Day 3. Classification. Emines.

Exercise 1 (Logistic Regression) Let (X, Y) be a couple of random variables with values in $\mathbb{R}^p \times \{0,1\}$ and $(X_i, Y_i)_{i=1,\dots,n}$ an i.i.d. sample with same distribution as (X, Y).

A classical approach is to assume a parametric model for the conditional probability $\mathbb{P}[Y = 1|X = x]$. The most popular model in $\mathbb{R}^d$ is probably the *logistic model*, where

$$\mathbb{P}[Y = 1|X = x] = \frac{\exp(\langle \beta^*, x\rangle)}{1 + \exp(\langle \beta^*, x\rangle)} \quad \text{for all } x \in \mathbb{R}^p, \tag{1}$$

with $\beta^* \in \mathbb{R}^p$. In this case, we have $\mathbb{P}[Y = 1|X = x] > 1/2$ if and only if $\langle \beta^*, x\rangle > 0$, so the frontier between $\{h_* = 1\}$ and $\{h_* = 0\}$ is again an hyperplane, with orthogonal direction β^*.

We can estimate the parameter β^* by maximizing the conditional likelihood of $(Y_1, \dots, Y_n)$ given that $(X_1, \dots, X_n) = (x_1, \dots, x_n)$:

$$\widehat{\beta} \in \operatorname*{argmax}_{\beta \in \mathbb{R}^d} \prod_{i=1}^{n} \left[\left(\frac{\exp(\langle \beta, x_i\rangle)}{1 + \exp(\langle \beta, x_i\rangle)} \right)^{Y_i} \left(\frac{1}{1 + \exp(\langle \beta, x_i\rangle)} \right)^{1-Y_i} \right],$$

and compute the classifier $\widehat{h}_{\mathsf{logistic}}(x) = \mathbf{1}_{\langle \widehat{\beta}, x\rangle > 0}$ for all $x \in \mathbb{R}^p$.

1. Check that the gradient and the Hessian $H_n(\beta)$ of

   $$\ell_n(\beta) = -\sum_{i=1}^{n} \left[Y_i \langle x_i, \beta\rangle - \log(1 + \exp(\langle x_i, \beta\rangle)) \right]$$

   are given by

   $$\nabla \ell_n(\beta) = -\sum_{i=1}^{n} \left(Y_i - \frac{e^{\langle x_i, \beta\rangle}}{1 + e^{\langle x_i, \beta\rangle}} \right) x_i \quad \text{and} \quad H_n(\beta) = \sum_{i=1}^{n} \frac{e^{\langle x_i, \beta\rangle}}{\left(1 + e^{\langle x_i, \beta\rangle}\right)^2} x_i x_i^\top.$$

2. We assume $H_n(\beta)$ to be non-singular. What can we say about the function ℓ_n?

In order to select useful features, we estimate β with the penalized criterion

$$\widehat{\beta}_\lambda \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \{\ell_n(\beta) + \lambda |\beta|_1\},$$

where $\lambda > 0$ is a regularization parameter.

Building on the Taylor expansion $\ell_n(\beta') = \ell_n(\beta) + \langle \nabla \ell_n(\beta), \beta' - \beta\rangle + O(\|\beta' - \beta\|^2)$, we compute $\widehat{\beta}_\lambda$ with the following iterations (for a given $\phi > 0$).

INIT: $\beta^0 = 0$, $t = 0$
ITERATE (until convergence)
$\beta^{t+1} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \{\ell_n(\beta^t) + \langle \nabla \ell_n(\beta^t), \beta - \beta^t\rangle + \frac{\phi}{2}\|\beta - \beta^t\|^2 + \lambda |\beta|_1\}$
$t \leftarrow t + 1$
OUTPUT: β^t

3. Check that $\beta^{t+1} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \{\|\beta - \beta^t + \phi^{-1} \nabla \ell_n(\beta^t)\|^2 + \frac{2\lambda}{\phi} |\beta|_1\}$.

4. Conclude that $\beta^{t+1} = S_{\lambda/\phi}(\beta^t - \phi^{-1} \nabla \ell_n(\beta^t))$, where $S_\mu(x) = [x_j(1 - \mu/|x_j|)_+]_{j=1,\dots,p}$.

Exercise 2 Let $h(\beta) = (\beta - u)^2 + c|\beta|$ where $u > 0$. Show that the argmin of h can be written as

$$\beta^\star = u \left(1 - \frac{c}{2|u|}\right)^+$$