ECT_HW3

2020

第一大題

第一大題

用Weka軟體對BreastCancerNUM.arff, BreastCancerNOM.arff分別建立Linear Regression model及Logistic model, 選擇“Use training set”, 設定Attribute: type為diagnosis, 在過程中對重要步驟截圖加以說明, 並回答以下問題：

第一大題(a)-題目

(a)　LinearRegression model中, 多少因素與乳癌為正相關?多少為負相關? (10%)

第一大題(a)-解答

```
diagnosis =

     -0.0709 * radius_mean +
     -0.0047 * perimeter_mean +
      0.0008 * area_mean +
      3.8735 * compactness_mean +
     -1.1601 * concavity_mean +
     -1.9936 * concave points_mean +
     -0.969  * radius_se +
      0.0363 * perimeter_se +
      0.0029 * area_se +
    -19.4375 * smoothness_se +
      3.0168 * concavity_se +
     -0.0105 * texture_worst +
     -0.0058 * perimeter_worst +
      0.0002 * area_worst +
     -0.3631 * concavity_worst +
     -1.8387 * concave points_worst +
     -0.811  * symmetry_worst +
     -3.909  * fractal_dimension_worst +
      3.133
```

正相關：6

負相關：12

第一大題(b)-題目

(b)　舉出三個LinearRegression model認為與乳癌無關的因素 (10%)

第一大題(b)-解答

自由作答，三個未出現在function中的Attribute即可

第一大題(c)-題目

(c)　比較Logistic model與Linear Regression model的均方根誤差，兩者差距多少?哪個模型表現較好? (30%)

第一大題(c)-解答

Logistic

Correctly Classified Instances	569	100	%
Incorrectly Classified Instances	0	0	%
Kappa statistic	1		
Mean absolute error	0		
Root mean squared error	0.0003		
Relative absolute error	0.0085 %		
Root relative squared error	0.0538 %		
Total Number of Instances	569		

Linear

Correlation coefficient	0.8761
Mean absolute error	0.1841
Root mean squared error	0.2331
Relative absolute error	39.3731 %
Root relative squared error	48.2196 %
Total Number of Instances	569

差距0.2331-0.0003=0.2328
Logistic表現較好

第二大題

第二大題

用python對BreastCancer.csv建立LinearRegression model及Logistic model, 在過程中對重要步驟截圖加以說明,並回答以下問題：

logistic_reg.fit(x, y.values.ravel())

第二大題(a)-題目

(a)　以coef_ function印出各Attribute的相關係數, 請問

LinearRegression model中, Attribute ‘area_mean’的係數是多少?(10%)

第二大題(a)-解答

```
#x:input
x=data.loc[:,['radius_mean','texture_mean','perimeter_mean','area_mean','smoothness_mean','compactness_mean','concavity_mean','c
```

```
In [58]: reg.coef_

Out[58]: array([[ 2.17772056e-01, -4.54546867e-03, -2.37398610e-02,
                 -3.17834750e-04, -8.46891371e-02,  4.22203525e+00,
                 -1.39799728e+00, -2.14183303e+00, -1.02709200e-01,
                 -3.32616096e-02, -4.34955932e-01,  6.75847233e-03,
                  2.25202577e-02,  9.23217886e-04, -1.58543207e+01,
                 -6.49034090e-02,  3.56546799e+00, -1.05679513e+01,
                 -1.69734069e+00,  7.14644016e+00, -1.95183121e-01,
                 -7.15937520e-03,  2.43505057e-03,  1.01122332e-03,
                 -5.42856861e-01, -6.71582941e-02, -3.81191215e-01,
                 -4.64309895e-01, -5.56787546e-01, -4.30348309e+00]])
```

第二大題(b)-題目

(b)　LinearRegression model中, 影響最大的因素為何?(20%)

第二大題(b)-解答

- 取絕對值最大：smoothness_se

```
array([[ 2.17772056e-01, -4.54546867e-03, -2.37398610e-02,
        -3.17834750e-04, -8.46891371e-02,  4.22203525e+00,
        -1.39799728e+00, -2.14183303e+00, -1.02709200e-01,
        -3.32616096e-02, -4.34955932e-01,  6.75847233e-03,
         2.25202577e-02,  9.23217886e-04, -1.58543207e+01,
        -6.49034090e-02,  3.56546799e+00, -1.05679513e+01,
        -1.69734069e+00,  7.14644016e+00, -1.95183121e-01,
        -7.15937520e-03,  2.43505057e-03,  1.01122332e-03,
        -5.42856861e-01, -6.71582941e-02, -3.81191215e-01,
        -4.64309895e-01, -5.56787546e-01, -4.30348309e+00]])
```

第二大題(c)-題目

(c)　以score() function印出兩模型的正確率，請問何者較高?(20%)

第二大題(c)-解答

LinearRegression

```
reg.score(x, y)
```

```
0.7743246526421793
```

LogisticRegression

```
logistic_regr.score(x, y)
```

```
0.9595782073813708
```

Logistic較高