PC2. Ecole Polytechnique. MAP 569. Machine Learning II.

Exercise 1 (Linear Discriminant analysis) Let (X, Y) be a couple of random variables with values in $\mathbb{R}^p \times \{0,1\}$ and a distribution

$$\mathbb{P}(Y = k) = \pi_k > 0 \quad \text{and} \quad \mathbb{P}(X \in dx | Y = k) = g_k(x)\, dx, \quad k \in \{0,1\},\ x \in \mathbb{R}^p, \tag{1}$$

where $\pi_0 + \pi_1 = 1$ and g_0, g_1 are two probability densities in $\mathbb{R}^p$.

We define the classifier $h_* : \mathbb{R}^p \to \{0,1\}$ by

$$h_*(x) = \mathbf{1}_{\{\pi_1 g_1(x) > \pi_0 g_0(x)\}},\ x \in \mathbb{R}^p.$$

1. What is the distribution of X?

2. Prove that the classifier h_* fulfills

$$\mathbb{P}(h_*(X) \neq Y) = \min_h \mathbb{P}(h(X) \neq Y).$$

3. We assume in the following that

$$g_k(x) = (2\pi)^{-p/2} \sqrt{\det(\Sigma_k^{-1})} \exp\left(-\frac{1}{2}(x - \mu_k)^\top \Sigma_k^{-1} (x - \mu_k)\right), \qquad k = 0, 1,$$

with Σ_0, Σ_1 non-singular and $\mu_0, \mu_1 \in \mathbb{R}^p$, $\mu_0 \neq \mu_1$. Prove that when $\Sigma_0 = \Sigma_1 = \Sigma$, the condition $\pi_1 g_1(x) > \pi_0 g_0(x)$ is equivalent to

$$(\mu_1 - \mu_0)^\top \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_0}{2}\right) > \log(\pi_0/\pi_1).$$

Interpret geometrically this result.

4. Assume now that π_k, μ_k, Σ are unknown, but we have a sample $(X_i, Y_i)_{i=1,\dots,n}$ i.i.d. with distribution (1). When $n > p$, propose a classifier $\hat{h} : \mathbb{R}^p \to \{0,1\}$.

5. We come back to the case where π_k, μ_k, Σ are known. If $\pi_1 = \pi_0$, check that

$$\mathbb{P}(h_*(X) = 1 | Y = 0) = \Phi(-d(\mu_1, \mu_0)/2)$$

where Φ is the cumulative distribution function of a standard Gaussian and $d(\mu_1, \mu_0)$ is the Mahalanobis distance defined by $d(\mu_1, \mu_0)^2 = (\mu_1 - \mu_0)^\top \Sigma^{-1} (\mu_1 - \mu_0)$.

6. When $\Sigma_1 \neq \Sigma_0$, what is the nature of the frontier between $\{h_* = 1\}$ and $\{h_* = 0\}$?

Exercise 2 (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).

Since the Bayes classifier only depends on the conditional distribution of Y given X, we can avoid to model the full distribution of X as in the previous exercise. 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

$$\mathbb{P}[Y = 1 | X = x] = \frac{\exp(\langle \beta^*, x \rangle)}{1 + \exp(\langle \beta^*, x \rangle)} \quad \text{for all } x \in \mathbb{R}^p, \tag{2}$$

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 $(X_1, \dots, X_n) = (x_1, \dots, x_n)$:

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

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

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

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

   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 $S_\mu(x) = [x_j(1 - \mu/|x_j|)_+]_{j=1,\dots,p}$.