HW2 Report

資工三 B05902001 廖彥綸

本次作業所有 train 皆使用前 17500 組 data 進行 training 剩下的 2500 組作 validation。

Problem 1. 請簡單描述你實作之 logistic regression 以及 generative model 於此 task 的表現,並試著討論可能原因。

未進行任何處理的情況下 logistic regression 在 validation 和 kaggle 的準確度分別為 0.8144 和 0.8116。而 generative model 分別得到 0.6904 和 0.6910。以此看來在本作業使用 logistic regression 的效果勝過 generative model。可能原因為：很多資料的屬性並不連續，所以 generative model 得到比較差的效果。logistic regression 也更適合處理將資料改為 one-hot 等步驟。

Problem 2. 請試著將 input feature 中的 gender, education, marriage status 等改為 one-hot encoding 進行 training process,比較其模型準確率及其可能影響原因。

更改成 one-hot encoding 前的程式 validation 為 0.8200，kaggle 的得分為 0.8182，將那三項用 one-hot encoding 取代後 validation 為 0.8204，kaggle 的得分則是 0.8188，有小幅度的提升。

原本的資料性別為 0 或 1；教育程度為 1 至 6 的分布；結婚為 1、2、3。但用這種方式表示讓資料產生距離的概念。像是結婚那一欄會被認為是狀態 1 較狀態 2 接近，而距離狀態 3 遠。但實際上可能不存在這種接近與否的關係。改為用 one-hot encoding 可以改善這類的影響。

Problem 3. 請試著討論哪些 input features 的影響較大(實驗方法沒有特別限制,但請簡單闡述實驗方法)。

刪除各行資料，如果刪除的資料對分類影響不大或是沒有影響，得到的正確率變化較小，甚至上升(減少不必要的資料)。如果刪除的資料對決定分類的判斷很重要，正確率應明顯下滑。以下資料為測試不同行被刪除時的結果，原始資料為 validation：0.8204，kaggle：0.8188。

status	validation	kaggle
Drop sex	0.8212	0.8190
Drop education	0.8196	0.8196
Drop LIMIT_BAL	0.8228	0.8180
Drop age	0.8196	0.8176
Drop PAY(all)	0.7808	0.7822
Drop BILL_AMT(all)	0.8240	0.8176
Drop PAY_AMT(all)	0.8200	0.8178
Drop 6 (第六項)	0.8200	0.8178

刪除 sex、education、LIMIT_BAL 時得到的結果與初始相距不大，甚至有近不，這三項可能對分類結果影響較小。而 PAY 是影響較大的項目，當 PAY 被刪除後結果有大幅的下降。

Problem 4. 請實作特徵標準化 (feature normalization),並討論其對於模型準確率的影響與可能原因。

使用 normalization 和 scaling：validation 為 0.8200，kaggle 的得分則是 0.8182。
不使用 normalization 但使用 scaling 時 validation 為 0.8200，kaggle 的 0.8182。
都不使用時 validation 為 0.7804，kaggle 的得分則是 0.7856。

這次的資料部分項的相差距大，從數十萬至負數都有。導至每個 feature 進行 gradient descend 時權重不同，normalization 和 scaling 皆可有效處理這種情形，但如果不做任何相關的處理，結果會受到巨大的影響。

5.

將新的資料放入函式

$$I = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

let $z = \frac{(x-\mu)}{\sigma}$

$$I = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

$$I^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy$$

$$I^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dxdy$$

let $x = rcos\theta, y = rsin\theta$

$$I^2 = \frac{1}{2\pi} \int_{0}^{2\pi} \int_{-\infty}^{\infty} e^{-\frac{r^2}{2}} rdrd\theta$$

$$= \frac{1}{2\pi} \int_{0}^{2\pi} d\theta = \frac{1}{2\pi} 2\pi = 1$$

6.

a.

$$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_k}$$

b.

$$\frac{\partial E}{\partial z_j} = \frac{\partial E}{\partial z_k} \frac{\partial z_k}{\partial y_j} = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_k} \sum_j \frac{\partial z_k}{\partial y_j} \frac{\partial y_j}{\partial z_j}$$

c.

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial z_k} \frac{\partial z_k}{\partial y_j} \frac{\partial z_j}{\partial w_{ij}} = \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_k} \sum_j \frac{\partial z_k}{\partial y_j} \frac{\partial y_j}{\partial z_j} \sum_i \frac{\partial z_j}{\partial w_{ij}}$$