學號：B02705027 系級： 資管四 姓名：陳信豪

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

答：

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

答：

* 先將資料做抽取，取1維、2維和3維，也就是會有106\*3個features
* 將所有資料做標準化
* 做 train test split，80% training， 20% testing (validation)
* 做 Adagrad Gradient Descent ，其中 loss function 為 [ yhat – sigmoid (f(x)) ]^2

def gradient(dataset, w):

g = np.zeros(len(w))

for x,y in dataset:

x = np.array(x)

error = sigmoid(w.T.dot(x))

g[0] -= 2 \* (y - error) \* 1

g[1:] -= 2 \* (y - error) \* x[1:]

return g

lr += gradient(train\_dataset, w) \*\* 2

eps = 1e-8

w -= eta \* gradient(train\_dataset, w) / (np.sqrt(lr) + eps)

* 將每次 iteration 結果的準確率和參數 (w) 存起來
* 如果 validation dataset準確率小於 2000 次以前的 iteration 結果，那麼便停止訓練
* 找到 validation dataset 準確率最高的那一次作為結果
* 結果為在第 1381 次 iteration 有最佳 validation 準確率 86.35%



3.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：

mean = np.mean(dataset, axis=0)

mean[0] = 0

train\_mean = np.tile(mean,(len(dataset),1))

std = np.std(dataset, axis=0)

std[0] = 1

train\_std = np.tile(std,(len(dataset),1))

dataset = dataset - train\_mean

dataset = dataset/train\_std

|  |  |
| --- | --- |
| 沒有做標準化  Cost function 難以收斂，準確率無法攀升 | 有做標準化  Cost function 得以收斂，準確率攀升 |
|  |  |

4. 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：

error = y - sigmoid(w.T.dot(x))

**error += lamb \* sum(w \*\* 2)**

**validation 準確率**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | lamb = 0 | lamb = 0.0005 | lamb = 0.005 | lamb = 0.05 |
| Train Accuracy | 0.85601965602 | 0.854115479115 | 0.361240786241 | 0.342997542998 |
| Valid Accuracy | 0.848965051287 | 0.847613782937 | 0.378662244334 | 0.357349057183 |

隨著 lamb 增加，準確率卻不斷下降

5.請討論你認為哪個attribute對結果影響最大？

將 logistic regression 結果的各項參數值取絕對值後，做排序，並找出最大的值

In: (np.argsort(abs(w[1:]))[::-1])[:10]

Out: array([ 78, 184, 290, 3, 0, 110, 240, 28, 134, 220])

In: (np.argsort(abs(w[1:]))[::-1] % 106)[:10]

Out: array([78, 78, 78, 3, 0, 4, 28, 28, 28, 8])

資料中的第 78 個feature，**Holand-Netherlands** ，為對結果影響最大的 attribute

其一維、二維、三維是整體的前三名