學號：b04902053系級： 資工二 姓名：鄭淵仁

1. **請說明你實作的generative model，其訓練方式和準確率為何？**

**答：**

* 訓練方式：

我把資料標準化，之後再依照上課投影片裡面的公式進行計算，也就是使用了Gaussian distribution的機率模型並在naïve Bayes的假設下訓練資料。

* 準確率：

我自己在資料上切validation set來測試，得到的準確率是0.841164547632。而在kaggle上面的public score是0.84128，private score是0.84633。

1. **請說明你實作的discriminative model，其訓練方式和準確率為何？**

**答：**

* 訓練方式：

我取助教提供的X\_train裡面所有的attribute的1維、2維、3維以及sin函式做為feature，並把資料標準化，然後跑5000次regression。

* 準確率：

我自己在資料上切validation set來測試，得到的準確率是0.857564031693。而在kaggle上面的public score是0.85700，private score是0.85874。

1. **請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。**

**答：**

* **先將feature標準化，再實作probability generative：**

我自己在資料上切validation set來測試，發現如果不先把資料標準化的話，準確是0.841440943431；如果先把資料標準化的話，準確率會是0.841256679565。

從上述數據可以看出：有沒有標準化對於probability generative的實作結果幾乎沒有影響。

我想原因是因為把資料標準化，不會影響到feature和結果的機率分布。如果把feature和結果的機率分布想成多維度的圖的話，「把資料標準化」這個動作，只是將這個圖延著不同的維度壓縮、延伸而已，而這個圖上面的點也會跟著壓縮、延伸，結果還會是一樣的。

* **先將feature標準化，再實作logistic regression：**

我自己在資料上切validation set來測試，發現如果不先把資料標準化的話，準確率會是0.776027271052；如果先把資料標準化的話，準確率會是0.857564031693。

從上述數據可以看出：不使用標準化得出來的結果，準確率明顯較差。

我想原因是因為：大多數feature的值都在0、1之間，但是有少數幾個 feature的值比1還要大得多（例如：age、fnlwgt、capital\_gain、capital\_loss、hours\_per\_week），所以訓練的時候，這些值比較容易影響係數，結果就會讓收斂的方向朝向有偏差的方向進行，使得結果容易卡在local minimum而無法到達global minimum。

1. **請實作logistic regression的正規化(regularization)，並討論其對於你的模型準確率的影響。**

**答：**

我取助教提供的X\_train裡面所有的attribute的1維、2維、3維以及sin函式做為feature，並把資料標準化，再加上正規化，再開始訓練資料。

跑出來的正規化程度（λ）與準確率的關係如下表所示：

|  |  |
| --- | --- |
| 正規化程度（λ） | 準確率 |
| 0 | 0.85756 |
| 0.01 | 0.85756 |
| 0.1 | 0.85775 |
| 1 | 0.85793 |
| 2 | 0.85839 |
| 3 | 0.85968 |
| 5 | 0.84918 |
| 10 | 0.80127 |
| 100 | 0.68463 |

從表格中可以發現：

* + 當λ很小（≤ 0.01）時，正規化的效用很小。
  + 當λ大約等於3時，正規化的效用最大，然而對整體準確率而言，影響還是微乎其微。
  + 而當λ太大（≥ 5）時，正規化之後的準確率反而降低了。

1. **請討論你認為哪個attribute對結果影響最大？**

**答：**

我先取全部的attribute，標準化處理後，再刪除其中一個attribute，再使用logistic regression訓練出結果，並用validation set計算出準確率，結果如下表（已經依照準確率由小排到大）：

|  |  |
| --- | --- |
| 刪去的attribute | 準確率 |
| capital\_gain | 0.8441 |
| education | 0.8514 |
| age | 0.8516 |
| occupation | 0.8540 |
| capital\_loss | 0.8541 |
| hours\_per\_week | 0.8554 |
| fnlwgt | 0.8556 |
| relationship | 0.8558 |
| race | 0.8566 |
| workclass | 0.8574 |
| sex | 0.8580 |
| native\_country | 0.8582 |
| marital\_status | 0.8584 |

從表格中可以發現：刪除capital\_gain的準確率最小（而且特別小），所以captial\_gain對結果的準確率影響最大。