Homework 1 - Introduction to Deep Learning

Kaleb Roncatti de Souza, Gabriel Teixeira Callado and Amanda de Sousa Martins

January 16, 2020

1 Exercise 1. Online gradient descent: impact of an update

1.1 Compute the gradient of the pointwise loss $w.r.t$ $\mathbf{w}$

Answer:

We are interested in computing $\frac{\partial l(c_i, xi, \theta_i)}{\partial \mathbf{w}}$ =?. We know that $y_i = \frac{1}{1+e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)}}$ and $l(c_i, xi, \theta_i) = -(c_i \log(y_i) + (1-c_i)\log(1-y_i))$. We will then calculate the derivative of $l(c_i, xi, \theta_i)$ to compute the j-th term of the gradient. For simplification reasons we are going to distinguish the cases $c = 0$ and $c = 1$. Later we will return to the generic case.

For c=0

$$l(0, xi, \theta_i) = -\log(1-y_i) = -\log\left[1 - \frac{1}{1+e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)}}\right]$$

Using the Chain's rule to derive we will get:

$$\frac{\partial l(c_i, xi, \theta_i)}{\partial \mathbf{w}} = -\left[\frac{1+e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)}}{e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)}}\right]\cdot\left[\frac{(-1)(-1)}{(1+e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)})^2}\right].(-x_{ij}).e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)} = \frac{x_{ij}}{(1+e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)})} = x_{ij}.y_i$$

For c=1

$$l(1, xi, \theta_i) = -\log y_i = -\log\left[\frac{1}{1+e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)}}\right]$$

Again, through the Chain's Rule:

$$\frac{\partial l(c_i, xi, \theta_i)}{\partial \mathbf{w}} = -\left[\frac{1+e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)}}{1}\right]\cdot\left[\frac{-1}{(1+e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)})^2}\right].(-x_{ij}).e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)} = \frac{-x_{ij}.e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)}}{(1+e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)})}$$

$$= \frac{-x_{ij}.(e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)}+1-1)}{(1+e^{-(w_0+\mathbf{w}^T\cdot\mathbf{x}_i)})} = -x_{ij} + x_{ij}.y_i = x_{ij}.(-1+y_i)$$

Now, we must come back to the generic case where we did not choose the c values:

$$\frac{\partial l(c_i, xi, \theta_i)}{\partial \mathbf{w}} = c_i(x_{ij}(-1+y_i)) + (1-c_i)(x_{ij}y_i) = -c_i x_{ij} + c_i x_{ij} y_i + x_{ij} y_i - c_i x_{ij} y_i = x_{ij}(y_i - c_i)$$

If we define $err = y_i - ci$ as the deviation between the hypothesis and the expected answer we will finally get what we expected from the practical course:

$$\boxed{\frac{\partial l(c_i, xi, \theta_i)}{\partial \mathbf{w}} = x_{ij}.err}$$

1.2 Write the update in both cases (c = 1 and c = 0). Provide an interpretation of these updates.

Answer:

Picking $\eta > 0$ and calculating $\frac{\partial l(c_k, xk, \theta_k)}{\partial \mathbf{w}}$, we will be able to compute the updated w_j value:

$$w_j^{k+1} = w_j^k - \eta \frac{\partial l^k}{\partial w_j} = w_j^k - \eta x_{ij}(y_i - c_i)$$

If we are in the **c=1** condition, we will have $w_j^{k+1} = w_j^k - \eta x_{ij}(y_i - 1)$, but we know that $0 \leq y_i \leq 1$ since it is the sigmoide function. It implies that the term $\eta x_{ij}(y_i - 1)$ will always be ≥ 0. Hence, w_j^{k+1} will be greater or equal to w_j^k, representing the **positive increment**. Similarly, for the case **c=0**, we will have $w_j^{k+1} = w_j^k - \eta x_{ij} y_i$, but we know that $0 \leq y_i \leq 1$ since it is the sigmoide function. It implies that the term $\eta x_{ij} y_i$ will always be ≤ 0. Hence, w_j^{k+1} will be less or equal to w_j^k, representing the **negative increment**.

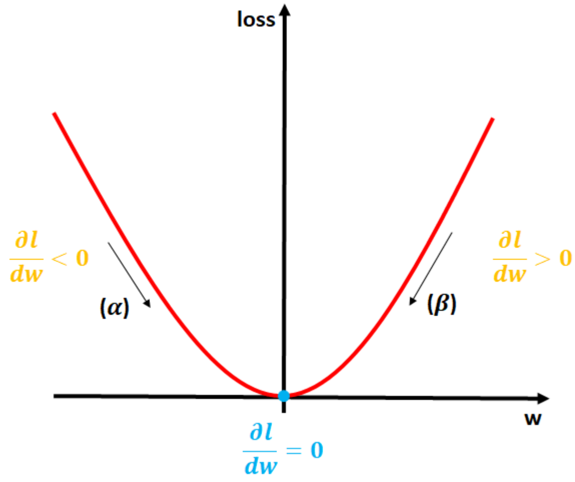
loss
$\frac{\partial l}{dw} < 0$
(α)
(β)
$\frac{\partial l}{dw} > 0$
w
$\frac{\partial l}{dw} = 0$

Figure 1: Schematic representation of the loss as a convex function. (α) In the left side, we need a positive increment in w to arrive at the minimum. (β) In the right side, we need a negative increment in w to arrive at the minimum.

We know that the loss function (l) is always convex and our interest through the logistic regression method is minimize the loss function, in other words, $\frac{\partial l}{\partial w} = 0$. Then, if we are in a region where $\frac{\partial l}{\partial w} < 0$, the increment needs to be positive to reach the minimum, since we are in the left side of the Figure 1. However, if we are in a $\frac{\partial l}{\partial w} > 0$ region, the increment needs to be negative to reach the minimum, since we are in the right side.

1.3 Show that the update indeed improves the pointwise loss:

$$l(c_i, xi, \theta_{i+1}) \leq l(c_i, xi, \theta_i)$$

We need to show that $l(c_i, xi, \theta_{i+1}) \leq l(c_i, xi, \theta_i)$. As seen in the last section, we have for c=1 : $w_j^{k+1} = w_j^k - \eta x_{ij}(y_i - c_i) \implies w_j^{k+1} = w_j^k + \alpha$. In the same way, for the case c=0, we have $w_j^{k+1} = w_j^k - \eta x_{ij}(y_i - c_i) \implies w_j^{k+1} = w_j^k - \beta$, with $\alpha \geq 0$ and $\beta \geq 0$. For the first case, c=1, the pointwise loss is described by the equation:

$$l(1, xi, \theta_{i+1}) = -\log \left[\frac{1}{1 + e^{-(w_0 + \mathbf{w}_{i+1}^T \cdot \mathbf{x}_i)}} \right]$$

We can evaluate $e^{-(w_0+\mathbf{w}_{i+1}^T\cdot\mathbf{x}_i)}$ to show the improve of the pointwise loss :

$$e^{-(w_0+\mathbf{w}_{i+1}^T\cdot\mathbf{x}_i)} = e^{-(w_0+(\mathbf{w}_i^T+\alpha)\cdot\mathbf{x}_i)} \leq e^{-(w_0+\mathbf{w}_i^T\cdot\mathbf{x}_i)}$$

$$\implies \frac{1}{1+e^{-(w_0+(\mathbf{w}_i^T+\alpha)\cdot\mathbf{x}_i)}} \geq \frac{1}{1+e^{-(w_0+\mathbf{w}_i^T\cdot\mathbf{x}_i)}}$$

$$\implies \log\frac{1}{1+e^{-(w_0+(\mathbf{w}_i^T+\alpha)\cdot\mathbf{x}_i)}} \geq \log\frac{1}{1+e^{-(w_0+\mathbf{w}_i^T\cdot\mathbf{x}_i)}}$$

$$\implies -\log\frac{1}{1+e^{-(w_0+(\mathbf{w}_i^T+\alpha)\cdot\mathbf{x}_i)}} \leq -\log\frac{1}{1+e^{-(w_0+\mathbf{w}_i^T\cdot\mathbf{x}_i)}}$$

$$\implies \boxed{l(c_i, xi, \theta_{i+1}) \leq l(c_i, xi, \theta_i)}$$

2 Exercise 2. Unrolling the updates

Answer:

For the first training example,

$$w_{k,(1)} = w_{k,(0)} - \eta\frac{\partial l(c_0, x_0, \theta_0)}{\partial w_k}$$
$$w_{k,(1)} = w_{k,(0)} - \eta x_{k,(0)}(y_{(0)} - c_{k,(0)})$$

For the next step, we calculate:

$$w_{k,(2)} = w_{k,(1)} - \eta\frac{\partial l(c_1, x_1, \theta_1)}{\partial w_k}$$
$$w_{k,(2)} = w_{k,(1)} - \eta x_{k,(1)}(y_{(1)} - c_{k,(1)})$$

if we know $w_{k,(1)}$, we can replace and calculate $w_{k,(2)}$ in function of $w_{k,(0)}$:

$$w_{k,(2)} = w_{k,(0)} - \eta x_{k,(0)}(y_{(0)} - c_{k,(0)}) - \eta x_{k,(1)}(y_{(1)} - c_{k,(1)})$$

But $x_{k,(1)} = x_{k,(0)} = x_k$ and $c_{k,(1)} = c_{k,(0)} = c_k$ because these parameters are fixed since they are, respectively, the data that we are going to treat and the pre-defined class of this data. This way,

$$w_{k,(2)} = w_{k,(0)} - \eta x_{k,(0)}(y_{(0)} - c_{k,(0)}) - \eta x_{k,(1)}(y_{(1)} - c_{k,(1)})$$
$$w_{k,(2)} = w_{k,(0)} - \eta x_k(y_{(0)} + y_{(1)} - c_k - c_k)$$
$$w_{k,(2)} = w_{k,(0)} - \eta x_k(y_{(0)} + y_{(1)} - 2c_k)$$

We can see the pattern of the equation and determinate the i-th step $w_{k,(i)}$:

$$\boxed{w_{k,(i)} = w_{k,(0)} - \eta x_k \sum_{j=0}^{k-1}(y_{(j)} - ic_k)}$$

By the last formula, it is possible to see that the i-th step $w_{k,(i)}$ is the sum of i steps - each one depends on the updated value of $y_{(i)}$. What really happens in practice is that w represents a line that separates the data in its classes. We begin with a "random line" with a random value for w, and through the logistic regression method we are able to perform iterations that improve this separation by finding another lines minimizing the loss function. An example can be seen in Figure 2.

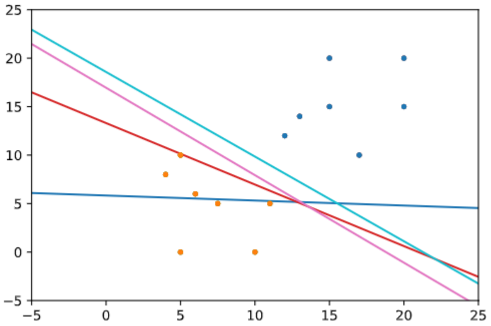

Figure 2: Schematic representation of different iterations line in the logistic regression method.