EXPERIMENTS ON LOGISTIC REGRESSION FOR SPAM PREDICTION

CMPS242 HW3, winter 2008, Carl Liu, ID: 1107417

Abstract. In this homework I study and implement k-fold cross validation model selection on logistic regression algorithm for spam email prediction. A small Matlab script was used to implement the algorithm and test the performance of different models. Running results of four models were evaluated at the end.

1. Introduction

Logistic regression is a good algorithm for text classification. It's not very difficult to train a logistic regression classifier to predict if an email is a spam or not. However to achieve the best performance, i.e. minimizing the testing error, is not a easy problem. In this homework, I use k-fold cross validation on selecting a good logistic regression model for spam prediction. By setting 4 different learning rate η, 4 models are compared with 5-fold cross validation. From the result, I found the model with $\eta = 0.7$ had the best performance. At the end, I also did some experiments on shrinking and stretching label. All results are listed in the appendix plots.

2. Algorithms

2.1. **Logistic Regression.** The Logistic Regression functions are:
Sigmoid function: $\hat{y} = \frac{1}{1+e^{-wx}}$
Loss function:

$$loss(y, \hat{y}) = \begin{cases} \ln(1 + e^{wx}) & \text{if y=0} \\ \ln(1 + e^{wx}) - wx & \text{if y=1} \end{cases}$$

Total loss:

$$\frac{\sum_{t=1}^{T} loss(y_t, \sigma(wx_t))}{T}$$

Gradient descent function: $w_t = w_t - \eta(\sigma(w.x_t) - y_t)x_t$

Date: February 14, 2008.
Key words and phrases. machine learning, logistic regression, cross validation.

2.2. **Regularization.** The loss and gradient are:
Total loss:

$$\frac{\sum_{t=1}^{T}(\lambda||w||^2 + loss(y_t, \sigma(wx_t)))}{T}$$

Gradient descent function: $w_t = w_t - (\lambda w + \eta(\sigma(w.x_t) - y_t)x_t)$

2.3. **5-fold cross validation.** The steps are:
1.Permute data, split data into 3/4 training and 1/4 testing set.
2.Partition training set into 5 parts, train all models on the 4/5 part and choose the best model on 1/5 part.
3.Evaluate the best model on the 1/4 test set.

2.4. **label shrinking/stretching.** The method is:
Transform label from [0,1] to [a,b] by setting a=0.01 and b=0.99:

$$\hat{y} = \begin{cases} 1 & \text{if } \sigma(wx) > \text{b} \\ 0 & \text{if } \sigma(wx) > \text{a} \\ \frac{\sigma(wx)-a}{b-a} & \text{otherwise} \end{cases}$$

3. Matlab codes

The batch gradient descent code is listed below:

```
for p = 1:passes
                %logistic regression, sigmoid
                hx = sigmoid(xlabel*weights);
                %sum_loss
                sum_loss=hx-ylabel;
                %gradient descent
                gradient=xlabel'*(sum_loss);
                weights=weights-eta0*gradient;

               %1-norm
                mean_sum_loss=abs(mean(gradient));
                norm1 = max(sum(abs(mean_sum_loss)));

                if mean_sum_loss < 1 && weights0(1,1)==0
                    weights0=weights;
                end

end
```

Following is the regularization part:

```
%sum_loss
sum_loss=hx-ylabel;
%regularized gradient descent
gradient=xlabel'*(sum_loss);
weights=weights-eta0*gradient;
rweight=lambda*weights;
```

Following is the code for label shrinking:

```
hhx = sigmoid(xlabel*weights);

%clipping
if hhx>clip_b
    hx=1;
elseif hhx<clip_b
    hx=0;
else
    hx=(hhx-clip_a)/(clip_b-clip_a);
end
```

4. Results

The figures below show the experiment results. Figure 1 is the original batch gradient descent version($\eta = 0.2$). We can see testing error is going up along with the stopping point. It shows over-fitting.Figure 2 is the regularization version, it shows with the control of a regularization factor $\lambda = 0.02$, testing error is going down, some over-fitting were removed. Figure 3 is the 5-fold cross validation for 4 models of the original logistic regression, we can see with $\eta = 0.7$, the performance is the best. Figure 4 is the label shrinking version, it doesn't show much over-fitting. From the experiment results, I think performance of the regularization version is the best. However by tuning the parameters of label shrinking version, the performance can be acceptable.

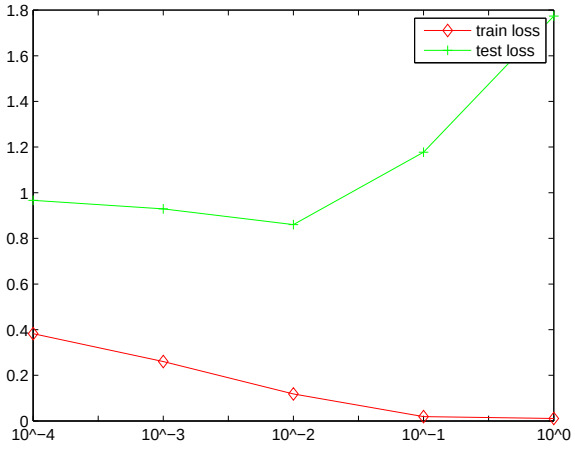

FIGURE 1. Gradient descent logistic regression (stopping points vs. loss

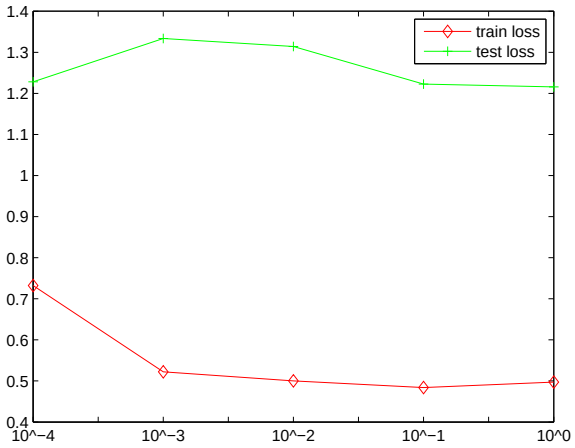

FIGURE 2. Gradient descent logistic regression - regularized

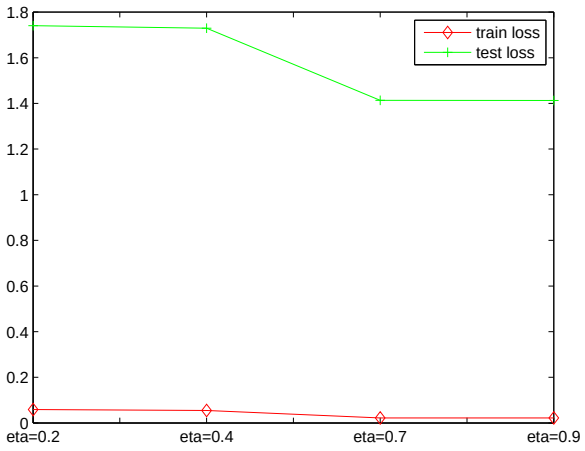

FIGURE 3. Gradient descent logistic regression - 5-fold cross validation

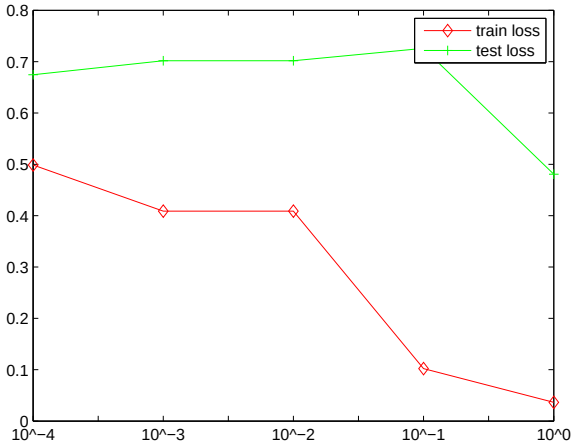

FIGURE 4. Gradient descent logistic regression - label shrinking