CS584 Assignment 3 Report

Weilun Zhao; A20329942

# Problem 1

## a) Implement the logistic regression algorithm

dataset: spambase

numOfData: 4601

numOfFeature: 57

Base 2-fold cross

|  |  |  |
| --- | --- | --- |
|  | Cross 1 | Cross 2 |
| TP | 185 | 186 |
| FN | 45 | 44 |
| FP | 0 | 0 |
| TN | 0 | 0 |
| Accuracy | 0.804347826087 | 0.808695652174 |
| Precision | 1.0 | 1.0 |
| Recall | 0.804347826087 | 0.808695652174 |
| Fmeasure | 0.89156626506 | 0.894230769231 |

Cross 2 has better performance than Cross; using cross theta to do whole the data train

|  |  |  |
| --- | --- | --- |
|  | Total Data-Set train | Total Data-Set testing |
| TP | 2662 | 376 |
| FN | 126 | 84 |
| FP | 140 | 0 |
| TN | 1212 | 0 |
| Accuracy | 0.935748792271 | 0.817391304348 |
| Precision | 0.950035688794 | 1.0 |
| Recall | 0.954806312769 | 0.817391304348 |
| Fmeasure | 0.952415026834 | 0.8995215311 |

## b) Non-Linear combination

dataset: spambase

numOfData: 4601

numOfFeature: 57

Original data feature:

Non-linear combination:

Combine the feature data and form new feature by using two adjacent feature data; the new have to 2m-1 feature rather than m.

New Data:

numOfData: 4601

numOfFeature: 113

Using 2-fold cross validation

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cross 1 | Cross 2 | Total Data |
| TP | 182 | 183 | 2635 |
| FN | 48 | 47 | 153 |
| FP | 0 | 0 | 167 |
| TN | 0 | 0 | 1185 |
| Accuracy | 0.791304347826 | 0.795652173913 | 0.92270531401 |
| Precision | 1.0 | 1.0 | 0.94039971449 |
| Recall | 0.791304347826 | 0.795652173913 | 0.94512195122 |
| Fmeasure | 0.883495145631 | 0.886198547215 | 0.942754919499 |

## c) K-Class logistic regression

dataset: iris

numOfData: 150

numOfFeature: 4

convert y class:

1: Iris Setosa

2: Iris Versicolour

3: Iris Virginica

|  |  |  |
| --- | --- | --- |
|  | Cross 1 | Cross 2 |
| TP | 7 | 7 |
| FN | 0 | 0 |
| FP | 0 | 0 |
| TN | 0 | 0 |
| Accuracy | 1.0 | 1.0 |
| Precision | 1.0 | 1.0 |
| Recall | 1.0 | 1.0 |
| Fmeasure | 1.0 | 1.0 |
| Confusion Matrix | [[7]] | [[7]] |

|  |  |  |  |
| --- | --- | --- | --- |
| Total Data Set | | | |
|  | Class 1 | Class 2 | Class 3 |
| TP | 35 | 45 | 49 |
| FN | 0 | 5 | 1 |
| FP | 0 | 1 | 5 |
| TN | 100 | 35 | 0 |
| Confusion Matrix | [[35 0 0]  [ 0 45 5]  [ 0 1 49]] | | |
| Accuracy | 1.0 | 0.93023255814 | 0.890909090909 |
| Precision | 1.0 | 0.978260869565 | 0.907407407407 |
| Recall | 1.0 | 0.9 | 0.98 |
| Fmeasure | 1.0 | 0.9375 | 0.942307692308 |

# Problem 2

## a) Compare the difference with maximum likelihood estimation

Base on the question: all the elements using sigmoid activation

The error function:

The maximum likelihood:

So the update equation:

## b)

DataSet: iris

numOfData: 150

numOfFeature: 4

convert y class:

1: Iris Setosa

2: Iris Versicolour

3: Iris Virginica

|  |  |  |  |
| --- | --- | --- | --- |
|  | Class 1 | Class 2 | Class 3 |
| TP | 50 | 0 | 0 |
| FN | 0 | 50 | 50 |
| FP | 100 | 0 | 0 |
| TN | 0 | 100 | 50 |
| Accuracy | 0.333333333333 | 0.666666666667 | 0.5 |
| Precision | 0.333333333333 | 0 | 0 |
| Recall | 1.0 | 0 | 0 |
| Fmeasure | 0.5 | 0 | 0 |
| Confusion Matrix | [[50 0 0]  [50 0 0]  [50 0 0]] | | |

SKLearn evaluation

Test Data = 1/10 total data;

|  |  |  |  |
| --- | --- | --- | --- |
|  | Class 1 | Class 2 | Class 3 |
| TP | 0 | 0 | 0 |
| FN | 0 | 0 | 15 |
| FP | 0 | 15 | 0 |
| TN | 15 | 0 | 0 |
| Accuracy | 1.0 | 0 | 0 |
| Precision | 0 | 0 | 0 |
| Recall | 0 | 0 | 0 |
| Fmeasure | 0 | 0 | 0 |
| Confusion Matrix | [[ 0 0]  [15 0]] | | |

# Conclusion:

The performance of logistic regression with K-class is better than MLP based on the performance evaluation; the accuracy of logistic regression performance is much higher than the MPL. And the scikit-learn library of MPL classifier do not provide high accuracy prediction based on train data.