CS584 Assignment 2 Report

Weilun Zhao; A20329942

# Problem 1: 1D 2-class Gaussian discriminant analysis

## a) Using Skin Segmentation Date set

The 1D feature is the blue color degree from 0 to 256 in Skin Segmentation dataset. The y column is the describe label and classified by two different number 1 and 2.

## b) Analysis

And in this model j = 1, 2

Define the member ship function:

And

In the dataset:

Number of instance : 245057

; 127.997481951; 3875.76629985

## c) Measurement

In the first class y = 1:

The confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 27174 | 74473 |
| N | 23685 | 119725 |

|  |  |
| --- | --- |
| Accuracy | 0.599448291622 |
| Precision | 0.267336960264 |
| Recall | 0.534300713738 |
| F-measure (b=1) | 0.356366306899 |

In the second class y =2:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 119725 | 23685 |
| N | 74473 | 27174 |

|  |  |
| --- | --- |
| Accuracy | 0.599448291622 |
| Precision | 0.834844153127 |
| Recall | 0.616509953759 |
| F-measure (b=1) | 0.709254520035 |

# Problem2: nD 2-class Gaussian Discriminant Analysis

## a) Using seeds Data set

The column of y is the examined group comprised kernels belonging to two different varieties of wheat: Kama as 1, Rosa as 2

Attribute Info:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Feature# | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Name of feature | Area A | Perimeter | Compactness | Length of kernel | Width of kernel | Asymmetry  Coefficient | Length of kernel groove |

## b) Analysis

And in this model j = 1, 2

Define the member ship function:

The u1 for class 1:

[ 14.33442857 14.29428571 0.88007 5.50805714 3.24462857 2.66740286 5.08721429]

The u2 for class 2:

[ 18.33428571 16.13571429 0.88351714 6.14802857 3.67741429 3.6448 6.0206 ]

(The cov matrix is quite large in the data file)

## c) Measurement

In the class y =1:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 68 | 0 |
| N | 2 | 70 |

|  |  |
| --- | --- |
| Accuracy | 0.985714285714 |
| Precision | 1.0 |
| Recall | 0.971428571429 |
| F-measure (b=1) | 0.985507246377 |

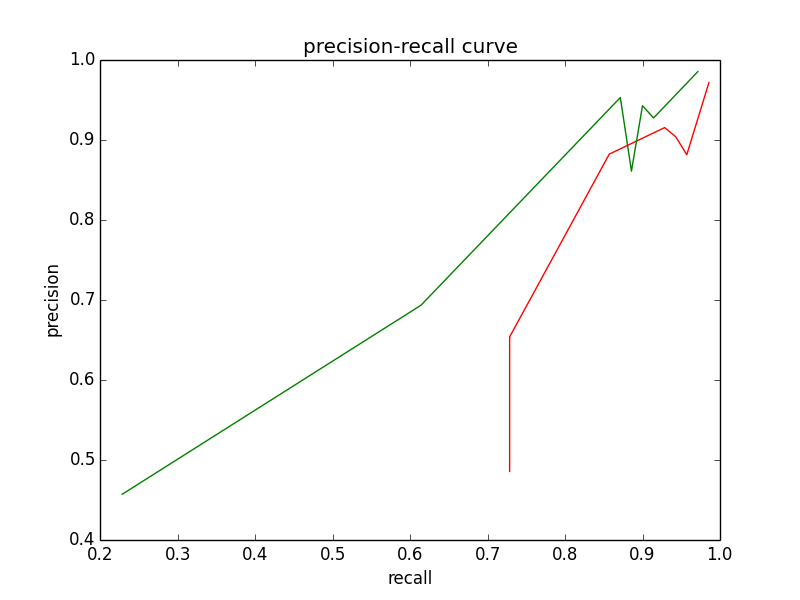
In the class y =2:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 70 | 2 |
| N | 0 | 68 |

|  |  |
| --- | --- |
| Accuracy | 0.985714285714 |
| Precision | 0.972222222222 |
| Recall | 1.0 |
| F-measure (b=1) | 0.985915492958 |

## d) Precision-recall curve

Three are 7 features in the dataset using them to get the different confusion matrix, get different precision-recall point, and plot them. (red line for class 1; green line for class2)



The area under the precision-recall curve for class 1: 0.21519380214

The area under the precision-recall curve for class 2: 0.527496830108

## 3. nD k-class Gaussian Discriminant Analysis

## a) Data Set

Using Iris Data Set (3 classes)

numFeature: 4

data size: 150

Attribute Information:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature# | 1 | 2 | 3 | 4 |
| Name of Feature | Sepal length in cm | Sepal width in cm | Petal length in cm | Petal width in cm |

Class:

Iris-Setosa as class 1; Iris-Versicolour as class 2; Iris-Virginica as class 3

## b) Analysis

The Theory is same with the second question, but the y can be 1, 2, 3

The expectation value for each feature in each class:

The u1 of class 1: [ 5.006 3.418 1.464 0.244]

The u2 of class 2: [ 5.936 2.77 4.26 1.326]

The u3 of class 3: [ 6.588 2.974 5.552 2.026]

## c) Measure

In the class y =1:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 49 | 0 |
| N | 1 | 100 |

|  |  |
| --- | --- |
| Accuracy | 0.993333333333 |
| Precision | 1.0 |
| Recall | 0.98 |
| F-measure (b=1) | 0.989898989899 |

In the class y =2:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 42 | 12 |
| N | 8 | 49 |

|  |  |
| --- | --- |
| Accuracy | 0.81981981982 |
| Precision | 0.777777777778 |
| Recall | 0.84 |
| F-measure (b=1) | 0.807692307692 |

In the class y =3:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 39 | 8 |
| N | 11 | 1 |

|  |  |
| --- | --- |
| Accuracy | 0.677966101695 |
| Precision | 0.829787234043 |
| Recall | 0.78 |
| F-measure (b=1) | 0.80412371134 |

# Problem4: Naïve Bayes with Bernoulli Feature

## a) Data Set

Spambase Data Set

numFeature: 57

data size: 4601

Update the original feature:

In the word and char frequent part, the value is the frequent of certain word or character. And I change the frequent value; if the value is not equal to 0, then update the value to 1. And 1 means certain character or word in the spam/unspam class.

Classification: spam as 1; unspam as 0

## b) Analysis

The probability mass function f for Bernoulli feature:

And in the Naïve Bayes classifier, the probabilistic model:

And

Do likelihood evaluation with Bernoulli Feature

## c) Measurement

In the class y =1:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 1478 | 190 |
| N | 335 | 2598 |

|  |  |
| --- | --- |
| Accuracy | 0.885894370789 |
| Precision | 0.886091127098 |
| Recall | 0.815223386652 |
| F-measure (b=1) | 0.84918126975 |

In the class y =0:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 2598 | 335 |
| N | 190 | 1478 |

|  |  |
| --- | --- |
| Accuracy | 0.885894370789 |
| Precision | 0.885782475281 |
| Recall | 0.931850789096 |
| F-measure (b=1) | 0.908232826429 |

# Problem5: Naïve Bayes with Binomioal Feature

## a) Likelihood analysis

The probability mass function f for Binomioal feature :

And in the Naïve Bayes classifier, the probabilistic model:

And in the multiple dimensional model

The log likelihood formula:

;

;

## b) Data Set

Using Car Evaluation Data Set

numFeature: 6

data size: 1728

The attributes description:

buying   
maint: vhigh, high, med, low.   
doors: 2, 3, 4, 5more.   
persons: 2, 4, more.   
lug\_boot: small, med, big.   
safety: low, med, high.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature number | Feature name | Original value | Changed value |
| 1 | Buying | vhigh, high, med, low | 4,3,2,1 |
| 2 | maint | vhigh, high, med, low | 4,3,2,1 |
| 3 | doors | 2, 3, 4, 5more | 2,3,4,5 |
| 4 | persons | 2, 4, more | 2,4,5 |
| 5 | lug\_boot | small, med, big | 1,2,3 |
| 6 | safety | low, med, high | 1,2,3 |

## Analysis

In the y column

Unacc as 1, acc as 2, good as 3, vgood 4

Counting Matrix for each class in the orginal data set

|  |  |  |  |
| --- | --- | --- | --- |
| unacc | acc | good | vgood |
| 1210 | 384 | 69 | 65 |

## Measure

In the class y =1:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 932 | 518 |
| N | 278 | 0 |

|  |  |
| --- | --- |
| Accuracy | 0.539351851852 |
| Precision | 0.64275862069 |
| Recall | 0.770247933884 |
| F-measure (b=1) | 0.700751879699 |

In the class y =2:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 0 | 110 |
| N | 384 | 1234 |

|  |  |
| --- | --- |
| Accuracy | 0.71412037037 |
| Precision | 0 |
| Recall | 0 |
| F-measure (b=1) | 0 |

In the class y =3:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 0 | 34 |
| N | 69 | 494 |

|  |  |
| --- | --- |
| Accuracy | 0.827470686767 |
| Precision | 0 |
| Recall | 0 |
| F-measure (b=1) | 0 |

In the class y =3:

|  |  |  |
| --- | --- | --- |
|  | P | N |
| P | 0 | 134 |
| N | 65 |  |

|  |  |
| --- | --- |
| Accuracy | 0 |
| Precision | 0 |
| Recall | 0 |
| F-measure (b=1) | 0 |

Output value for each class

|  |  |  |  |
| --- | --- | --- | --- |
| unacc | acc | good | vgood |
| 1450 | 110 | 34 | 134 |