Logistic Regression: back-propagation derivation

Harsha Vardhan

February 6, 2022

Abstract

This document contains derivation of the gradients for a logistic-regression classifier, using back-propagation. For the implementation of the classifier, see the accompanying notebooks.

1 Network Architecture

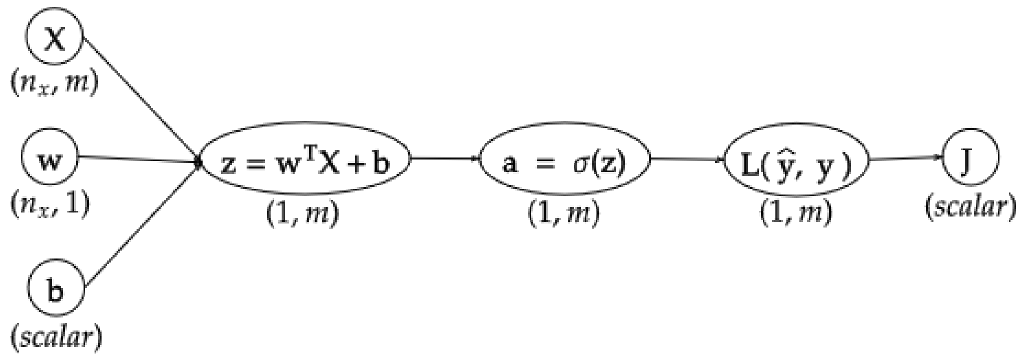

Figure 1: the computation graph along with the dimensions of each nodes' output

2 Forward Propagation

The equations for forward propagation are as follows:

$$
\begin{aligned}
\mathbf{z} &= \mathbf{w}^T\mathbf{X} + b\vec{1}_{(1,m)} \\
\mathbf{a} = \hat{\mathbf{y}} &= \frac{1}{1+e^{-\mathbf{z}}} \\
\mathbf{L}(\hat{\mathbf{y}}, \mathbf{y}) &= -\mathbf{y}\log(\hat{\mathbf{y}}) - (1-\mathbf{y})\log(1-\hat{\mathbf{y}}) \\
J &= \frac{1}{m}\sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)}) = \frac{1}{m}\mathbf{L}\vec{1}_{(m,1)}
\end{aligned}
$$

where, m	is the number of training-examples
$\mathbf{w}$	is a $(n_x, 1)$ dimensional vector
b	is a scalar-value
$\mathbf{L}$	is the loss function, and is a $(1, m)$ dimensional vector
J	is the cost-function

3 Optimization: gradient-descent

The optimization is performed according to the following equations:

$$
\begin{aligned}
\mathbf{w} &:= \mathbf{w} - \alpha\nabla_{\mathbf{w}}(J) \\
b &:= b - \alpha\nabla_{b}(J)
\end{aligned}
$$

where, α is the learning-rate/step-size.

3.1 Back-propagation

The gradients in the above equations are derived using back-propagation, as follows:

Since, J is a scalar, we can write $J = \mathrm{tr}(J) = J^\intercal = \mathrm{tr}(J^\intercal)$. And the derivative can be computed as follows:

$$\begin{aligned} \mathrm{d}J &= \mathrm{d}(\mathrm{tr}(J)) \\ &= \mathrm{tr}(\mathrm{d}J) \end{aligned}$$

The objective while computing the above derivative, is to massage the expression to the following form:

$$\mathrm{d}y = \mathrm{tr}(\mathbf{A}\mathrm{d}\mathbf{X})$$

then,

$$\frac{\mathrm{d}y}{\mathrm{d}\mathbf{X}} = \mathbf{A}$$

See [1] and [2] for more information.

3.1.1 Computing $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{L}^\intercal}$

$$\begin{aligned} \mathrm{d}J &= \mathrm{tr}(\mathrm{d}J^\intercal) \\ &= \mathrm{tr}(\mathrm{d}(\frac{1}{m}(\vec{1}_{(m,1)})^\intercal \mathbf{L}^\intercal)) \\ &= \mathrm{tr}(\frac{1}{m}(\vec{1}_{(m,1)})^\intercal \mathrm{d}\mathbf{L}^\intercal) \\ \implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{L}^\intercal} &= \frac{1}{m}(\vec{1}_{(m,1)})^\intercal \end{aligned}$$

3.1.2 Computing $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{a}^\intercal}$

$$\begin{aligned} \mathrm{d}J &= \mathrm{tr}(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{L}^\intercal}\frac{\mathrm{d}\mathbf{L}^\intercal}{\mathrm{d}\mathbf{a}^\intercal}\mathrm{d}\mathbf{a}^\intercal) \\ &= \mathrm{tr}\left(\frac{1}{m}(\vec{1}_{(m,1)})^\intercal \begin{bmatrix} \frac{\partial L^{(1)}}{\partial a^{(1)}} & 0 & \dots & 0 \\ 0 & \frac{\partial L^{(2)}}{\partial a^{(2)}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\partial L^{(m)}}{\partial a^{(m)}} \end{bmatrix} \mathrm{d}\mathbf{a}^\intercal\right) \\ \implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{a}^\intercal} &= \frac{1}{m}\left[\frac{\partial L^{(1)}}{\partial a^{(1)}}, \frac{\partial L^{(2)}}{\partial a^{(2)}}, \dots, \frac{\partial L^{(m)}}{\partial a^{(m)}}\right] \end{aligned}$$

where,

$$
\begin{aligned}
\frac{\mathrm{d}\mathbf{L}^\intercal}{\mathrm{d}\mathbf{a}^\intercal} &= \left[\frac{\partial}{\partial a^{(1)}}, \frac{\partial}{\partial a^{(2)}}, \ldots, \frac{\partial}{\partial a^{(m)}}\right] \otimes \begin{bmatrix} L^{(1)} \\ L^{(2)} \\ \vdots \\ L^{(m)} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial L^{(1)}}{\partial a^{(1)}} & 0 & \ldots & 0 \\ 0 & \frac{\partial L^{(2)}}{\partial a^{(2)}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{\partial L^{(m)}}{\partial a^{(m)}} \end{bmatrix}
\end{aligned}
$$

3.1.3 Computing $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{z}^\intercal}$

$$
\begin{aligned}
\mathrm{d}J &= \mathrm{tr}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{a}^\intercal}\frac{\mathrm{d}\mathbf{a}^\intercal}{\mathrm{d}\mathbf{z}^\intercal}\mathrm{d}\mathbf{z}^\intercal\right) \\
&= \mathrm{tr}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{a}^\intercal}\begin{bmatrix} \frac{\partial a^{(1)}}{\partial z^{(1)}} & 0 & \ldots & 0 \\ 0 & \frac{\partial a^{(2)}}{\partial z^{(2)}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{\partial a^{(m)}}{\partial z^{(m)}} \end{bmatrix}\mathrm{d}\mathbf{z}^\intercal\right) \\
&= \mathrm{tr}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{a}^\intercal} \circ \mathrm{diag}^{-1}\left(\frac{\mathrm{d}\mathbf{a}^\intercal}{\mathrm{d}\mathbf{z}^\intercal}\right)\mathrm{d}\mathbf{z}^\intercal\right) \\
\implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{z}^\intercal} &= \frac{\mathrm{d}J}{\mathrm{d}\mathbf{a}^\intercal} \circ \mathrm{diag}^{-1}\left(\frac{\mathrm{d}\mathbf{a}^\intercal}{\mathrm{d}\mathbf{z}^\intercal}\right)
\end{aligned}
$$

where,

$$
\begin{aligned}
\frac{\mathrm{d}\mathbf{a}^\intercal}{\mathrm{d}\mathbf{z}^\intercal} &= [\frac{\partial}{\partial z^{(1)}}, \frac{\partial}{\partial z^{(2)}}, \ldots, \frac{\partial}{\partial z^{(m)}}] \otimes \begin{bmatrix} a^{(1)} \\ a^{(2)} \\ \vdots \\ a^{(m)} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial a^{(1)}}{\partial z^{(1)}} & 0 & \ldots & 0 \\ 0 & \frac{\partial a^{(2)}}{\partial z^{(2)}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{\partial a^{(m)}}{\partial z^{(m)}} \end{bmatrix}
\end{aligned}
$$

3.1.4 Computing $\frac{\mathrm{d}J}{\mathrm{d}\mathbf{w}}$, and $\frac{\mathrm{d}J}{\mathrm{d}b}$

$$\mathrm{d}J = \mathrm{tr}(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{z}^\intercal}\mathrm{d}\mathbf{z}^\intercal)$$

where, $\mathrm{d}\mathbf{z}^\intercal$ may be expanded as,

$$\mathrm{d}\mathbf{z}^\intercal = \mathrm{d}\mathbf{X}^\intercal\mathbf{w} + \mathbf{X}^\intercal\mathrm{d}\mathbf{w} + (\vec{1}_{(1,m)})^\intercal\mathrm{d}b$$

Here, $\mathrm{d}\mathbf{X}^\intercal = 0$, since, $\mathbf{X}$ is the input. And when differentiating w.r.t $\mathbf{w}$, we have $\mathrm{d}b = 0$. So,

$$\mathrm{d}J = \mathrm{tr}(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{z}^\intercal}\mathbf{X}^\intercal\mathrm{d}\mathbf{w})$$

$$\implies \frac{\mathrm{d}J}{\mathrm{d}\mathbf{w}} = \frac{\mathrm{d}J}{\mathrm{d}\mathbf{z}^\intercal}\mathbf{X}^\intercal$$

when differentiating w.r.t b, we have $\mathrm{d}\mathbf{w} = 0$. So,

$$\mathrm{d}J = \mathrm{tr}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{z}^\intercal}(\vec{1}_{(1,m)})^\intercal\mathrm{d}b\right)$$

$$\implies \frac{\mathrm{d}J}{\mathrm{d}b} = \frac{\mathrm{d}J}{\mathrm{d}\mathbf{z}^\intercal}(\vec{1}_{(1,m)})^\intercal$$

3.2 Jacobian or Gradient?

In the above derivations, we have used the numerator layout while performing matrix-derivatives. Therefore, the derivatives are jacobians and not gradients. And the corresponding gradients are simply transpose of jacobians, i.e.,

$$\begin{aligned}\nabla_{\mathbf{w}}(J) &= \left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{w}}\right)^\intercal &&= \mathbf{X}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{z}^\intercal}\right)^\intercal \\ \nabla_{b}(J) &= \left(\frac{\mathrm{d}J}{\mathrm{d}b}\right)^\intercal &&= \vec{1}_{(1,m)}\left(\frac{\mathrm{d}J}{\mathrm{d}\mathbf{z}^\intercal}\right)^\intercal\end{aligned}$$

References

[1] T. Minka, "Old and new matrix algebra useful for statistics," Sep. 1997. [Online]. Available: `https://www.microsoft.com/en-us/research/publication/old-new-matrix-algebra-useful-statistics/`.

[2] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Second. John Wiley, 1999, ISBN: 0471986321 9780471986324 047198633X 9780471986331.