Homework 4. Supervised learning

In Lab 5 you will use supervised learning to solve two classification problems. To prepare for the lab, in the homework you will compute some of the necessary expressions that needed to implement one of the learning algorithms—logistic regression.

(Binary) logistic regression provides a discriminative model that estimates the probability of selecting each of two actions, $\mathcal{A} = \{0, 1\}$, given a set of examples, $\mathcal{D} = \{(\boldsymbol{x}_1, a_1), \ldots, (\boldsymbol{x}_N, a_N)\}$, with $a_n \in \mathcal{A}, n = 1, \ldots, N$ and $\boldsymbol{x}_n \in \mathbb{R}^P, n = 1, \ldots, N$. In particular, each state $\boldsymbol{x}_n$ is described as a vector with P attributes, where attribute p is a real value. Logistic regression assumes that the probability of action $a = 1$ takes the form

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

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

It is a common practice to write

$$\boldsymbol{w}^\top \boldsymbol{x} + b = \underbrace{\begin{bmatrix} \boldsymbol{w}^\top & b \end{bmatrix}}_{\text{new } \boldsymbol{w}} \underbrace{\begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix}}_{\text{new } \boldsymbol{x}}.$$

In other words, we augment $\boldsymbol{x}$ with a component $x_{P+1} \equiv 1$ so that the parameter b can be treated as one of the weights. Using this simplification, (1) becomes just

$$\pi(1 \mid \boldsymbol{x}; \boldsymbol{w}) = \frac{1}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{x}}},$$

where now $\boldsymbol{w}, \boldsymbol{x} \in \mathbb{R}^{P+1}$. We adopt this practice throughout.

Training logistic regression consists in finding the parameters $\boldsymbol{w}$ that minimize the *negative log likelihood* of the data. In other words, and assuming the examples (x_n, a_n) to be independent, we want to compute

$$\boldsymbol{w}^* = \underset{\boldsymbol{w} \in \mathbb{R}^{P+1}}{\operatorname{argmin}} \left\{ -\log \mathbb{P}\left[\mathcal{D}; \boldsymbol{w}\right] \right\} = \underset{\boldsymbol{w} \in \mathbb{R}^{P+1}}{\operatorname{argmin}} \left\{ -\log \prod_{n=1}^{N} \mathbb{P}\left[a_n \mid \boldsymbol{x}_n; \boldsymbol{w}\right] \right\}. \tag{2}$$

You will start by writing (2) as a function of the data in $\mathcal{D}$ and the probability $\pi(1 \mid \boldsymbol{x}, \boldsymbol{w})$; you will then compute the operators needed to implement a local search algorithm (namely, the Newton-Raphson algorithm).

Exercise 1.

(a) Show that

$$\boldsymbol{w}^* = \operatorname*{argmax}_{\boldsymbol{w} \in \mathbb{R}^{P+1}} \ell(\mathcal{D}; \pi),$$

where

$$\ell(\mathcal{D}; \pi) \stackrel{\text{def}}{=} \sum_{n=1}^{N} \left[a_n \log \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w}) + (1 - a_n) \log(1 - \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w})) \right].$$

(b) Show that the gradient of $\ell(\mathcal{D}; \pi)$ with respect to $\boldsymbol{w}$ (i.e., the vector of first-order derivatives) is given by

$$\mathbf{g} = \sum_{n=1}^{N} \boldsymbol{x}_n (a_n - \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w}))$$

(c) Show that the Hessian of $\ell(\mathcal{D}; \pi)$ with respect to $\boldsymbol{w}$ (i.e., the matrix of second order derivatives) is given by

$$\mathbf{H} = -\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w})(1 - \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w})).$$

Solution 1:

(a) We have that

$$\pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w})^{a_n} = \begin{cases} \pi(a_n \mid \boldsymbol{x}_n; \boldsymbol{w}) & \text{if } a_n = 1; \\ 1 & \text{otherwise.} \end{cases}$$

and

$$(1 - \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w}))^{1-a_n} = \begin{cases} \pi(a_n \mid \boldsymbol{x}_n; \boldsymbol{w}) & \text{if } a_n = 0; \\ 1 & \text{otherwise.} \end{cases}$$

Therefore,

$$\pi(a_n \mid \boldsymbol{x}_n; \boldsymbol{w}) = \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w})^{a_n} (1 - \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w}))^{1-a_n}.$$

Replacing in (2) and using the properties of the logarithm yields the desired result.

(b) We have that

$$\begin{aligned}\mathbf{g} &= \nabla \ell(\mathcal{D}; \boldsymbol{w}) \\ &= \sum_{n=1}^{N} -a_n \frac{-\boldsymbol{x}_n e^{-\boldsymbol{w}^\top \boldsymbol{x}_n}}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{x}_n}} + (1 - a_n) \left(-\boldsymbol{x}_n - \boldsymbol{x}_n \frac{-\boldsymbol{x}_n e^{-\boldsymbol{w}^\top \boldsymbol{x}_n}}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{x}_n}} \right)\end{aligned}$$

$$\begin{aligned}&= \sum_{n=1}^{N} a_n \boldsymbol{x}_n (1 - \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w}) + (1 - a_n) \boldsymbol{x}_n \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w}) \\ &= \sum_{n=1}^{N} \boldsymbol{x}_n (a_n - \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w})).\end{aligned}$$

(c) We have that

$$\begin{aligned}\mathbf{H} &= \nabla^2 \ell(\mathcal{D}; \boldsymbol{w}) \\ &= \sum_{n=1}^{N} \boldsymbol{x}_n \frac{-\boldsymbol{x}_n^\top e^{-\boldsymbol{w}^\top \boldsymbol{x}_n}}{(1 + e^{-\boldsymbol{w}^\top \boldsymbol{x}_n})^2} \\ &= -\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w})(1 - \pi(1 \mid \boldsymbol{x}_n; \boldsymbol{w})).\end{aligned}$$