ΥΣ19 ARTIFTICIAL INTELLIGENCE II (DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING). FALL SEMESTER 2020, HOMEWORK 2

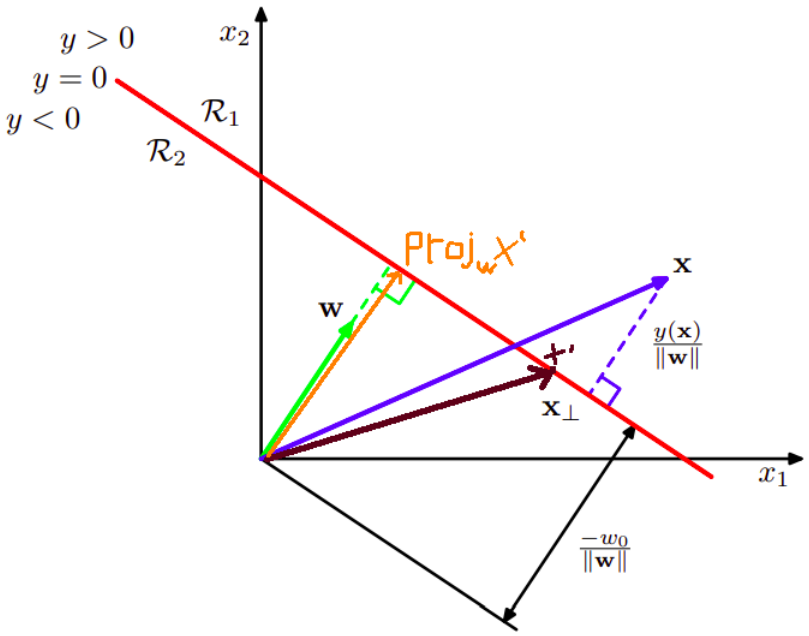
National and Kapodistrian University of Athens, DiT. 24/11/2020

Dimitrios Foteinos | sdi1700181

Contents

1 Proofs of (4.5 and 4.6) equations of the book : Patter Recognition and Machine Learning

So, for the first exercise of this homework, we had to prove the below equations (Considering the image):



For this linear function:

$$\mathbf{y(x) = w^T x + w_0}$$

We had first to prove that, if x is a point on the decision surface, then $y(x) = 0$, and so the normal distance from the origin to the decision surface is given by:

$$\frac{w^T x}{||w||} = -\frac{w_0}{||w||}$$

Main Rule: To find a distance from a given point , to a line (y function), we simply project the point onto the vector, perpendicular to line, and find the lengh of this projection.

In our image, we want to find the distance from the black line to the red one (decision boundary). To do this, we simply select a point x' on the line y, and find the length of **projection** of vector $\vec{x'}$ onto vector $\vec{w}$, which is:

$$||\mathbf{proj}_{\tilde{\mathbf{w}}}\tilde{\mathbf{x}'}|| = ||\tilde{\mathbf{x}'}|| \cdot \mathbf{cos}(\tilde{\mathbf{x}'}, \tilde{\mathbf{w}}) = ||\tilde{\mathbf{x}'}|| \cdot \frac{\tilde{\mathbf{w}} \cdot \tilde{\mathbf{x}'}}{||\tilde{\mathbf{w}}|| \cdot ||\tilde{\mathbf{x}'}||} = \frac{\tilde{\mathbf{w}} \cdot \tilde{\mathbf{x}'}}{||\tilde{\mathbf{w}}||}$$

Numerator is the same as $w^T x'$, and as a result, distance is $\frac{w^T x'}{||w||}$. We also know that our point x', lies on decision line, so equality $w^T x' + w_0 = 0$ holds, or $w^T x' = -w_0$. If we divide with the norm of the non-zero weight vector w, we has the following equation prooved (4.5) :

$$d = \frac{w^T x}{||w||} = \frac{-w_0}{||w||}$$

And thus, this quantity is equal with the distance from the origin to the decision surface. Also, as we know from the linear algebra, the projection of a vector onto another one, is equal to the vector itself multiplied with the cosine of the angle of the two vectors. Vector Projection. Also, we used the formula of the: Cosine Similarity

And now, the next equation we want to prove is this one: $\mathbf{x} = \mathbf{x}_\perp + \mathbf{r}\frac{\mathbf{w}}{||\mathbf{w}||}$

The text's approach is to decompose **x** into components relative to the decision boundary: **x**$\perp$ on the boundary, and something perpendicular to the boundary. Being perpendicular to the boundary, it is in the direction of **w** , but we don't know how far, so they introduce it's length as the unknown r.

Now, as we mentioned before, y is zero on the boundary so they have $y(x_\perp) = 0$. And now, just to fill in some of the details of what they say:

$$x = x_\perp + r\frac{w}{||w||}$$

$$w^T x + w_0 = w^T x_\perp + w_0 + r\frac{w^T w}{||w||}$$

$$y(x) = y(x_\perp) + r\frac{||w||^2}{||w||}$$

$$y(x) = 0 + r \cdot ||w||$$

and so $\mathbf{r} = \mathbf{y}(\mathbf{x})/||\mathbf{w}||$

Alternatively, we can think that equation as: $\mathbf{r}\frac{\mathbf{w}}{||\mathbf{w}||} = \mathbf{x} - \mathbf{x}_\perp$.

And so , if we set $z = r\frac{w}{||w||}$, we can say that geometrically, $x_\perp$ is the orthogonal projection of x onto the subspace y(x), and z is a vector orthogonal to $x_\perp$. For this problem, we used our knowledge from basic linear algebra, and more specifically we used the Orthogonal Decomposition theorem.

2 Compute partial derivative of matrix z = xW

Our mission for this exercise is that we want to compute $\frac{\partial z}{\partial x}$.

We can think of z as a function of z taking an m-dimensional vector to an n-dimensional vector. So it's Jacobian will be n x m. We also observe that:

$$z_i = \sum_{k=1}^{m} W_{ik} x_k$$

So an entry $(\frac{\partial z}{\partial x})_{ij}$ of the Jacobian will be:

$$(\frac{\partial z}{\partial x})_{ij} = \frac{\partial z_i}{\partial x_j} = \frac{\partial}{\partial x_j}\sum_{k=1}^{m} W_{ik}x_k = \sum_{k=1}^{m} W_{ik}\frac{\partial}{\partial x_j}x_k = W_{ij}$$

because $\frac{\partial}{\partial x_j}x_k = 1$ if $k = j$ and 0 if otherwise. So we see that $\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \mathbf{W}$

3 Compute partial derivative of function ŷ = σ ($\mathbf{x}^T w$)

First of all, as we know, the sigmoid function is like:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

.

In our case:

$$\hat{y} = \frac{1}{1 + e^{-x^T w}}$$

And we want to compute: $\frac{\partial \hat{y}}{\partial w}$.

So first, we set: $u = x^T w$, and we suppose that: $(.)' = \partial./\partial w$, to simplify our grammar.
So, the equation goes like this:

$$\frac{\partial \hat{y}}{\partial w} = (\frac{1}{1 + e^{-x^T w}})' = (\frac{1}{1 + e^{-u}})'$$

$$(=) - \frac{(1 + e^{-u})'}{(1 + e^{-u})^2} = \frac{-(e^{-u})'}{(1 + e^{-u})^2} = \frac{-(-u)' \cdot e^{-u}}{(1 + e^{-u})^2} = \frac{(u)' \cdot e^{-u}}{(1 + e^{-u})^2}$$

And thus, we have only to compute: $(u)'$.
To do this, we must first do some analysis.

We know in fact that, x and w are both **column** vectors, with elements: n x 1. So, x^T it's going to be a **row** vector, with elements: 1 x n. And so, u it's going to be a scalar value: $u = w_1 \cdot x_1 + ... + w_n \cdot x_n$. So, the partial derivatives with respect to w are going to be like this:

$$\frac{\partial u}{\partial w_1} = x_1, ..., \frac{\partial u}{\partial w_n} = x_n$$

And thus, we have:

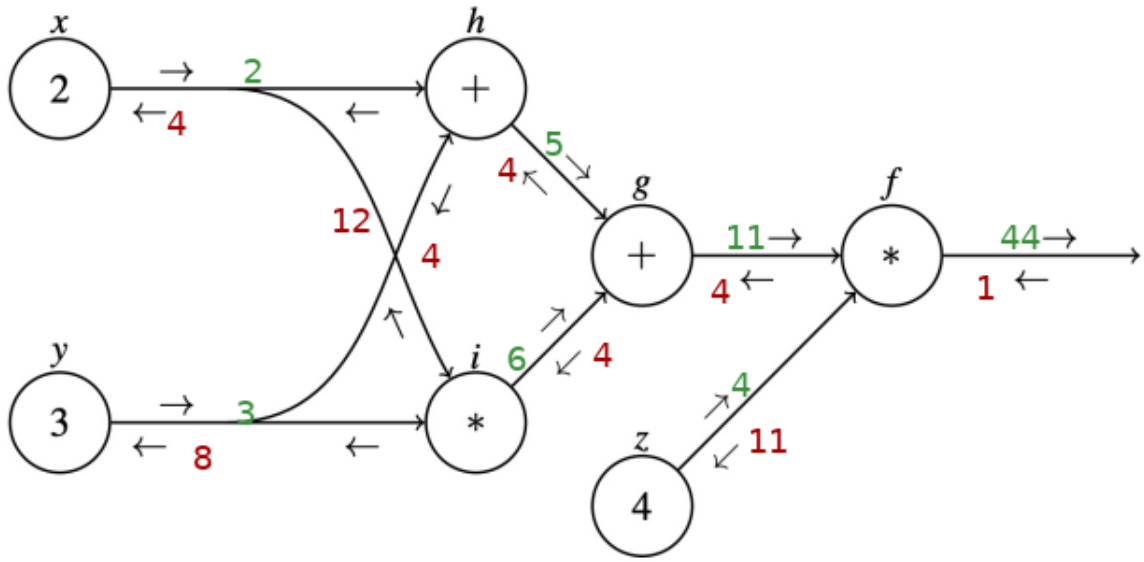
$$\mathbf{m} = (\mathbf{u})' = \frac{\partial \mathbf{u}}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial u}{\partial w_1} \\ \frac{\partial u}{\partial w_2} \\ \vdots \\ \frac{\partial u}{\partial w_n} \end{bmatrix}$$

So, the final result is:

$$\frac{(m) \cdot e^{-x^T w}}{(1 + e^{-x^T w})^2}$$

4 Forward/Backward Propagation at a computational graph

In the following image, with green, we have the **values** of the nodes in our graph, and with red, we have the **gradients** of the nodes in our graph. To prove these results, we used the: Chain Rule.



Explanation:

We have successively:

- $h = x + y = 5$
- $i = x \cdot y = 6$
- $g = h + i = 11$
- $f = z \cdot g = 44$

Having these equations, we can find now the gradients.
More specifically, we seek: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \quad for (x, y, z) = (2, 3, 4)$

And thus:
$\frac{\partial f}{\partial f} = 1, \frac{\partial f}{\partial g} = 4, \frac{\partial f}{\partial z} = 11$

Going in the deeper layers, we have:
$\frac{\partial f}{\partial h} = \frac{\partial f}{\partial g}\frac{\partial g}{\partial h} = 4, \frac{\partial f}{\partial i} = \frac{\partial f}{\partial g}\frac{\partial g}{\partial i} = 4$

And now, to find the gradients of node x and node y, we can follow 2 different ways:
$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial h}\frac{\partial h}{\partial x} = 4$ or $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial i}\frac{\partial i}{\partial x} = 12$ and for y: $\frac{\partial f}{\partial y} = \frac{\partial f}{\partial h}\frac{\partial h}{\partial y} = 4$ or: $\frac{\partial f}{\partial y} = \frac{\partial f}{\partial i}\frac{\partial i}{\partial y} = 8$