Homework 4. Supervised learning

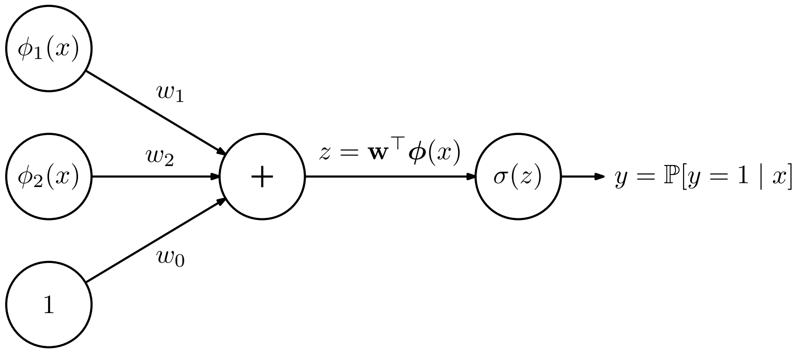

Figure 1: Logistic regression as a neural network with a single neuron.

In Lab 4 you will use supervised learning to solve a real-world classification problem. To prepare for the lab, in this homework you will perform some preliminary computations that will be useful in the lab.

Logistic regression is an approach to classification that estimates the probability of each of two actions, $\mathcal{A} = \{-1, 1\}$, given a set of examples, $\mathcal{D} = \{(x_1, a_1), \ldots, (x_N, a_N)\}$, with $a_n \in \mathcal{A}, n = 1, \ldots, N$ and where each state x_n is described by a number of features $\phi_1, \ldots, \phi_K$. In logistic regression, we assume that the probability of the action $a = 1$ takes the form

$$\pi(1 \mid x) \stackrel{\text{def}}{=} \mathbb{P}\left[\mathrm{a} = 1 \mid \mathrm{x} = x\right] = \sigma(z),$$

where σ is the logistic function

$$\sigma(z) = \frac{1}{1 + \exp(-z)},$$

and z is a weighted sum of the input features, i.e.,

$$z = w_0 + \sum_{k=1}^{K} w_k \phi_k(x) = \mathbf{w}^\top \boldsymbol{\phi}(x),$$

with

$$\mathbf{w} = \begin{bmatrix} w_0 & w_1 & \dots & w_K \end{bmatrix}^\top, \qquad \phi(x) = \begin{bmatrix} 1 & \phi_1(x) & \dots & \phi_K(x) \end{bmatrix}^\top.$$

The logistic regression classifier can also be seen as a very simple neural network with a single neuron, as depicted in Fig. 1.

In practice, training logistic regression consists of finding the parameters $\mathbf{w}$ that minimize the empirical risk

$$\hat{L}_N(\pi) = -\frac{1}{N}\sum_{n=1}^{N} \log \pi(a_n \mid x_n)$$

which can be done, for example, using gradient descent.

Exercise 1.

(a) Show that

$$\hat{L}_N(\pi) = \frac{1}{N}\sum_{n=1}^{N} \log\big(\exp(-a_n \mathbf{w}^\top \boldsymbol{\phi}(x_n)) + 1\big).$$

(b) Using the expression from (a), show that

$$\mathbf{g} = \nabla_{\mathbf{w}} \hat{L}_N(\pi) = \frac{1}{N}\sum_{n=1}^{N} a_n \boldsymbol{\phi}(x_n)(\pi(a_n \mid x_n) - 1).$$

(c) Show that

$$\mathbf{H} = \nabla^2_{\mathbf{w}} \hat{L}_N(\pi) = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\phi}(x_n)\boldsymbol{\phi}^\top(x_n)\pi(a_n \mid x_n)(1 - \pi(a_n \mid x_n)).$$

Solution 1:

(a) We have that

$$\hat{L}_N(\pi) = -\frac{1}{N}\sum_{n=1}^{N} \log \pi(a_n \mid x_n). \tag{1}$$

Let us then consider the term inside the summation. By definition, for $a = 1$,

$$\log \pi(a \mid x) = \log\left(\frac{1}{1 + \exp(-z)}\right),$$

with $z = \mathbf{w}^\top \boldsymbol{\phi}(x)$. Equivalently,

$$\log \pi(a \mid x) = -\log(1 + \exp(-z)). \tag{2}$$

Conversely, for $a = -1$,

$$\begin{aligned}\log \pi(a \mid x) &= \log\left(1 - \frac{1}{1+\exp(-z)}\right)\\ &= \log\left(\frac{\exp(-z)}{1+\exp(-z)}\right).\end{aligned}$$

Dividing both numerator and denominator by $\exp(-z)$, we get

$$\log \pi(a \mid x) = \log\left(\frac{1}{\exp(z)+1}\right) = -\log(\exp(z)+1). \tag{3}$$

Combining (3) and (4) yields

$$\log \pi(a \mid x) = -\log(1+\exp(-az)) \tag{4}$$

which, replacing in (1) leads to the desired result.

(b) We can use the chain rule of derivatives to compute $\nabla_{\mathbf{w}} \hat{L}_N(\pi)$. We have that

$$\frac{d}{dz}\big(\log(\exp(-az)+1)\big) = \frac{-a\exp(-az)}{1+\exp(-az)}.$$

Noting that

$$\frac{\exp(-az)}{1+\exp(-az)} = 1 - \pi(a \mid x)$$

yields

$$\frac{d}{dz}\big(\log(\exp(-az)+1)\big) = a(\pi(a \mid x) - 1).$$

Finally, since $\nabla_{\mathbf{w}} z = \boldsymbol{\phi}(x)$, the desired result follows.

(c) We have that

$$\nabla^2_{\mathbf{w}} \hat{L}_N(\pi) = \frac{1}{N}\sum_{n=1}^{N} a_n \boldsymbol{\phi}(x_n)\nabla^\top_{\mathbf{w}}\pi(a_n \mid x_n).$$

Again using the chain rule of derivatives,

$$\frac{d}{dz}\left(\frac{1}{1+\exp(-z)}\right) = \frac{\exp(-z)}{(1+\exp(-z))^2} = \pi(a \mid x)(1-\pi(a \mid x)).$$

Using the fact that $\nabla_{\mathbf{w}} z = \boldsymbol{\phi}(x)$, we get that

$$\nabla_{\mathbf{w}}\pi(a \mid x) = \boldsymbol{\phi}(x)\pi(a \mid x)(1-\pi(a \mid x).$$

Putting all together, the result follows.