Homework Set 5, CPSC 8420, Spring 2022

Sherrer, Will

Due 04/28/2022, Thursday, 11:59PM EST

Problem 1

Recall the classification models we discussed in class: **LDA**, **SVM** and **Logistic Regression**, seems all of them work on binary classification task. However, in real-world applications, multi-classification is everywhere, thus in this problem we explore how to extend vanilla **Logistic Regression** for multi-classification. Assume we have K different classes and the input $\mathbf{x} \in \mathcal{R}^d$, and the probability to each class is defined as:

$$P(Y=k|X=\mathbf{x}) = \frac{exp(\mathbf{w}_k^T\mathbf{x})}{1+\sum_{l=1}^{K-1} exp(\mathbf{w}_l^T\mathbf{x})} \quad for \quad k=1,2,\dots,K{-}1; P(Y=K|X=\mathbf{x}) = \frac{1}{1+\sum_{l=1}^{K-1} exp(\mathbf{w}_l^T\mathbf{x})} \tag{1}$$

If we define $\mathbf{w}_K = \mathbf{0}$, then we can combine the two cases above as one:

$$P(Y=k|X=\mathbf{x}) = \frac{exp(\mathbf{w}_k^T\mathbf{x})}{1+\sum_{l=1}^{K-1} exp(\mathbf{w}_l^T\mathbf{x})} \quad for \quad k=1,\dots,K \tag{2}$$

1. What and how many parameters are there to be optimized?
   The parameters are $\{w_1, w_2...., w_{K-1}\}$ where they are a d-dimensional vector. There are a total of $(K-1)*d$ parameters

2. The training data is given as: $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, please simplify the log likelihood function to your best:

   $$L(\mathbf{w}_1, \dots, \mathbf{w}_{K-1}) = \sum_{i=1}^{n} lnP(Y=y_i|X=\mathbf{x}_i) \tag{3}$$

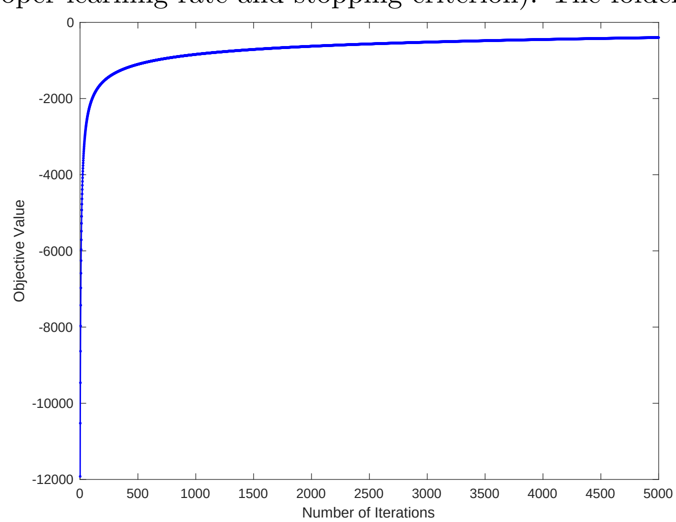
   $$L(\mathbf{w}_1, \dots, \mathbf{w}_{K-1}) = \sum_{i=1}^{n} lnP(Y=y_i|X=\mathbf{x}_i)$$
   $$= \sum_{i=1}^{n} ln\frac{exp(\mathbf{w}_k^T\mathbf{x})}{1+\sum_{l=1}^{K-1} exp(\mathbf{w}_l^T\mathbf{x})}$$
   $$= \sum_{i=1}^{n}[w_{yi}^T x - ln(1+\sum_{l=1}^{K-1} exp(w_l^T x))]$$

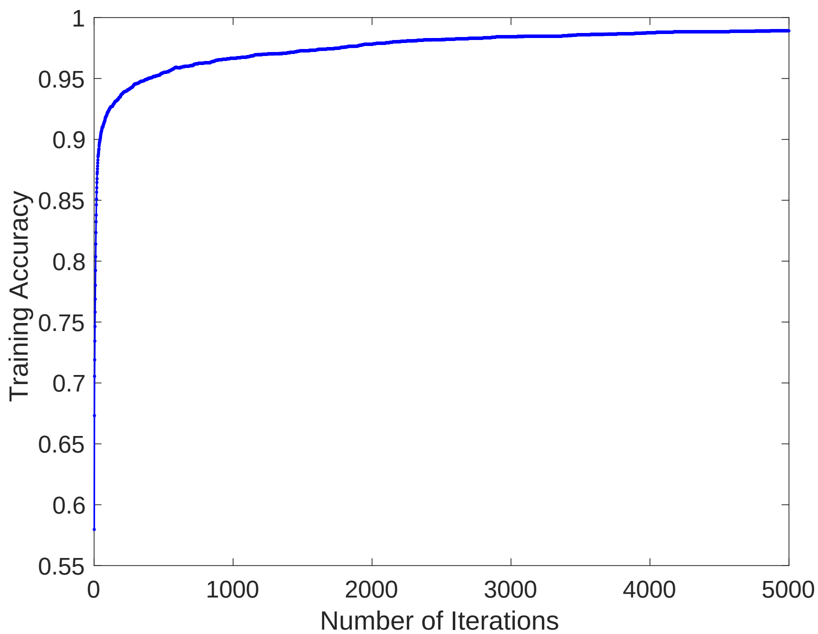
3. Now please find the gradient of L w.r.t. $\mathbf{w}_k$.
$$\nabla L(w_k) = \sum_{i=1}^{n}[I(y_i = k)x_i - \frac{exp(\mathbf{w}_k^T\mathbf{x})}{1 + \sum_{l=1}^{K-1} exp(\mathbf{w}_l^T\mathbf{x})} x_i]$$
$$= \sum_{i=1}^{m}[I(y_i = k) - P(y = k|X = x_i)]x_i$$

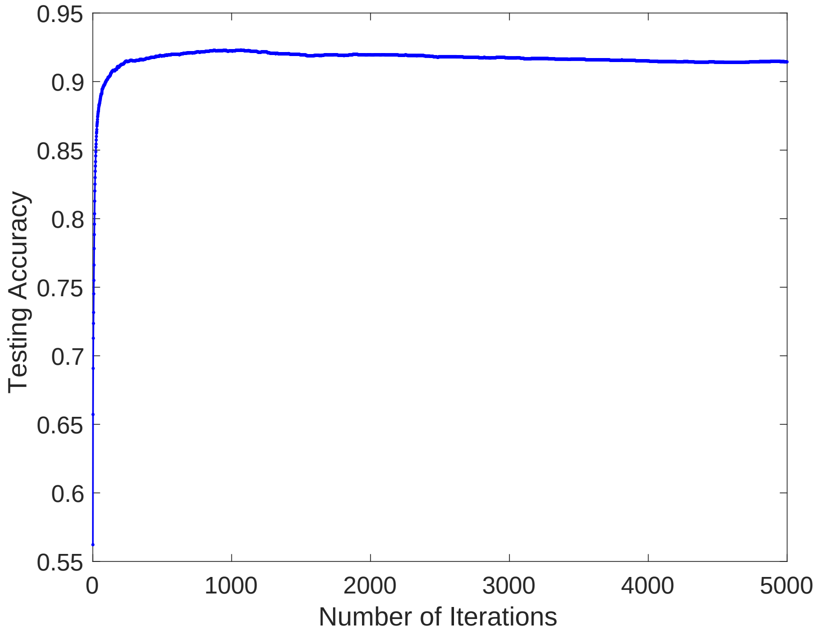
4. If we add regularization term and formulate new objective function as:
$$f(\mathbf{w}_1, \ldots, \mathbf{w}_{K-1}) = L(\mathbf{w}_1, \ldots, \mathbf{w}_{K-1}) - \frac{\lambda}{2}\sum_{l=1}^{K-1} \|\mathbf{w}_l\|_2^2, \tag{4}$$
now please determine the new gradient.
$\nabla f(w_k) = \nabla(w_k) - \lambda(w_k)$

5. You are given *USPS* handwritten recognition digit dataset, with image size 16×16. For each digit (*i.e.* 0,1,...,9) there are 600 training samples in addition to 500 testing ones. You may use: imshow(reshape($\mathbf{x}$,16,16)) to view the image in Matlab. (Non-Matlab user may utilize .txt files to conduct experiments.)

   (a) Please use gradient ascent algorithm (you are expected to complete log_grad.m) to train the model and plot 1) vanilla objective function L in Eq.(3); 2) training accuracy and 3) testing accuracy with updates respectively. Also indicate the final testing accuracy score. (Please choose a proper learning rate and stopping criterion). The folder include figures for your reference.







(b) Now if we add the regularization term as Eq.(4), please show the final accuracy when $\lambda = \{0, 1, 10, 100, 200\}$ respectively.

λ	0	1	10	100	200
Acc.	0.9144	0.9158	0.9192	0.8974	0.8818

(c) What conclusion can we draw from the above experiments?
To conclude the experiments show that it is possible for the regularization to avoid overfitting the data, namely when $\lambda = 10$, however the regularization can't be too large as to not disturb completely.

log Grad Code

```
function GG=log_grad(y, X, B)

%compute gradient

K = size(B, 2);

GG=zeros(size(B));
for i=1:size(X, 1)
    x=0;
    for j=1:K
        w_t = exp(dot(B(:, j), X(i, :)));
        x = x + w_t;
    end
    x=x+1;
    for k=1:K
        GG(:,k)=GG(:, k) + ((y(i)==k) - exp(dot(B(:,k),X(i,:)))/x) * X(
            i,:)';
    end
end
```