

# Homework Set 5, CPSC 8420, Spring 2022

Last Name, First Name

**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:

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

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

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

1. What and how many parameters are there to be optimized?
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:

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

3. Now please find the gradient of  $L$  w.r.t.  $\mathbf{w}_k$ .
4. If we add regularization term and formulate new objective function as:

$$f(\mathbf{w}_1, \dots, \mathbf{w}_{K-1}) = L(\mathbf{w}_1, \dots, \mathbf{w}_{K-1}) - \frac{\lambda}{2} \sum_{l=1}^{K-1} \|\mathbf{w}_l\|_2^2, \quad (4)$$

now please determine the new gradient.

5. You are given *USPS* handwritten recognition digit dataset, with image size  $16 \times 16$ . For each digit (*i.e.* 0,1,...,9) there are 600 training samples in addition to 500 testing ones. You may use: `imshow(reshape(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.
- (c) What conclusion can we draw from the above experiments?