Homework2 Report – Income Prediction

B04703117 財金五 謝昊辰

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

   Generative Model:

   Private accuracy: 0.78933，Public accuracy: 0.79275

   Logistic Model:

   Private accuracy: 0.84289，Public accuracy: 0.84557

   ➔ Logistic regression 較佳。

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。

   若沒有 feature normalization 的情況下，private score 和 public score 分別為 0.76280 和 0.76474
   僅對於連續的資料做 feature normalization，private score 和 public score 可以上升至 0.85493 及 0.85945
   推測可能的原因是 fnlwgt 這個 data 的值比其他 data 來得大，做 feature normalization 可以降低 fnlwgt 對於 model 的影響力，相對加重其他資料的重要性。

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

   這次是利用 keras 實作，總共只有 input layer 和 output layer
   activation function 分別是用 relu 和 softmax
   loss function 利用 cross entropy，optimizer 為 SGD，learning rate 設 0.1
   model 的 batch size 為 100，epoch 設 20
   public 的 accuracy 為 0.86277，private 的 accuracy 為 0.85345

4.

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

$$\rightarrow \ \ln L = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \times (\ln \pi_k + \ln \mathrm{P}(x|C_k))$$

$$\rightarrow \frac{\partial \ln L}{\partial \pi_k} = \sum_{n=1}^{N} t_{nk} \times \frac{1}{\pi_k} = 0 \rightarrow \pi_k = \frac{1}{N} \sum_{n=1}^{N} t_{nk} = \frac{N_k}{N}$$

5. Let A = Σ,

$$\frac{\partial \log(\det A)}{\partial \sigma_{i,j}} = \frac{1}{\det(A)} \frac{\partial \det(\mathrm{A})}{\partial \sigma_{i,j}} = \frac{1}{\det(A)} adj(A)_{ji} = \frac{1}{\det(A)} \det(A)\, {A^{-1}}_{ji}$$

$$= (A^{-1})_{ji} = e_j \Sigma^{-1} {e_i}^T$$

6.

$$L = \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(x_n | \mu_k, \Sigma)]^{t_{nk}}$$

$$\rightarrow \ln L = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \times \ln \pi_k \mathcal{N}(x_n | \mu_k, \Sigma)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \times [\ln \pi_k - \frac{1}{2}(x_n - \mu_k)^T \Sigma^{-1} (x_n - \mu_k) - \frac{1}{2} \ln|\Sigma| + C]$$

$$\rightarrow \frac{\partial \ln L}{\partial \mu_\mathrm{k}} = \sum_{n=1}^{N} t_{nk} \times (x_n - \mu_k) = 0 \rightarrow \sum_{n=1}^{N} t_{nk} x_n = \sum_{n=1}^{N} t_{nk} \mu_k = N_k \mu_k$$

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

All term about Σ,

$$\rightarrow -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln|\Sigma| - \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} (x_n - \mu_k)^T \Sigma^{-1} (x_n - \mu_k)$$

$$= -\frac{N}{2} \ln|\Sigma| - \frac{1}{2} \Sigma^{-1} \sum_{k=1}^{K} N_k S_k$$

Setting the derivative with respect to Σ equal to zero,

$$-\frac{N}{2} e_j \Sigma^{-1} {e_i}^T + \frac{1}{2} \Sigma^{-2} \sum_{k=1}^{K} N_k S_k = 0$$

$$\rightarrow e_j \Sigma^{-1} {e_i}^T = \Sigma^{-2} \sum_{k=1}^{K} \frac{N_k}{N} S_k \rightarrow \Sigma = \sum_{k=1}^{K} \frac{N_k}{N} S_k$$