CMU 10-701, Coding Assignment 3

Joseph Datz

October 2020

Question 2.1 We need to train $d(K-1)$ parameters. The d comes from the dimensions of the data, the $K-1$ comes from the amount of classes, sans the class with the fixed constraint. In this particular example, d is the $16^2 = 256$ pixels, and there are 10 classes, so the total parameter count is $256 * (10-1) = 2304$ parameters.

Question 2.2.

$$\sum_{i=1}^{n} \ln P(Y = y_i | X = \mathbf{x}_i) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta(Y = y_k) \ln \frac{e^{w_k^T \mathbf{x}_i}}{1 + \sum_{l=1}^{K-1} e^{w_l^T \mathbf{x}_i}}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \delta(Y = y_k)[w_k^T \mathbf{x}_i - \ln(1 + \sum_{l=1}^{K-1} e^{w_l^T \mathbf{x}_i})]$$

Question 2.3. For a particular parameter w_{kj},

$$\frac{\partial L}{\partial w_{kj}} = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta(Y = y_k)[x_{ij} - \frac{x_{ij} e^{w_k^T \mathbf{x}_i}}{1 + \sum_{l=1}^{K-1} e^{w_l^T \mathbf{x}_i}}]$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \delta(Y = y_k) x_{ij} [1 - \frac{e^{w_k^T \mathbf{x}_i}}{1 + \sum_{l=1}^{K-1} e^{w_l^T \mathbf{x}_i}}] = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta(Y = y_k) x_{ij} [1 - P(Y = y_i | X = \mathbf{x}_i)]$$

This of course can be edited to show the gradient for class k instead of a particular parameter.

Question 2.4. As with the previous question, we will derive with respect to a particular parameter w_{kj}:

$$\frac{\partial f}{\partial w_{kj}} = \frac{\partial L}{\partial w_{kj}} + \lambda w_{kj}$$

Where $\frac{\partial L}{\partial w_{kj}}$ is the solution to Question 2.3.