Machine Learning - 89511

Assignment 2 - Solution

Boaz Ardel - 203642806

April 2018

Department of Computer Science

Bar-Ilan University

´LATEX

Contents

Chapter 1

Theoretical Part

1.1 Q1 - Multi-class and Logistic Regression

A Logistic regression for multi-class scenario as defined and developed in class:
$P(y=i \mid x_t) = softmax(Wx_t + b)_i = \frac{e^{w_i x_t + b_i}}{\sum_{j=1}^{k} e^{w_j x_t + b_j}}$

B We would need to solve the following optimization problem over the training set:
$\arg\min_{W,b}\{L(W;b)\}$
Thus minimizing the negative log likelihood defined as:
$L(W;b) = -\sum_t \log(P(y=y_t \mid x_t))$

C The update rule of this optimization problem using stochastic gradient descent (SGD) is consisting of the following:
$W^t \leftarrow W^{t-1} - \eta \frac{\partial L(W;b)}{\partial W^{t-1}}$
$b^t \leftarrow b^{t-1} - \eta \frac{\partial L(W;b)}{\partial b^{t-1}}$
first finding the gradients - $\frac{\partial L(W;b)}{\partial W_i}, \frac{\partial L(W;b)}{\partial b_i}$:
$L(W;b) = -\sum_t \ln(P(y=y_t \mid x_t)) = \ldots \qquad \mid (P(y=i \mid x_t) = softmax(Wx_t + b)_i)$
$\ldots = -\sum_t \ln\left(\frac{e^{w_{y_t} x_t + b_{y_t}}}{\sum_{j=1}^{k} e^{w_j x_t + b_j}}\right) = \sum_t \left(\ln\left(\sum_{j=1}^{k} e^{w_j x_t + b_j}\right) - \ln\left(e^{w_{y_t} x_t + b_{y_t}}\right)\right) =$
$= \sum_t \left(\ln\left(\sum_{j=1}^{k} e^{w_j x_t + b_j}\right) - w_{y_t} x_t + b_{y_t}\right)$

$$\frac{\partial L(W;b)}{\partial W_i} = \begin{cases} \sum_t \left(\frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} (e^{w_{y_t} x_t + b_{y_t}}) x_t - x_t\right), & i = y_t \\ \sum_t \left(\frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} (e^{w_i x_t + b_i}) x_t\right), & i \neq y_t \end{cases}$$

$$\frac{\partial L(W;b)}{\partial b_i} = \begin{cases} \sum_t \left(\frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} (e^{w_{y_t} x_t + b_{y_t}}) - 1\right), & i = y_t \\ \sum_t \left(\frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} (e^{w_i x_t + b_i})\right), & i \neq y_t \end{cases}$$

In each update we calculate the gradients with a single example chosen i.i.d from the rest:

$$W^t \leftarrow W^{t-1} - \eta \begin{cases} \left(\frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} (e^{w_{y_t} x_t + b_{y_t}}) x_t - x_t\right), & i = y_t \\ \left(\frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} (e^{w_i x_t + b_i}) x_t\right), & i \neq y_t \end{cases}$$

$$b^t \leftarrow b^{t-1} - \eta \begin{cases} \left(\frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} (e^{w_{y_t} x_t + b_{y_t}}) - 1\right), & i = y_t \\ \left(\frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} (e^{w_i x_t + b_i})\right), & i \neq y_t \end{cases}$$

Chapter 2

Practical Part

The code is in attached files as instructed.

The following graph shows us the differences between normal distribution and our $\hat{y}$ which represents our 'learned' distribution.

We can see after running the code for different epochs that we converging as the epochs increase as expected.