Homework 4. Supervised learning

In Lab 4 you will use supervised learning to solve a real-world classification problem. To prepare for the lab, in the homework you will go over a toy problem that you can run by hand.

Logistic regression consists of estimating the probability of each of two actions, $\mathcal{A} = \{0, 1\}$, given a set of examples, $\mathcal{D} = \{(x_1, a_1), \dots, (x_N, a_N)\}$, with $a_n \in \mathcal{A}, n = 1, \dots, N$ and where each state x_n is described by a number of features $\phi_1, \dots, \phi_K$. In logistic regression, we assume that

$$\pi(1 \mid x) \stackrel{\text{def}}{=} \mathbb{P}\left[\mathrm{a} = 1 \mid \mathrm{x} = x\right] = \frac{1}{1 + e^{-(\boldsymbol{w}^\top \boldsymbol{\phi}(x) + w_0)}}, \tag{1}$$

where $\boldsymbol{w}$ and w_0 are the parameters to be learned.

Training logistic regression consists of finding the parameters $\boldsymbol{w}, w_0$ that minimize the *negative log likelihood* of the data, i.e.,

$$J(\pi) = -\log \prod_{n=1}^{N} \pi(a_n \mid x_n), \tag{2}$$

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

Consider the following dataset, comprising 5 points described by two attributes, ϕ_1 and ϕ_2.

$\phi_1(x)$	$\phi_2(x)$	a
1.0	3.0	0
2.0	2.0	1
2.0	1.0	0
3.0	1.0	0
5.0	4.0	1

Exercise 1.

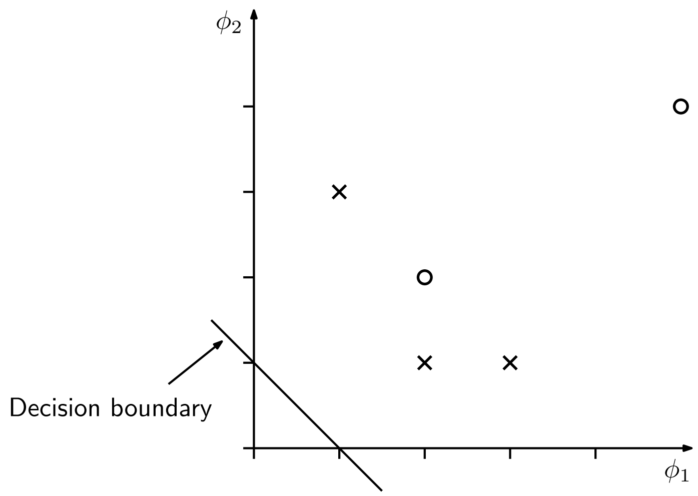
(a) Plot the training data provided. Is the data linearly separable? Why?

(b) Initializing the weights of the logistic regression classifier to 0, run 1 iteration of gradient descent with a step-size $\alpha = 1$, and indicate the resulting weights.

(c) Indicate the equation defining the decision boundary corresponding to the weights obtained after the update in Question (b), and indicate it in the plot of Question (a). Does it properly classify the points in the training set?

Solution 1:

(a) The data is not linearly separable, as no "straight line" (hyperplane) can perfectly discriminate the examples in the two classes.



(b) We have that, for the logistic regression classifier,

$$\frac{\partial J(\pi)}{\partial w_0} = \frac{1}{N}\sum_{n=1}^{N}(\pi(1 \mid x_n) - a_n)$$
$$\frac{\partial J(\pi)}{\partial w_1} = \frac{1}{N}\sum_{n=1}^{N}\phi_1(x_n)(\pi(1 \mid x_n) - a_n)$$
$$\frac{\partial J(\pi)}{\partial w_2} = \frac{1}{N}\sum_{n=1}^{N}\phi_2(x_n)(\pi(1 \mid x_n) - a_n).$$

In our case, this yields

$$\frac{\partial J(\pi)}{\partial w_0} = \frac{1}{5}(0.5 - 0.5 + 0.5 + 0.5 - 0.5) = 0.1$$
$$\frac{\partial J(\pi)}{\partial w_1} = \frac{1}{5}(0.5 - 1.0 + 1.0 + 1.5 - 2.5) = -0.1$$
$$\frac{\partial J(\pi)}{\partial w_2} = \frac{1}{5}(1.5 - 1.0 + 0.5 + 0.5 - 2.0) = -0.1$$

which corresponds to the updated weights

$$w_0 = -0.1, \qquad w_1 = 0.1, \qquad w_2 = 0.1.$$

(c) The decision boundary corresponds to the solution of

$$-0.1 + 0.1\phi_1(x) + 0.1\phi_2(x) = 0,$$

and is plotted in the diagram above. It classifies all points in class in class 1, which is to be expected after only 1 gradient descent iteration.