Machine Learning Ex2 theoretical part Hodaya Koslowsky 313377673

1) The probability of being correct for example t is:

$$P(y=i|x_t) = softmax(wx_t + b)_i = \frac{e^{w_i x_t + b_i}}{\sum_{j=1}^{k} e^{w_j x_t + b_j}}$$

2) The Loss function is:

$$L(w;b) = \arg min \sum_t -\log(\frac{e^{w_i x_t + b_i}}{\sum_{j=1}^{k} e^{w_j x_t + b_j}})$$

3) The update rule will be according to the derivative of the loss function w.r.t- *w*, and w.r.t- *b*. For SGD we will calculate the derivative for a single example t:

$$\frac{\partial L}{\partial w} = \frac{\partial\left(-\log\left(\frac{e^{w_i x_t + b_i}}{\sum_{j=1}^{k} e^{w_j x_t + b_j}}\right)\right)}{\partial w} = \frac{\partial\left(-\log\left(e^{w_i x_t + b_i}\right) + \log(\sum_{j=1}^{k} e^{w_j x_t + b_j})\right)}{\partial w}$$

$$= \frac{\partial\left(-(w_i x_t + b_i) + \log(\sum_{j=1}^{k} e^{w_j x_t + b_j})\right)}{\partial w} = \frac{\partial\left(\log\left(\sum_{j=1}^{k} e^{w_j x_t + b_j}\right) - w_i x_t - b_i\right)}{\partial w}$$

When we derive w.r.t the vector of the correct class i:

$$\frac{\partial\left(\log\left(\sum_{j=1}^{k} e^{w_j x_t + b_j}\right) - w_i x_t - b_i\right)}{\partial w_i} = \frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} e^{w_i x_t + b_i} * x_t - x_t$$

$$\frac{e^{w_i x_t + b_i}}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} * x_t - x_t = x_t * softmax(wx_t + b)_i - x_t$$

$$= \boldsymbol{x_t(softmax(wx_t + b)_i - 1)}$$

When we derive all the other w_a vectors, when a != i, we get:

$$\frac{\partial\left(\log\left(\sum_{j=1}^{k} e^{w_j x_t + b_j}\right) - w_i x_t - b_i\right)}{\partial w_a} = \frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} e^{w_a x_t + b_a} * x_t = \frac{e^{w_a x_t + b_a}}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} x_t$$

$$= \boldsymbol{x_t(softmax(wx_t + b)_a)}$$

Now let's derive the loss w.r.t b: When we are in the b of the correct class, b_i:

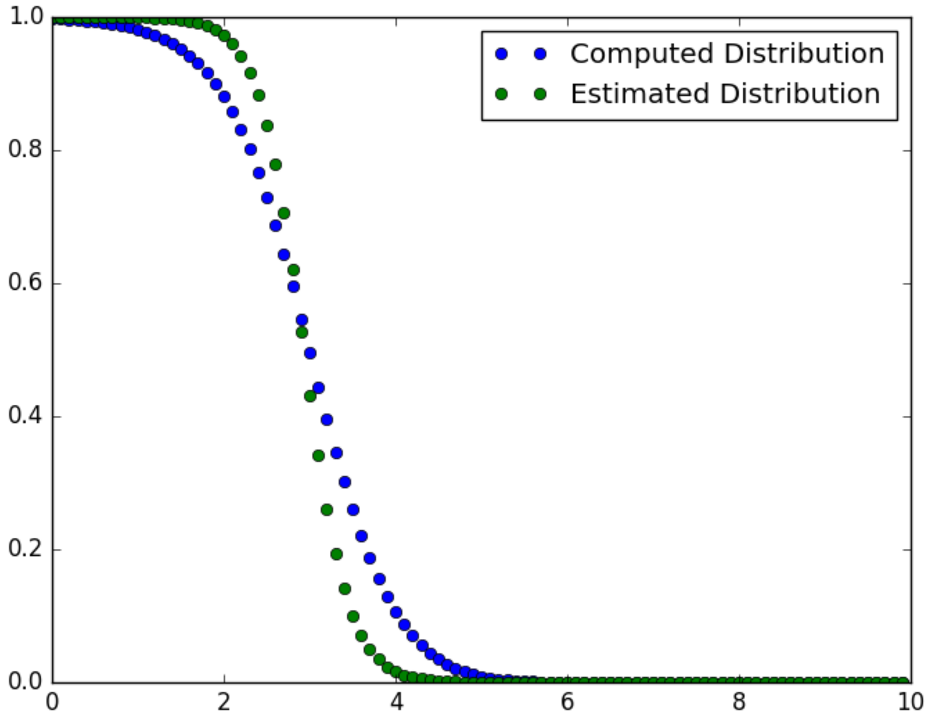
$$\frac{\partial\left(\log\left(\sum_{j=1}^{k} e^{w_j x_t + b_j}\right) - w_i x_t - b_i\right)}{\partial b_i} = \frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} e^{w_i x_t + b_i} - 1$$

$$= \boldsymbol{softmax(wx_t + b)_i - 1}$$

And w.r.t all others b_a:

$$\frac{\partial\left(\log\left(\sum_{j=1}^{k} e^{w_j x_t + b_j}\right) - w_i x_t - b_i\right)}{\partial b_a} = \frac{1}{\sum_{j=1}^{k} e^{w_j x_t + b_j}} e^{w_a x_t + b_a}$$

$$= \boldsymbol{softmax(wx_t + b)_a}$$



1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
Computed Distribution
Estimated Distribution