Machine Learning- Exercise 2
Theoretical Part

Name: Tomer Gill (תומר גיל)
ID: 318459450
U2 Username: gilltom

Date: 31/03/2018

Contents

Multiclass Logistic Regression

(a) $P(Y=y|X=\boldsymbol{x}) = softmax(\boldsymbol{Wx}+\boldsymbol{b})_{[y]} = \frac{e^{w_{[y]}x+b_{[y]}}}{\sum_i e^{w_{[i]}x+b_{[i]}}}$, where $[i]$ indicates the i-th index of the vector / matrix, **W** the weights matrix and **b** the bias vector.

(b) $\Theta^* = \underset{\Theta}{argmin}\frac{1}{m}\sum_{t=1}^{m} -\log\big(P(Y_t=y|X_t=\boldsymbol{x})\big) = \underset{\Theta}{argmin}\sum_t -\log\big(softmax(\boldsymbol{Wx}+\boldsymbol{b})_{[y]}\big) \overset{\boldsymbol{w}:=\boldsymbol{W}_{[y]},b:=\boldsymbol{b}_{[y]}}{\triangleq} \underset{\Theta}{argmin}\sum_t -\log\left(\frac{e^{\boldsymbol{wx}+b}}{\sum_i e^{\boldsymbol{W}_{[i]}\boldsymbol{x}+\boldsymbol{b}_{[i]}}}\right) = \underset{\Theta}{argmin}\sum_t -\frac{\boldsymbol{wx}+b}{\log(\sum_i \boldsymbol{W}_{[i]}\boldsymbol{x}+\boldsymbol{b}_{[i]})}$

(c) I'll calculate the gradients of the average loss for all the training examples w.r.t each weight ($w_1,\dots,w_n -\ n\ is\ the\ number\ of\ classes$) and the bias ($b$):

Weights: $\nabla_{w_i}loss(\boldsymbol{W};\boldsymbol{b}) = \nabla_{w_i}\frac{1}{m}\sum_{t=1}^{m} -\log\big(P(Y_t=y|X_t=\boldsymbol{x})\big) = \frac{1}{m}\sum_{t=1}^{m}\nabla_{w_i} -\log\big(P(Y_t=y|X_t=\boldsymbol{x})\big) = \frac{1}{m}\sum_{t=1}^{m}\nabla_{w_i} -\log\left(\frac{e^{\boldsymbol{w_y}\boldsymbol{x}+b_y}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}}\right) = \frac{1}{m}\sum_{t=1}^{m} -\nabla_{w_i}\log\big(e^{\boldsymbol{w_y}\boldsymbol{x}+b_y}\big)+\nabla_{w_i}\log\big(\sum_{j=1}^{n} e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}\big) = \frac{1}{m}\sum_{t=1}^{m} -\nabla_{w_i}(\boldsymbol{w_y x}+b_y)+\frac{1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}}\nabla_{w_i}\big(\sum_{j=1}^{n} e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}\big)$

We'll call the derivative of the loss of example t as $\nabla_{w_i}loss_t(\boldsymbol{W};\boldsymbol{b}) = -\nabla_{w_i}(\boldsymbol{w_y x}+b_y)+\frac{1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}}\nabla_{w_i}\big(\sum_{j=1}^{n} e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}\big)$, so in total $\nabla_{w_i}loss(\boldsymbol{W};\boldsymbol{b}) = \frac{1}{m}\sum_{t=1}^{m} loss_t(\boldsymbol{W};\boldsymbol{b})$. For $i=y$ (where y is the correct tag), $loss_t(\boldsymbol{W};\boldsymbol{b}) = -\nabla_{w_y}(\boldsymbol{w_y x}+b_y)+\frac{1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}}\nabla_{w_y}\big(\sum_{j=1}^{n} e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}\big) = -\boldsymbol{x}+\frac{1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}}\big(\sum_{j=1}^{n}\nabla_{w_y}e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}\big) = -\boldsymbol{x}+\frac{1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}}\Big(\nabla_{w_y}e^{\boldsymbol{W}_0\boldsymbol{x}+\boldsymbol{b}_0}+\cdots\nabla_{w_y}e^{\boldsymbol{W}_y\boldsymbol{x}+\boldsymbol{b}_y}+\cdots\nabla_{w_y}e^{\boldsymbol{W}_n\boldsymbol{x}+\boldsymbol{b}_n}\Big) = -\boldsymbol{x}+\frac{1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j\boldsymbol{x}+\boldsymbol{b}_j}}\big(e^{\boldsymbol{W}_y\boldsymbol{x}+\boldsymbol{b}_y}\boldsymbol{x}\big) =$

$\boldsymbol{x}\left(\frac{e^{\boldsymbol{W}_y \boldsymbol{x}+\boldsymbol{b}_y}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} - 1\right) = \boldsymbol{x}\left(softmax_{[y]}(\boldsymbol{Wx}+\boldsymbol{b}) - 1\right)$. For $i \neq y$: $loss_t(\boldsymbol{W}; \boldsymbol{b}) = -\nabla_{w_i}(\boldsymbol{w_y x} + b_y) + \frac{1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} \nabla_{w_i}\left(\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}\right) = -0 + \frac{1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} \left(\sum_{j=1}^{n} \nabla_{w_y} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}\right) = \frac{e^{\boldsymbol{W}_i \boldsymbol{x}+\boldsymbol{b}_i}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} \boldsymbol{x} = softmax(\boldsymbol{Wx}+\boldsymbol{b})_{[i]} * \boldsymbol{x}$.

In conclusion, $\nabla_{w_i} loss(\boldsymbol{W}; \boldsymbol{b}) = \frac{1}{m}\sum_{t=1}^{m} \nabla_{w_i} loss_t(\boldsymbol{W}; \boldsymbol{b})$,

where $\nabla_{w_i} loss_t(\boldsymbol{W}; \boldsymbol{b}) = \begin{cases} -\boldsymbol{x} + \frac{e^{\boldsymbol{W}_y \boldsymbol{x}+\boldsymbol{b}_y}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} \boldsymbol{x}, & i = y \\ \frac{e^{\boldsymbol{W}_i \boldsymbol{x}+\boldsymbol{b}_i}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} \boldsymbol{x}, & i \neq y \end{cases}$.

Bias: $\nabla_{b_i} loss(\boldsymbol{W}; \boldsymbol{b}) = \nabla_{b_i} \frac{1}{m} \sum_{t=1}^{m} -\log\left(P(Y_t = y | X_t = \boldsymbol{x})\right) = \frac{1}{m}\sum_{t=1}^{m} \nabla_{b_i} - \log\left(P(Y_t = y | X_t = \boldsymbol{x})\right) = \frac{1}{m}\sum_{t=1}^{m} \nabla_{b_i} - \log\left(\frac{e^{\boldsymbol{w}_y \boldsymbol{x}+b_y}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}}\right) = \frac{1}{m}\sum_{t=1}^{m} - \nabla_{b_i}\log\left(e^{\boldsymbol{w}_y \boldsymbol{x}+b_y}\right) + \nabla_{b_i}\log\left(\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}\right) = \frac{1}{m}\sum_{t=1}^{m} -\nabla_{b_i}(\boldsymbol{w_y x} + b_y) + \frac{1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} \nabla_{b_i}\left(\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}\right)$, so like before we'll look at the case $i = y$ for a single example t:
$-\nabla_{b_y}(\boldsymbol{w_y x} + b_y) + \frac{1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} \nabla_{b_y}\left(\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}\right) = -1 + \frac{1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} \left(\nabla_{b_y} e^{\boldsymbol{W}_0 \boldsymbol{x}+\boldsymbol{b}_0} + \cdots \nabla_{b_y} e^{\boldsymbol{W}_y \boldsymbol{x}+\boldsymbol{b}_y} + \cdots\ \nabla_{b_y} e^{\boldsymbol{W}_n \boldsymbol{x}+\boldsymbol{b}_n}\right) = -1 + \frac{e^{\boldsymbol{W}_y \boldsymbol{x}+\boldsymbol{b}_y} * 1}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} = softmax(\boldsymbol{Wx}+\boldsymbol{b})_{[y]} - 1$, and for $i \neq y$ it's the same without the -1.

For conclusion, $\nabla_{b_i} loss(\boldsymbol{W}; \boldsymbol{b}) = \frac{1}{m}\sum_{t=1}^{m} \nabla_{b_i} loss_t(\boldsymbol{W}; \boldsymbol{b})$,

where $\nabla_{b_i} loss_t(\boldsymbol{W}; \boldsymbol{b}) = \begin{cases} \frac{e^{\boldsymbol{W}_y \boldsymbol{x}+\boldsymbol{b}_y}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} - 1, & i = y \\ \frac{e^{\boldsymbol{W}_i \boldsymbol{x}+\boldsymbol{b}_i}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}}, & i \neq y \end{cases}$.

Now, for the update rules: the general SGD update rule is $\boldsymbol{w_i} = \eta \nabla_{\boldsymbol{w_i}} loss(\boldsymbol{W}; \boldsymbol{b})$ for each row $\boldsymbol{w_i}$ in the weight matrix $\boldsymbol{W}$, and $b_i = \eta \nabla_{b_i} loss(\boldsymbol{W}; \boldsymbol{b})$ for each element b_i in the bias vector $\boldsymbol{b}$. So the update rules are:

$$\boldsymbol{w_i} = \begin{cases} -\eta \boldsymbol{x} + \eta \frac{e^{\boldsymbol{W}_y \boldsymbol{x}+\boldsymbol{b}_y}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} \boldsymbol{x}, & i = y \\ \eta \frac{e^{\boldsymbol{W}_i \boldsymbol{x}+\boldsymbol{b}_i}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} \boldsymbol{x}, & i \neq y \end{cases}$$

$$b_i = \begin{cases} \eta \frac{e^{\boldsymbol{W}_y \boldsymbol{x}+\boldsymbol{b}_y}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}} - \eta, & i = y \\ \eta \frac{e^{\boldsymbol{W}_i \boldsymbol{x}+\boldsymbol{b}_i}}{\sum_{j=1}^{n} e^{\boldsymbol{W}_j \boldsymbol{x}+\boldsymbol{b}_j}}, & i \neq y \end{cases}$$

Practical Part – Graph