

Machine Learning- Exercise 2

Theoretical Part

Name: Tomer Gill (תומר גיל)
 ID: 318459450
 U2 Username: gilltom
 Date: 31/03/2018

Contents

- 1..... Multiclass Logistic Regression
 3..... Practical Part – Graph

Multiclass Logistic Regression

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

(b) $\Theta^* = \underset{\Theta}{\operatorname{argmin}} \frac{1}{m} \sum_{t=1}^m -\log(P(Y_t = y|X_t = \mathbf{x})) =$
 $\underset{\Theta}{\operatorname{argmin}} \sum_t -\log(\text{softmax}(\mathbf{W}\mathbf{x} + \mathbf{b})_{[y]}) \stackrel{\mathbf{w} := \mathbf{w}_{[y]}, \mathbf{b} := b_{[y]}}{=} \underset{\Theta}{\operatorname{argmin}} \sum_t -\log\left(\frac{e^{\mathbf{w}\mathbf{x} + b}}{\sum_i e^{\mathbf{w}_{[i]}\mathbf{x} + b_{[i]}}}\right) = \underset{\Theta}{\operatorname{argmin}} \sum_t -\mathbf{w}\mathbf{x} - b + \log(\sum_i \mathbf{W}_{[i]}\mathbf{x} + \mathbf{b}_{[i]}),$
 where Θ is \mathbf{W} and \mathbf{b} .

(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):

$$\begin{aligned} \text{Weights: } \nabla_{w_i} \text{loss}(\mathbf{W}; \mathbf{b}) &= \nabla_{w_i} \frac{1}{m} \sum_{t=1}^m -\log(P(Y_t = y|X_t = \mathbf{x})) = \\ \frac{1}{m} \sum_{t=1}^m \nabla_{w_i} -\log(P(Y_t = y|X_t = \mathbf{x})) &= \frac{1}{m} \sum_{t=1}^m \nabla_{w_i} -\log\left(\frac{e^{\mathbf{w}_y\mathbf{x} + b_y}}{\sum_{j=1}^n e^{\mathbf{w}_j\mathbf{x} + b_j}}\right) = \\ \frac{1}{m} \sum_{t=1}^m -\nabla_{w_i} \log(e^{\mathbf{w}_y\mathbf{x} + b_y}) + \nabla_{w_i} \log(\sum_{j=1}^n e^{\mathbf{w}_j\mathbf{x} + b_j}) &= \\ \frac{1}{m} \sum_{t=1}^m -\nabla_{w_i} (\mathbf{w}_y\mathbf{x} + b_y) + \frac{1}{\sum_{j=1}^n e^{\mathbf{w}_j\mathbf{x} + b_j}} \nabla_{w_i} (\sum_{j=1}^n e^{\mathbf{w}_j\mathbf{x} + b_j}) \end{aligned}$$

We'll call the derivative of the loss of example t as $\nabla_{w_i} \text{loss}_t(\mathbf{W}; \mathbf{b}) =$

$$-\nabla_{w_i} (\mathbf{w}_y\mathbf{x} + b_y) + \frac{1}{\sum_{j=1}^n e^{\mathbf{w}_j\mathbf{x} + b_j}} \nabla_{w_i} (\sum_{j=1}^n e^{\mathbf{w}_j\mathbf{x} + b_j}), \text{ so in total } \nabla_{w_i} \text{loss}(\mathbf{W}; \mathbf{b}) =$$

$$\frac{1}{m} \sum_{t=1}^m \text{loss}_t(\mathbf{W}; \mathbf{b}). \text{ For } i = y \text{ (where } y \text{ is the correct tag), } \text{loss}_t(\mathbf{W}; \mathbf{b}) =$$

$$-\nabla_{w_y} (\mathbf{w}_y\mathbf{x} + b_y) + \frac{1}{\sum_{j=1}^n e^{\mathbf{w}_j\mathbf{x} + b_j}} \nabla_{w_y} (\sum_{j=1}^n e^{\mathbf{w}_j\mathbf{x} + b_j}) = -\mathbf{x} +$$

$$\frac{1}{\sum_{j=1}^n e^{\mathbf{w}_j\mathbf{x} + b_j}} (\sum_{j=1}^n \nabla_{w_y} e^{\mathbf{w}_j\mathbf{x} + b_j}) = -\mathbf{x} + \frac{1}{\sum_{j=1}^n e^{\mathbf{w}_j\mathbf{x} + b_j}} (\nabla_{w_y} e^{\mathbf{w}_0\mathbf{x} + b_0} +$$

$$\dots \nabla_{w_y} e^{\mathbf{w}_y\mathbf{x} + b_y} + \dots \nabla_{w_y} e^{\mathbf{w}_n\mathbf{x} + b_n}) = -\mathbf{x} + \frac{1}{\sum_{j=1}^n e^{\mathbf{w}_j\mathbf{x} + b_j}} (e^{\mathbf{w}_y\mathbf{x} + b_y} \mathbf{x}) =$$

$$\mathbf{x} \left(\frac{e^{W_y \mathbf{x} + b_y}}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} - 1 \right) = \mathbf{x} (\text{softmax}_{[y]}(\mathbf{W}\mathbf{x} + \mathbf{b}) - 1). \text{ For } i \neq y: \text{loss}_t(\mathbf{W}; \mathbf{b}) =$$

$$-\nabla_{w_i}(\mathbf{w}_y \mathbf{x} + b_y) + \frac{1}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} \nabla_{w_i}(\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}) = -0 +$$

$$\frac{1}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} (\sum_{j=1}^n \nabla_{w_i} e^{W_j \mathbf{x} + b_j}) = \frac{e^{W_i \mathbf{x} + b_i}}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} \mathbf{x} = \text{softmax}(\mathbf{W}\mathbf{x} + \mathbf{b})_{[i]} * \mathbf{x}.$$

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

where $\nabla_{w_i} \text{loss}_t(\mathbf{W}; \mathbf{b}) = \begin{cases} -\mathbf{x} + \frac{e^{W_y \mathbf{x} + b_y}}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} \mathbf{x}, i = y \\ \frac{e^{W_i \mathbf{x} + b_i}}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} \mathbf{x}, i \neq y \end{cases}.$

Bias: $\nabla_{b_i} \text{loss}(\mathbf{W}; \mathbf{b}) = \nabla_{b_i} \frac{1}{m} \sum_{t=1}^m -\log(P(Y_t = y | X_t = \mathbf{x})) = \frac{1}{m} \sum_{t=1}^m \nabla_{b_i} -$

$$\log(P(Y_t = y | X_t = \mathbf{x})) = \frac{1}{m} \sum_{t=1}^m \nabla_{b_i} - \log\left(\frac{e^{W_y \mathbf{x} + b_y}}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}}\right) =$$

$$\frac{1}{m} \sum_{t=1}^m -\nabla_{b_i} \log(e^{W_y \mathbf{x} + b_y}) + \nabla_{b_i} \log(\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}) =$$

$$\frac{1}{m} \sum_{t=1}^m -\nabla_{b_i}(\mathbf{w}_y \mathbf{x} + b_y) + \frac{1}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} \nabla_{b_i}(\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}), \text{ so like before we'll}$$

look at the case $i = y$ for a single example t :

$$-\nabla_{b_y}(\mathbf{w}_y \mathbf{x} + b_y) + \frac{1}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} \nabla_{b_y}(\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}) = -1 +$$

$$\frac{1}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} (\nabla_{b_y} e^{W_0 \mathbf{x} + b_0} + \dots \nabla_{b_y} e^{W_y \mathbf{x} + b_y} + \dots \nabla_{b_y} e^{W_n \mathbf{x} + b_n}) = -1 +$$

$$\frac{e^{W_y \mathbf{x} + b_y} * 1}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} = \text{softmax}(\mathbf{W}\mathbf{x} + \mathbf{b})_{[y]} - 1, \text{ and for } i \neq y \text{ it's the same without the } -1.$$

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

where $\nabla_{b_i} \text{loss}_t(\mathbf{W}; \mathbf{b}) = \begin{cases} \frac{e^{W_y \mathbf{x} + b_y}}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} - 1, i = y \\ \frac{e^{W_i \mathbf{x} + b_i}}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}}, i \neq y \end{cases}.$

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

$$\mathbf{w}_i = \begin{cases} -\eta \mathbf{x} + \eta \frac{e^{W_y \mathbf{x} + b_y}}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} \mathbf{x}, i = y \\ \eta \frac{e^{W_i \mathbf{x} + b_i}}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} \mathbf{x}, i \neq y \end{cases}$$

$$b_i = \begin{cases} \eta \frac{e^{W_y \mathbf{x} + b_y}}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}} - \eta, i = y \\ \eta \frac{e^{W_i \mathbf{x} + b_i}}{\sum_{j=1}^n e^{W_j \mathbf{x} + b_j}}, i \neq y \end{cases}$$

Practical Part – Graph

