Homework 2 Report - Credit Card Default Payment Prediction

電機四 B04505025 陳在賢

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。

在固定其餘條件下（相同的 training data set、data preprocessing 程序及 testing data set），根據我的實測結果，經兩方法訓練出的模型會有完全相同的準確率，兩者皆為 82% (16400/20000)。

為了深究其原因，我將兩者訓練出的 weight vector print 出比較，發現對於大部分的權重，雖然兩者不盡相同卻都一致得相當小（10 的-2 至-4 數量級）。考量到 feature 的數值為介於 0 至 1 的小數（我有實做 feature normalization），再乘上這麼小的數值，對最終分數的影響也就微乎其微了。

因此不論兩種模型，最後都僅由對應權重值較大的一、兩個 feature 進行預測，也就使得即便權重差異大，但預測結果仍大同小異。

Problem 2. 請試著將 input feature 中的 gender, education, martial 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。

同樣固定其餘條件，並以 logistic regression 為實驗模型，會發現無論有無經過 one-hot encoding 處理，兩者準確率幾乎相同。

同樣觀察 weight vector，發現此三特徵正是第一題所提到，對應相當小權重值的 feature。換句話來說，最終結果幾乎與此三特徵無關，也就使得無論如何處理數據，對預測結果也都微乎其微。

然而，這不能否定 one-hot encoding 的功用，畢竟對某些數據此三特徵對預測有關鍵性的影響，而若此影響又不為線性的（例如：預測容易感冒的歲數族群，則答案應為 <15 歲或 >70 歲，並非線性得年紀越大/越小就越容易/越不易感冒），那麼 one-hot encoding 就相當重要了！

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）。

我一開始是對所有 feature 做 data preprocessing（對 sex, education, marriage 做 one-hot encoding，並對與錢相關的數據做特徵標準化），之後直接用所有數據做 logistic regression，並將 weight vector print 出，觀察哪些 feature 對最終預測有關鍵性的判斷。後來發現除了 PAY_0、PAY_1 外，其他 feature 對應的權重根本小到不太影響預測結果。

之後改用 PAY_0、PAY_1 及各種 domain knowledge（例如：用 PAY_X 觀察過去繳款紀錄，或用 BILL_AMT/LIMIT_AMT 觀察帳單金額是否過高，導致債主存在信用不良的風險）搭配各種 if-else 判斷，建出 decision tree，發現其實手刻 decision tree 就已能達 0.822 的準確率，甚至高於 logistic regression 的結果。

最終再將上數多個布林指標加入 data 中，當作新的 feature 做 training，再同樣將 weight vector print 出觀察，一如預期得發現新增的布林指標對應的權重相當大，而其他 feature 則幾乎不影響預測。

由此可知，這筆 data 並不適合用 logistic regression 等線性模型。此外還需要對 data 有基本了解，在 domain knowledge 的介入下，才得以增加準確率。

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。

以我實測的結果來看，實做特徵標準化後，會更容易得選擇 training model 的參數（如權重向量初始值、疊代次數等），同時使準確率有些許提昇。

我認為會有此結果是因為，若未做特徵標準化，容易使最終每項 feature 對應的權重差很多，（例如：有兩 feature 對預測的影響相似，但 feature 平均值大小差異很大，則較小的 feature 對應的權重會明顯較大）。此時若將初始權重向量設為 0，會因各權重須改變的量不同，使得疊帶次數或進步幅度 (ita) 等參數值，大大得影響最終模型準確率，同時也增加 training 的時間。

然而，兩者其實會隨著疊代次數及 training 時間的增加，而得近似的準確率。這是因為特徵標準化也只是對 feature 做平移及對常數乘積，這些方法 linear model 也可實現，因此有充分的 training 後，其實準確率不至於差太多。

Problem 5~6.

5.

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2}\cdot\frac{dx}{\sqrt{2}\sigma}$$

令 $u=\frac{x-\mu}{\sqrt{2}\sigma}$, $du=\frac{dx}{\sqrt{2}\sigma}$

$$=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-u^2}du = \frac{1}{\sqrt{\pi}}\cdot\sqrt{\pi} = 1$$,得證

pf:

$$\left(\int_{-\infty}^{\infty} e^{-x^2}dx\right)^2 = \int_{-\infty}^{\infty} e^{-x^2}dx\cdot\int_{-\infty}^{\infty} e^{-y^2}dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\,dxdy$$

[* 座標轉換: $x\in(-\infty,\infty)$, $y\in(-\infty,\infty)$ $\Rightarrow$ $R\in(0,\infty)$, $\theta\in(0,2\pi)$; $dxdy \Rightarrow rdrd\theta$
整張平面 微小面積]

$$=\int_0^{2\pi}\int_0^{\infty} e^{-r^2}\,r\,dr\,d\theta = 2\pi\int_0^{\infty} re^{-r^2}dr = \pi\int_0^{\infty} e^{-r^2}dr^2 = \pi\left(-e^{-r^2}\right)\Big|_{r=0}^{\infty} = \pi$$

$$\Rightarrow \int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}$$,得證

6.

(a)
$$\frac{\partial E}{\partial z_k} = \frac{\partial E}{\partial y_k}\cdot\frac{\partial y_k}{\partial z_k} = \frac{\partial E}{\partial y_k}\cdot\frac{\partial g(z_k)}{\partial z_k} = \frac{\partial E}{\partial y_k}\cdot g'(z_k)$$

(b)
$$\frac{\partial E}{\partial z_j} = \underbrace{\frac{\partial E}{\partial z_k}}_{(a)}\cdot\frac{\partial z_k}{\partial y_j}\cdot\frac{\partial y_j}{\partial z_j} = \left(\frac{\partial E}{\partial y_k}\cdot g'(z_k)\right)\left(\frac{\partial\left(\sum_j w_{jk}\cdot y_j\right)}{\partial y_j}\right)\cdot\left(\frac{\partial g(z_j)}{\partial z_j}\right)$$

$$=\frac{\partial E}{\partial y_k}\cdot g'(z_k)\cdot w_{jk}\cdot g'(z_j)$$

(c)
$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial z_j}\cdot\frac{\partial z_j}{\partial w_{ij}} = \frac{\partial E}{\partial z_j}\left(\frac{\partial\left(\sum_i w_{ij}y_i\right)}{\partial w_{ij}}\right)$$

$$=\frac{\partial E}{\partial y_k}\cdot g'(z_k)\cdot w_{jk}\cdot g'(z_j)\cdot y_i$$