Weekly Homework 3

Ava Chong
CS 1675: Intro to Machine Learning

February 21, 2019

Problem 1. Logistic regression model

(*a*) See normalized data
(*b*) N/A
(*c*) See code
(*d*) Miss-classification errors: Train = .295, Test = .4192

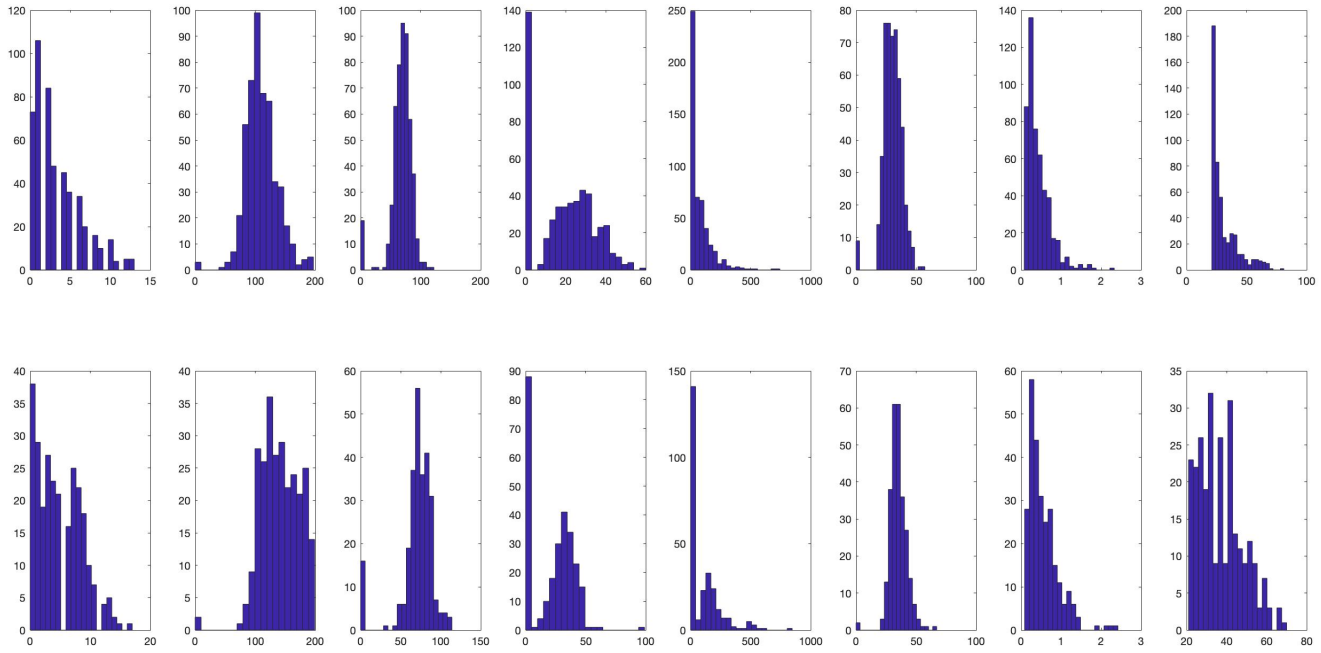
Confusion Matrix

Training		Testing	
273	93	79	52
66	107	55	43

Sensitivity: .5896
Specificity: .4526

(*e*) Experimenting with the learning rate and number of iterations yielded a test error that was consistently around 0.35. I also found that initializing the test weights produced more inconsistent results. The best miss-classification error I got for testing was .254 and for training it was .334 with 15,000 iterations.

Problem 2.1. Naive Bayes model: Exploratory data analysis
(*a*) Histograms
Top row are attributes 1-8 with Class = 0
Bottom row are attributes 1-8 with Class = 1



(*b*) Attributes 1 & 5 have distributions similar to the Gamma distribution and attributes 2, 3, 4, 6, 7 & 8 have distributions similar to a normal distribution.

Problem 2.2. Learning of the Naive Bayes classifier
(*a*) See Code
(*b*) Parameters:
Class priors: p(y=0) = .65, p(y=1) = .35
Class Conditionals: Standard Deviation is in columns 1 & 3 and the Mean is listed in columns 2 & 4

Class 0		Class 1	
3.29	0	4.87	0
109.9	26.14	141.25	31.93
68.18	18.06	70.82	21.49
19.66	14.89	22.16	17.67
68.79	0	100.33	0
30.30	7.68	35.14	7.26
0.429	0	0.55	0
31.19	0	37.06	0

Problem 2.3. Classification with the Naive Bayes model
(*a*) See Code
(*b*) Mean classification error for training = .223 and testing = .205

Confusion Matrix

Training		Testing	
294	80	142	28
45	120	19	40

Sensitivity: .89
Specificity: .59

(*c*) The Naive Bayes Algorithm had a significantly smaller miss-classification error (.205) than the Logistic Regression Algorithm (.419). This shows that for at least this data set, the Naive Bayes Algorithm worked better.

Problem 3. ROC analysis
I was not able to get this to work properly. However, the underlying principle to this problem is to to generate the value for p(y=1—x) for each point then pass the true value, the probabilities, and 1 to the ROC function and it does everything else.