1}^K \sum_{n=1}^N t_{nk} \log \mathcal{N}(x_n | \mu_k, \Sigma)$  和  $\mu_1, \mu_2, \dots, \mu_K, \Sigma$  有關，因此，只要求出  $\mu_1, \mu_2, \dots, \mu_K, \Sigma$  使得  $\sum_{k=1}^K \sum_{n=1}^N t_{nk} \log \mathcal{N}(x_n | \mu_k, \Sigma)$  有最大值，該  $\mu_1, \mu_2, \dots, \mu_K, \Sigma$  即可使得 log likelihood function 有最大值。  
令

$$\begin{aligned}
 l &= \sum_{k=1}^K \sum_{n=1}^N t_{nk} \log \mathcal{N}(x_n | \mu_k, \Sigma) \\
 &= \sum_{k=1}^K \sum_{n=1}^N t_{nk} \log \left( \frac{1}{\sqrt{(2\pi)^m \det \Sigma}} e^{-\frac{1}{2}(\mu_k - x_n)^T \Sigma^{-1}(\mu_k - x_n)} \right) \\
 &= \sum_{k=1}^K \sum_{n=1}^N t_{nk} \left( -\frac{1}{2}(\mu_k - x_n)^T \Sigma^{-1}(\mu_k - x_n) - \frac{1}{2} \log \det \Sigma - \frac{m}{2} \log 2\pi \right)
 \end{aligned}$$

因爲

$$\begin{aligned}
\frac{\partial l}{\partial \mu_i} &= \frac{\partial}{\partial \mu_i} \sum_{k=1}^K \sum_{n=1}^N t_{nk} \left( -\frac{1}{2} (\mu_k - x_n)^T \Sigma^{-1} (\mu_k - x_n) - \frac{1}{2} \log \det \Sigma - \frac{m}{2} \log 2\pi \right) \\
&= \frac{\partial}{\partial \mu_i} \sum_{n=1}^N t_{ni} \left( -\frac{1}{2} (\mu_i - x_n)^T \Sigma^{-1} (\mu_i - x_n) - \frac{1}{2} \log \det \Sigma - \frac{m}{2} \log 2\pi \right) \\
&= \sum_{n=1}^N (t_{ni} \cdot \frac{\partial}{\partial \mu_i} \left( -\frac{1}{2} (\mu_i - x_n)^T \Sigma^{-1} (\mu_i - x_n) - \frac{1}{2} \log \det \Sigma - \frac{m}{2} \log 2\pi \right)) \\
&= \sum_{n=1}^N t_{ni} \left( -\frac{1}{2} \cdot 2 \Sigma^{-1} (\mu_i - x_n) \right) = \sum_{n=1}^N (\Sigma^{-1} ((-t_{ni})(\mu_i - x_n))) \\
&= \Sigma^{-1} \cdot \sum_{n=1}^N (t_{ni} x_n - t_{ni} \mu_i) = \Sigma^{-1} \cdot \left( \sum_{n=1}^N t_{ni} x_n - \sum_{n=1}^N t_{ni} \mu_i \right) \\
&= \Sigma^{-1} \cdot \left( \sum_{n=1}^N t_{ni} x_n - \left( \sum_{n=1}^N t_{ni} \right) \mu_i \right) = \Sigma^{-1} \cdot \left( \sum_{n=1}^N t_{ni} x_n - N_i \mu_i \right)
\end{aligned}$$

因此，令  $\frac{\partial l}{\partial \mu_i} = 0$ ，可得

$$\begin{aligned}
\Sigma^{-1} \cdot \left( \sum_{n=1}^N t_{ni} x_n - N_i \mu_i \right) &= 0 \\
\sum_{n=1}^N t_{ni} x_n - N_i \mu_i &= 0 \\
\mu_i &= \frac{1}{N_i} \sum_{n=1}^N t_{ni} x_n
\end{aligned}$$

接著，由於第 2 題的證明中並未使用到任何 covariance matrix 的性質，因此事實上由第 2 題的證明，可得： $\forall A \in \mathbb{R}^{m \times m}$ ，若  $A$  爲 invertible，則有  $\frac{\partial}{\partial A_{ij}} \log \det A = (A^{-1})_{ji}$ ，故  $\frac{\partial}{\partial A} \log \det A = (A^{-1})^T$ 。

因此，當  $\Sigma$  爲 covariance matrix 且爲 invertible 時，可得

$$\begin{aligned}
\frac{\partial}{\partial \Sigma^{-1}} \log \det \Sigma &= \frac{\partial}{\partial \Sigma^{-1}} \log \frac{1}{\det \Sigma^{-1}} \\
&= -\frac{\partial}{\partial \Sigma^{-1}} \log \det \Sigma^{-1} \\
&= -((\Sigma^{-1})^{-1})^T = -\Sigma^T = -\Sigma
\end{aligned}$$

此外

$$\frac{\partial}{\partial \Sigma^{-1}} (\mu_k - x_n)^T \Sigma^{-1} (\mu_k - x_n) = (\mu_k - x_n)(\mu_k - x_n)^T$$

所以

$$\begin{aligned}
\frac{\partial l}{\partial \Sigma^{-1}} &= \frac{\partial}{\partial \Sigma^{-1}} \sum_{k=1}^K \sum_{n=1}^N t_{nk} \left( -\frac{1}{2} (\mu_k - x_n)^T \Sigma^{-1} (\mu_k - x_n) - \frac{1}{2} \log \det \Sigma - \frac{m}{2} \log 2\pi \right) \\
&= \sum_{k=1}^K \sum_{n=1}^N (t_{nk} \cdot \frac{\partial}{\partial \Sigma^{-1}} \left( -\frac{1}{2} (\mu_k - x_n)^T \Sigma^{-1} (\mu_k - x_n) - \frac{1}{2} \log \det \Sigma - \frac{m}{2} \log 2\pi \right)) \\
&= \sum_{k=1}^K \sum_{n=1}^N t_{nk} \left( -\frac{1}{2} (\mu_k - x_n) (\mu_k - x_n)^T - \frac{1}{2} (-\Sigma) \right) \\
&= \frac{1}{2} \sum_{k=1}^K \sum_{n=1}^N (t_{nk} \Sigma - t_{nk} (\mu_k - x_n) (\mu_k - x_n)^T) \\
&= \frac{1}{2} \sum_{k=1}^K \left( \sum_{n=1}^N t_{nk} \Sigma - \sum_{n=1}^N t_{nk} (\mu_k - x_n) (\mu_k - x_n)^T \right) \\
&= \frac{1}{2} \sum_{k=1}^K \left( \left( \sum_{n=1}^N t_{nk} \right) \Sigma - N_k S_k \right) = \frac{1}{2} \sum_{k=1}^K (N_k \Sigma - N_k S_k) \\
&= \frac{1}{2} \left( \sum_{k=1}^K N_k \Sigma - \sum_{k=1}^K N_k S_k \right) = \frac{1}{2} \left( N \Sigma - \sum_{k=1}^K N_k S_k \right)
\end{aligned}$$

因此，令  $\frac{\partial l}{\partial \Sigma^{-1}} = 0$ ，可得

$$\begin{aligned}
\frac{1}{2} \left( N \Sigma - \sum_{k=1}^K N_k S_k \right) &= 0 \\
\Sigma &= \frac{1}{N} \sum_{k=1}^K N_k S_k = \sum_{k=1}^K \frac{N_k}{N} S_k \quad \square
\end{aligned}$$