Logistic Regression

李孟潔

111522030

Result

- Loss function 採用 cross entropy loss:

$$-\sum\{y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)})\}$$

- 透過隨機梯度下降(SGD)的方式來找最佳theta:

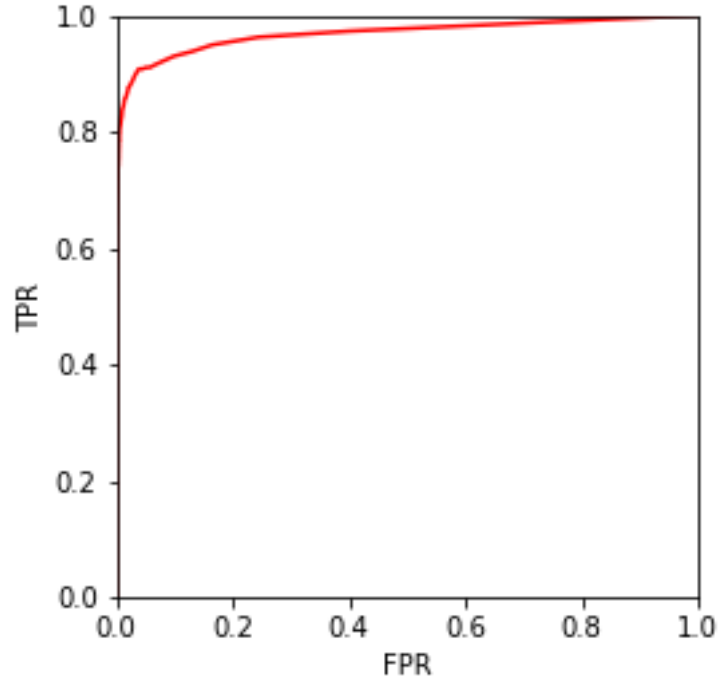
$$\boldsymbol{\theta}^{(k+1)} := \boldsymbol{\theta}^{(k)} - \alpha \boldsymbol{g}$$

$$\theta_i^{(k+1)} = \theta_i^{(k)} - \alpha \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}$$

$\frac{\partial}{\partial \theta_i} J(\theta) = \frac{1}{n}\sum_j (\hat{y}^{(j)} - y^{(j)})\, x_i^{(j)}$,
$\frac{1}{n}$ is usually ignored

Result

- 藉由不同的Threshold來找到TPR跟FPR的關係，就能畫出ROC curve。藉由觀察ROC curve來看分類的情形,當ROC curve底下的面積越大，代表預測的準確度越高!



FPR:
False Positive Rate, 當Specificity越高, FPR越低, model越能正確判斷負樣本

TPR:
True Positive Rate, 又稱Sensitivity or Recall, TPR越高,model越能正確判斷正樣本

Result

- 藉由上述方法找到的theta:

```
[[  0.        ]
 [ -9.60770821]
 [  0.44053858]
 [ 16.9074481 ]
 [  0.80578103]
 [ -5.81422069]
 [  2.75235616]
 [-11.04834491]
 [  3.6384396 ]]
```

- Accuracy, precision and recall for training data and the testing data:

```
Logreg train accuracy: 0.975081
Logreg train precision: 0.937838
Logreg train recall: 0.796785
Logreg test accuracy: 0.977763
Logreg test precision: 0.943838
Logreg test recall: 0.787760
```