

Solutions to Problem set 5

Problem 1. Logistic regression

Part d If we use $2/\sqrt{i}$ learning rate, after 2000 steps. Missclassification error of the train set is : 0.306122. Missclassification error of the test set is : 0.270742.

Training confusion matrix:

	Target0	Target1
Predict 0	263(TN)	89(FN)
Predict 1	76(FP)	111(TP)

Testing confusion matrix:

	Target0	Target1
Predict 0	121(TN)	22(FN)
Predict 1	40(FP)	46(TP)

Sensitivity = 0.6765. Specificity = 0.7516.

The learning curve is shown in Figure 1.

Problem 2. Naive Bayes model

This problem consisted of three parts. In the first, you were to run a histogram analysis on the Pima dataset and propose distributions appropriate for each of the features in that dataset as well as pick out and argue for two features which you found promising. Below are shown each histogram for features 1 . . . 8, with the distribution appropriate for it. Note that judgements as to how to interpret these histograms were, for the purposes of this assignment, somewhat but not entirely subjective. In particular, features 3 and 4 both had significant outliers which may have influenced some to argue for bimodal distributions rather than unimodal and normal, which is how we treat them here.

In addition to the histogram analysis, the first part of the assignment required you to measure the maximum and minimum values, means, and standard deviations for all features.

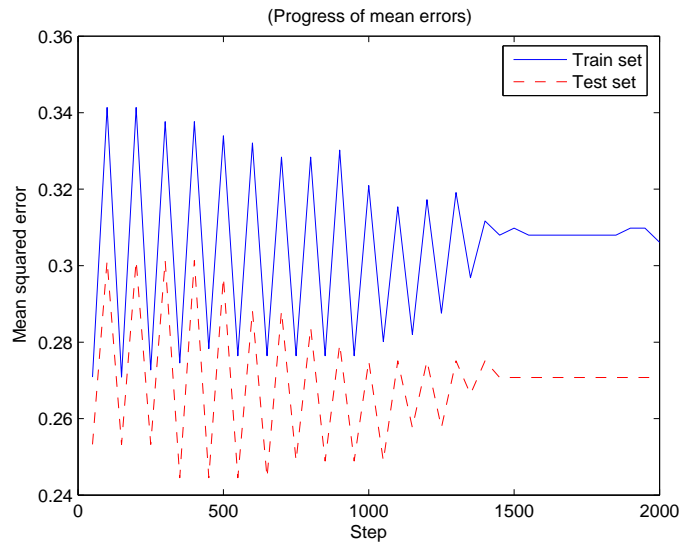


Figure 1: One initial weights, $\alpha = 2/\sqrt{i}$, 2000 steps.

Between these metrics and the histograms themselves you were to choose and argue for 2 features which looked promising to you. Reasonable answers were those which pointed to features which showed different behaviors given the class. Features which behaved the same regardless of class should be candidates for discard, or should be weighted lower.

Part a Max, Min, Std, and Mean values for Pima's eight features are given in the table below.

	F1	F2	F3	F4	F5	F6	F7	F8
max	17.0000	199.0000	122.0000	99.0000	846.0000	67.1000	2.4200	81.0000
min	0	0	0	0	0	0	0.0780	21.0000
mean	3.8451	120.8945	69.1055	20.5365	79.7995	31.9926	0.4719	33.2409
std	3.3696	31.9726	19.3558	15.9522	115.2440	7.8842	0.3313	11.7602

Problem 2.2. and 2.3.

Here you were required to implement and run a Naive Bayes generative model classifier on some pre-split Pima datasets, providing your classification errors and confusion matrix. These values, as it turned out, were:

$$NBErr_{train} = 0.2393$$

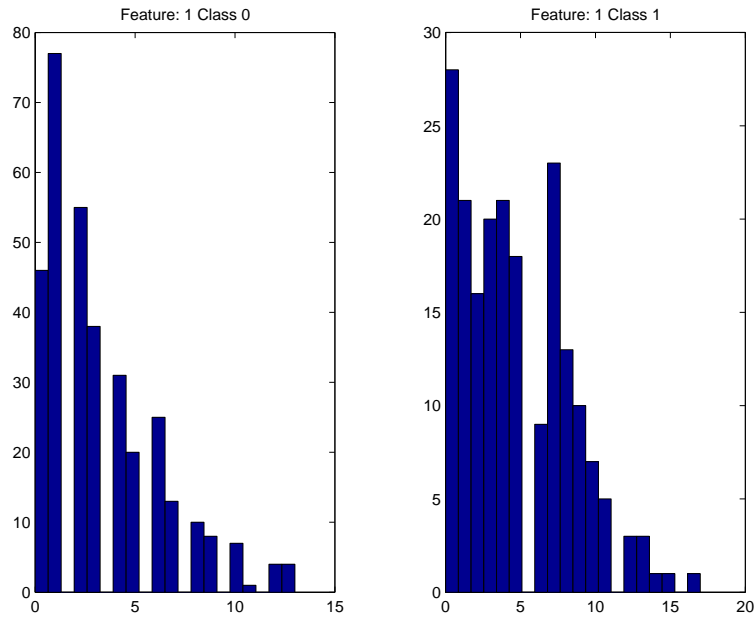


Figure 2: Feature 1, exponential distribution.

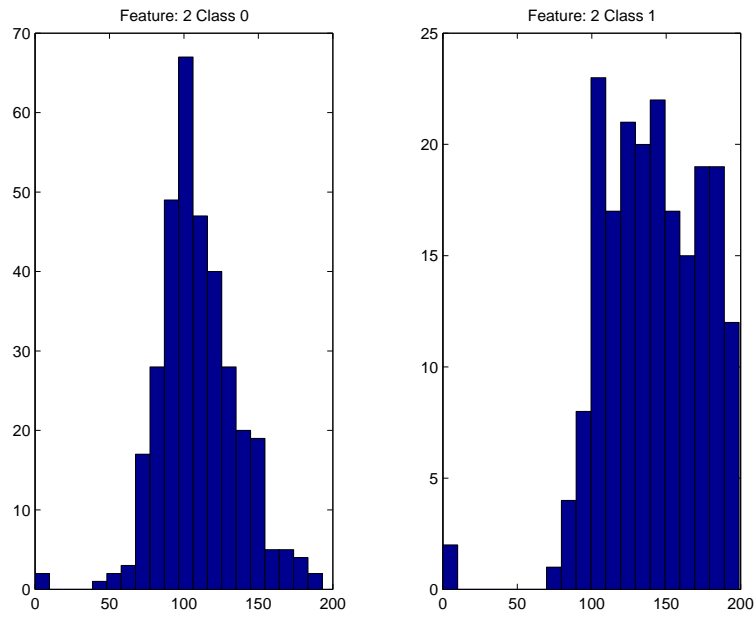


Figure 3: Feature 2, normal distribution.

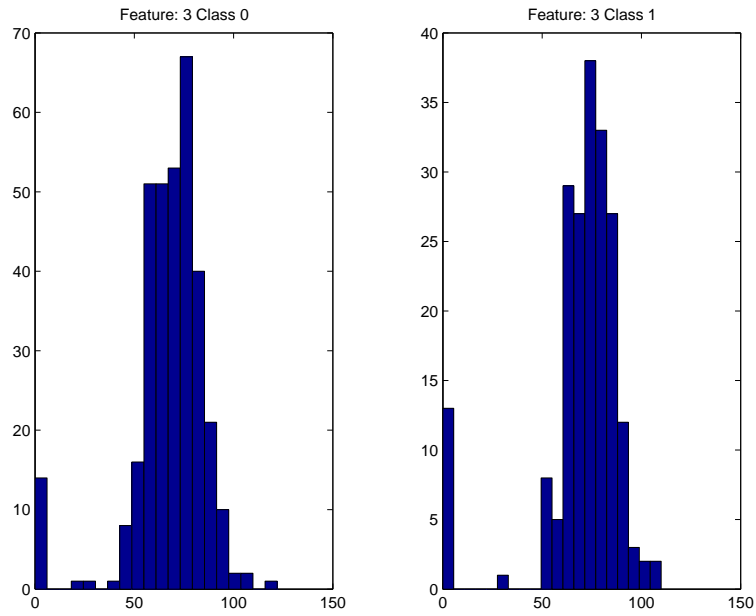


Figure 4: Feature 3, normal distribution.

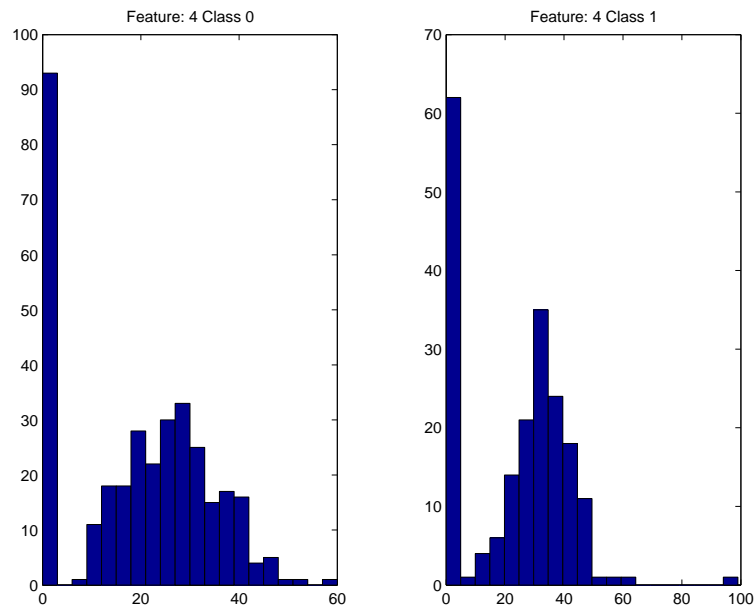


Figure 5: Feature 4, normal distribution.

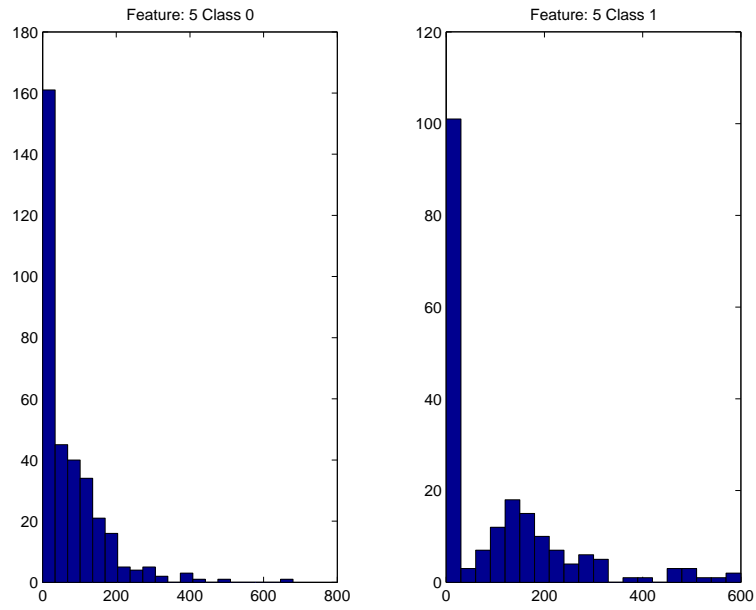


Figure 6: Feature 5, exponential distribution.

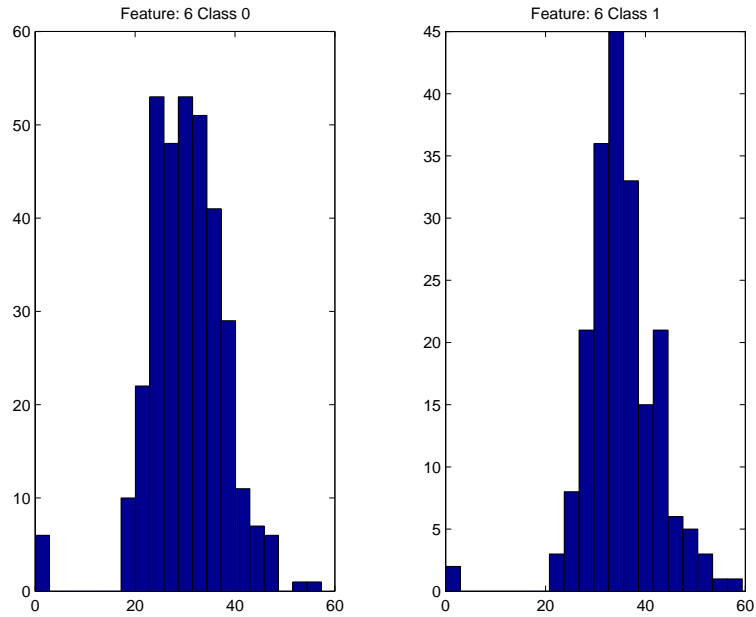


Figure 7: Feature 6, normal distribution.

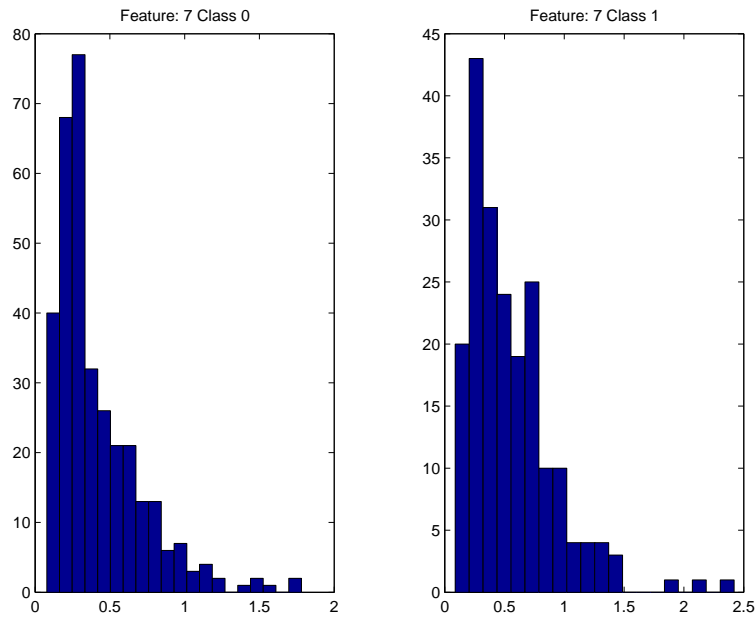


Figure 8: Feature 7, exponential distribution.

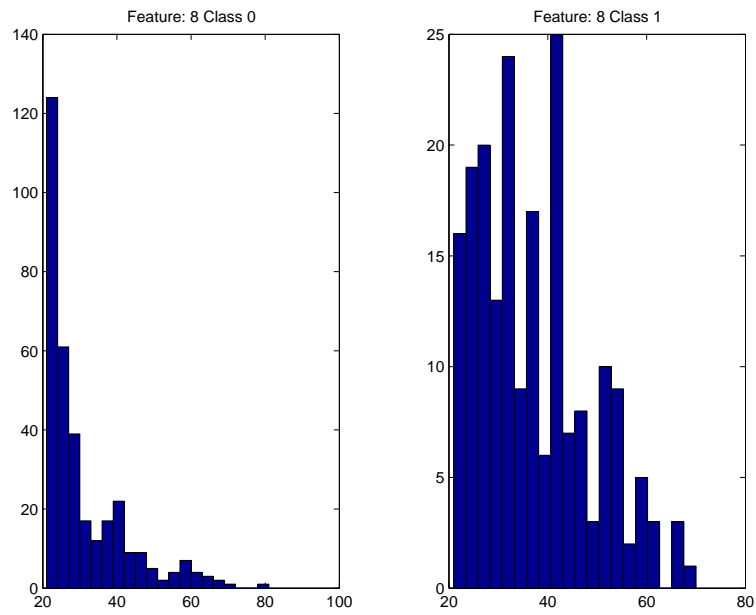


Figure 9: Feature 8, exponential distribution.

Training confusion matrix:

(289 50
79 121)

$NBErr_{test} = 0.2271$

Test confusion matrix:

(138 23
29 39)

The parameters of Naive Bayes learned were:

Class	Prior	Attr1	Attr2	Attr3	Attr4	Attr5	Attr6	Attr7	Attr8
μ_0	0.629	3.242	109.6	67.53	19.73	67.72	30.31	0.4164	31.10
μ_1	0.371	4.710	141.4	70.19	22.94	103.7	35.26	0.5491	37.12
σ_0	-	-	26.23	18.67	14.58	-	7.73	-	-
σ_1	-	-	33.67	21.62	17.83	-	7.33	-	-

Sensitivity = 0.5735. Specificity = 0.8571.

Problem 3. ROC analysis.

AUC can be computed by :

$$AUC = \sum_{x=0}^{N-1} \frac{1}{2} |X_{i+1} - X_i| (Y_{i+1} + Y_i)$$

where, $X_i = 1 - SP_i$ and $Y_i = SN_i$.

Part b

The ROC curve of the Logistic regression model is shown in Figure 10.

AUC = 0.7380.

Part c

The ROC curve of the Naive Bayes model is shown in Figure 11.

AUC = 0.8148.

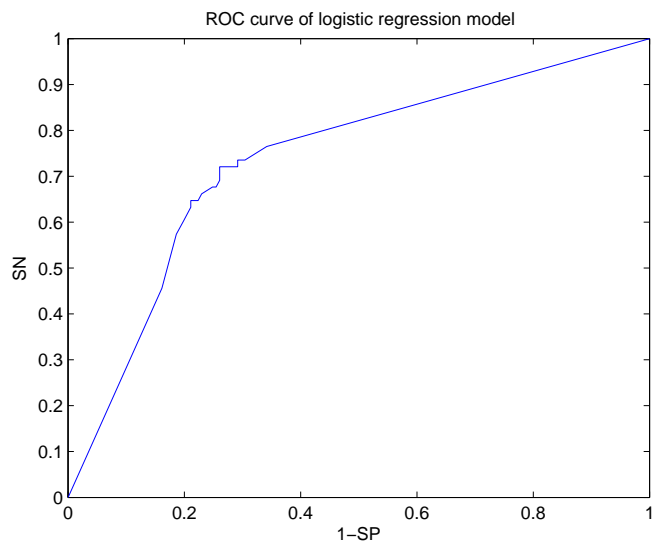


Figure 10: Logistic regression model

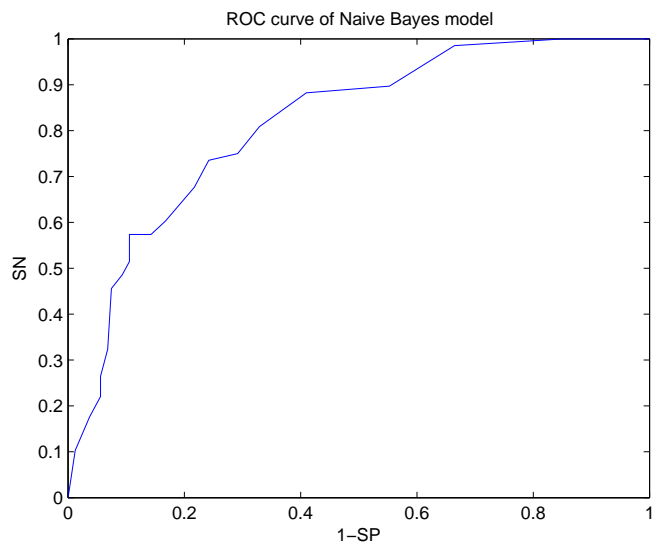


Figure 11: Naive Bayes model