EECS 440 Machine Learning Programming Problem 2 Writeup

Sixiao Zhang sxz603, Jianzhe Zhang jxz851, Junyi Zheng jxz990

(a) What is the accuracy of naïve Bayes with Laplace smoothing and logistic regression (λ =1) on the different learning problems? For each problem, perform a *t*-test to determine if either is superior with 95% confidence.

For voting, we set number of bins=100 in naïve bayes, and set iteration=100, learning rate=0.01 in logistic regression. The average accuracy of naïve bayes is 0.997, while the logistic regression 0.934. Figure 1 shows the result of the t-test.

```
The interval is [-0.295, 0.079], the null hypothesis can't be rejected
```

Fig. 1 Result of t-test on voting

For volcanoes, we set number of bins=100 in naïve bayes, and set iteration=10, learning rate=0.1 in logistic regression. The average accuracy of naïve bayes is 0.656, while the logistic regression 0.585. Figure 2 shows the result of the t-test.

```
The interval is [-0.319, 0.129], the null hypothesis can't be rejected
```

Fig. 2 Result of t-test on volcanoes

For spam, we set number of bins=100 in naïve bayes, and set iteration=10, learning rate=0.1 in logistic regression. The average accuracy of naïve bayes is 0.631, while the logistic regression 0.516. Figure 2 shows the result of the t-test.

```
The interval is [-0.129, -0.102], the null hypothesis can be rejected
```

Fig. 3 Result of t-test on spam

(b) Examine the effect of the number of bins when discretizing continuous features in naïve Bayes. Do this by comparing accuracy across several different values of this parameter using *volcanoes*.

To examine the effect of the number of bins, we use cross validation and set m=0. Table 1 shows the results. When number of bins is small, the accuracy does not change much. But when the number of bins is above 50, the accuracy starts to decrease rapidly. This might be due to the significant decrease of the number of instances in each bin, which causes the classifier to learn worse because of the limited size of training examples in each bin.

Table. 1 Accuracies on volcanoes with different nbins under the condition that m=0

Nbins	2	10	30	50	70	100	120
Accuracy	0.723	0.723	0.723	0.722	0.694	0.564	0.416

(c) Examine the effect of *m* in the naïve Bayes *m*-estimates. Do this by comparing accuracy across *m*=0, Laplace smoothing, *m*=10 and 100 on the given problems.

For voting, we set the number of bins=50. Table 2 shows the results.

Table. 2 Accuracies on voting with different m under the condition that nbins=50

m	0	-1(Laplace)	10	100
Accuracy	0.977	0.975	0.975	0.952

For volcanoes, we set the number of bins=100. Table 3 shows the results.

Table. 3 Accuracies on volcanoes with different m under the condition that nbins=100

m	0	-1(Laplace)	10	100
Accuracy	0.564	0.656	0.643	0.641

For spam, we set number of bins=100. Table 4 shows the results.

Table. 4 Accuracies on spam with different m under the condition that nbins=100

m	0	-1(Laplace)	10	100
Accuracy	0.549	0.631	0.629	0.631

From these results, we can see that if m is not 0, the accuracy can be larger (except for voting, maybe due to the limited size of dataset). So m-estimate can improve the performance of Naïve Bayes.

(d) Examine the effect of λ on logistic regression. Do this by comparing accuracy across λ=0, 1, and 10 for the given problems.

To examine the effect of λ, we run the algorithm on all of the three datasets under the condition that iteration=100 and learning rate=0.01.

Table. 5 Accuracies on voting with different λ under the condition that iter=100 and lr=0.01

λ	0	1	10
Accuracy	0.936	0.934	0.923

Table. 6 Accuracies on volcanoes with different λ under the condition that iter=100 and lr=0.01

λ	0	1	10
Accuracy	0.454	0.585	0.343

Table. 7 Accuracies on spam with different λ under the condition that iter=100 and lr=0.01

λ	0	1	10
Accuracy	0.518	0.516	0.565

From these results, we can see that, the effect of λ is different on three datasets. This is because that, when the penalty term is much smaller than the original loss, the penalty term can be ignored and won't cause much influence on the accuracy, but when the penalty term is too large, it becomes the main part of the loss and the original loss can be ignored, which will cause the accuracy to decrease. Only when the penalty term is in the same or near orders of magnitude can it improves the accuracy.

Time and Memory Requirements

When running the two algorithms, the time and memory required by the logistic regression is larger than the naïve bayes. This is mainly due to the gradient descent implemented in the logistic regression. It has to iterate many times, whereas the naïve bayes only needs to implement once and all the probabilities will be saved for further use.