CS1675 - Assignment 5

Zachary M. Mattis

February 21, 2019

I. Problem 1 - Logistic Regression Model

a. Normalization

`data_normalize.m`

b. Batch-Mode Gradient

`Log_regression.m`

c. Training Gradient

`main1.m`

d. Misclassification, Confusion, Sensitivity, Specificity

Dataset	**Misclassification Error**
Training	0.2988
Testing	0.2722

Table 1: Misclassification Error

Predict / Target	**1**	**1**
1	118	42
0	82	297

Table 2: Training Confusion Matrix

Predict / Target	**1**	**1**
1	46	27
0	22	134

Table 3: Test Confusion Matrix

Sensitivity	0.6765
Specificity	0.8323

Table 4: Test Sensitivity / Specificity

e. Efficiency

For a $\frac{2}{\sqrt{k}}$ schedule, the gradient converged after 30,000 epochs. For a $\frac{2}{k}$ schedule, the gradient converged after 200 epochs. Differing starting weights of ± 100 also converged in the same

amount of time, resulting in a test misclassification error of 0.2722 and training error of 0.2988.

II. Problem 2.1 - Naive Bayes Model

a. Data Analysis

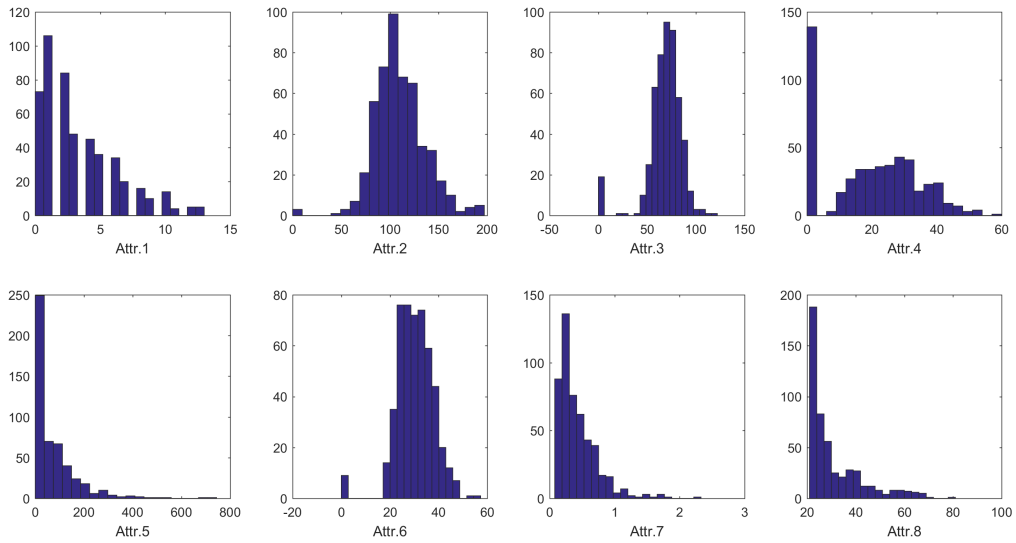

Figure 1: Class 0

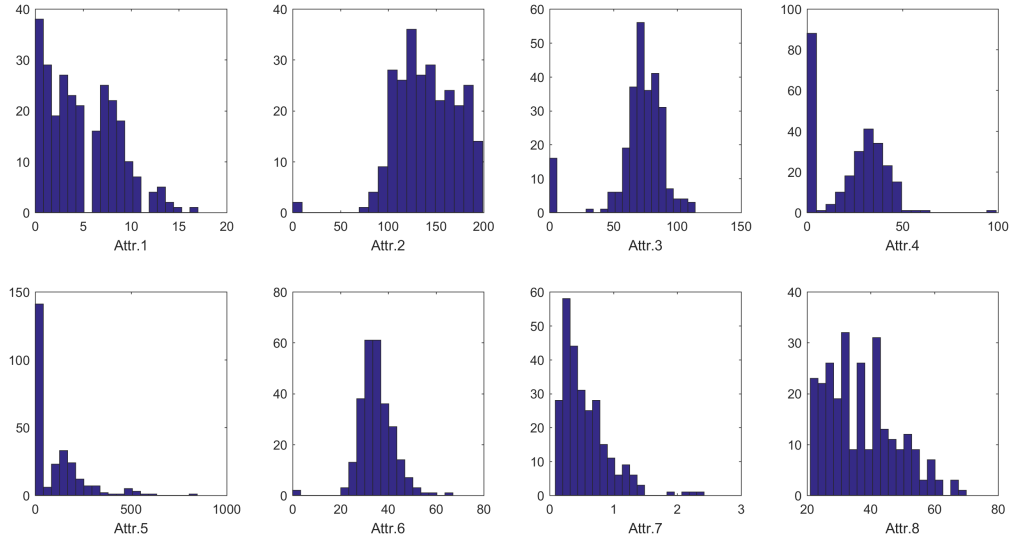

Figure 2: Class 1

b. Data Distribution

Attribute	Distribution
1	Exponential
2	Normal
3	Normal
4	Normal
5	Exponential
6	Normal
7	Exponential
8	Exponential

Table 5: Attribute Distribution

III. Problem 2.2 - Naive Bayes Learning

a. Naive Bayes Training Data

`main2_2.m`

b. Naive Bayes Parameters

Attribute	μ	σ	CI μ lower	CI μ upper	CI σ lower	CI σ upper
2	109.98	26.1412	107.6831	112.2769	24.6152	27.8705
3	68.184	18.0621	66.5969	69.7711	17.0086	19.258
4	19.664	14.8899	18.3557	20.9723	14.020	15.8749
6	30.3042	7.6899	29.6285	30.9799	7.2409	8.1986

Table 6: Class 0 Normal

Attribute	μ	σ	CI μ lower	CI μ upper	CI σ lower	CI σ upper
2	141.2575	31.9396	137.4161	145.0988	29.4451	34.8995
3	70.8246	21.4918	68.2398	73.4094	19.8133	23.4835
4	22.1642	17.6797	20.0379	24.2905	16.2989	19.3181
6	35.1425	7.2630	34.269	36.0160	6.6957	7.9360

Table 7: Class 1 Normal

Attribute	μ	CI μ lower	CI μ upper
1	3.2980	3.0270	3.6073
5	68.792	63.139	75.2436
7	0.4297	0.3944	0.4700
8	31.190	28.627	34.1151

Table 8: Class 0 Exponential

Attribute	μ	CI μ lower	CI μ upper
1	4.8657	4.3319	5.5051
5	100.3358	89.3289	113.5215
7	0.5505	0.4901	0.6228
8	37.0672	33.0009	41.9384

Table 9: Class 1 Exponential

IV. Problem 2.3 - Naive Bayes Classification

a. NB Prediction

`main2_3.m`

b. Misclassification, Confusion, Sensitivity, Specificity

Dataset	Misclassification Error
Training	0.5356
Testing	0.6241

Table 10: Misclassification Error

Predict / Target	1	1
1	168	156
0	32	183

Table 11: Training Confusion Matrix

Predict / Target	1	1
1	59	79
0	9	82

Table 12: Test Confusion Matrix

Sensitivity	0.8676
Specificity	0.5093

Table 13: Test Sensitivity / Specificity

c. Logistic Regression vs. Naive Bayes

Model	Training	Testing
Logistic Regression	0.2988	0.2722
Naive Bayes	0.5356	0.6241

Table 14: Misclassification Error

V. Problem 3 - ROC Analysis

a. *perfcurve()*

Receiver operating characteristic (ROC) curve or other performance curve for classifier output.

b. ROC

`roc_analysis.m`

c. AUC

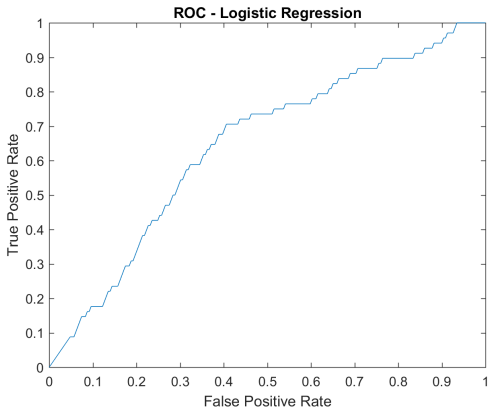

Figure 3

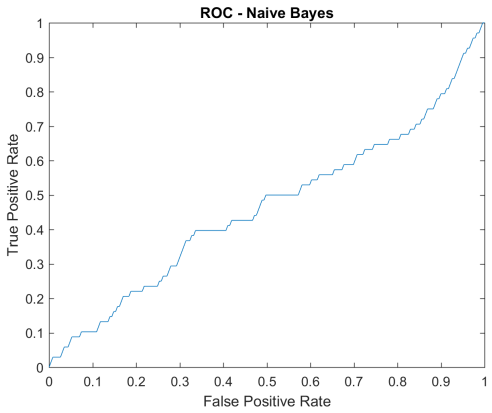

Figure 4

Model	AUC
Logistic Regression	0.8518
Naive Bayes	0.4450

Table 15: AUC

Based on the collected data for the ROC curves and the AUC statistics, logistic regression outperforms the Naive Bayes implementation, with a higher true positive rate and AUC.