Homework 2 Report

Professor Pei-Yuan Wu EE5184 - Machine Learning

學號：R07922162 系級：資工碩一 姓名：胡嘉祐

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

|  |  |  |
| --- | --- | --- |
|  | Kaggle score | loss |
| logistic regression | 0.81820 | 0.44994 |
| generative model | 0.82100 | 0.44777 |

實作的feat皆一樣，使用原本的feature加上 limit\_ball, pay\_0, bill\_amt,pay\_amt的二次方，以及limit\_ball, pay\_0, bill\_amt\_0的三次方，並對所有data做normalization。

Logistic regression: 初始w,b=0,learning rate =0.5 with adagrad,

直接使用20000筆data train,每100筆更新一次w,b 共做1000次iteration。

Generative model: 標準的slide提供的方法，並沒有做額外的延伸。

由此可見generative model 得到較好的結果，有可能是我的logistic regression寫得不夠完善，也有可能是data本身有noise，label並不正確，而generative model能自動忽視掉這個部份。

**Problem 2. (1%) 請試著將 input feature 中的 gender, education, martial status 等改 one-hot encoding 進行 training process，比較其模型準確率及其可能影響原因**。

做one-hot encoding，其結果有好一點點而已，因為後來有進行normalization，所以使one hot的效果並不顯著

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Kaggle score | loss | Kaggle score | loss |
| One hot | yes | yes | no | no |
| logistic regression | 0.81140 | 0.45103 | 0.81800 | 0.44538 |
| generative model | 0.79720 | 0.47336 | 0.81580 | 0.467629 |

**Problem 3. (1%) 請試著討論哪些 input features 的影響較大(實驗方法不限)。**

使用generative model進行實驗，並將以下項目作二次方並對所有data進行標準化。由此可見，對於準確性影響較大，影響較小，甚至有負面反應。

|  |  |  |
| --- | --- | --- |
|  | Kaggle score | loss |
| normal | 0.81580 | 0.46762 |
| Pay\_0 | 0.79560 | 0.46581 |
| Pay\_2 | 0.79040 | 0.46712 |
| Pay\_3 | 0.80020 | 0.46603 |
| Pay\_4 | 0.80040 | 0.46772 |
| Pay\_5 | 0.80540 | 0.46749 |
| Pay\_6 | 0.80120 | 0.46668 |
| PAY\_AMT1 | 0.81380 | 0.46774 |

|  |  |  |
| --- | --- | --- |
|  | Kaggle score | loss |
| normal | 0.81580 | 0.46762 |
| Limit ball | 0.81600 | 0.46814 |
| sex | 0.81600 | 0.46814 |
| education | 0.81560 | 0.46813 |
| marriage | 0.81740 | 0.46785 |
| age | 0.81480 | 0.46792 |
| Pay\_0 | 0.79560 | 0.46581 |
| Pay\_2 | 0.79040 | 0.46712 |
| Pay\_3 | 0.80020 | 0.46603 |
| Pay\_4 | 0.80040 | 0.46772 |
| Pay\_5 | 0.80540 | 0.46749 |
| Pay\_6 | 0.80120 | 0.46668 |
| Bill\_Amt1 | 0.81760 | 0.46819 |
| Bill\_Amt2 | 0.81740 | 0.46781 |
| Bill\_Amt3 | 0.81780 | 0.46797 |
| Bill\_Amt4 | 0.81700 | 0.46786 |
| Bill\_Amt5 | 0.81660 | 0.46792 |
| Bill\_Amt6 | 0.81680 | 0.46752 |
| PAY\_AMT1 | 0.81380 | 0.46774 |
| PAY\_AMT2 | 0.81620 | 0.46772 |
| PAY\_AMT3 | 0.81620 | 0.46774 |
| PAY\_AMT4 | 0.81580 | 0.46741 |
| PAY\_AMT5 | 0.81720 | 0.46745 |
| PAY\_AMT6 | 0.81600 | 0.46786 |

**Problem 4. (1%) 請實作特徵標準化 (feature normalization)，討論其對於你的模型準確率的影響。**

使用特徵標準化對於training的結果，會有非常顯著的影響

有些feature像是limit\_ball, bill\_amt, pay\_amt數值較其他數值而言大非常多，因此如果不做標準化，則會使準確度下降非常多，我的logistic的model，甚至無法training。

僅使用23個feature

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Kaggle score | loss | Kaggle score | loss |
| Feature normalization | yes | yes | no | no |
| Logistic regression | 0.82140 | 0.44538 | 0.21940 | 21.49004 |
| generative model | 0.81580 | 0.46762 | 0.78120 | 0.68878 |

**Problem 5. (1%)**

**Problem 6. (1%)**