Introduction to Machine Learning

LDA and logistic regression

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

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

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

\[
h^*(x) = 1\{\pi_1 g_1(x) > \pi_0 g_0(x)\}, \quad x \in \mathbb{R}^p.
\]

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

\[
g_k(x) = (2\pi)^{-n/2} \sqrt{\det(\Sigma_k)} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right), \quad k = 0, 1,
\]

with \(\Sigma_0, \Sigma_1\) 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

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

Interpret geometrically this result.

4. Assume now that \(\pi_k, \mu_k, \Sigma\) are unknown, but we have a sample \((X_i, Y_i)_{i=1,\ldots,n}\) 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 \(\pi_k, \mu_k, \Sigma\) are known. If \(\pi_1 = \pi_0\), check that

\[
\mathbb{P}(h^*(X) = 1 \mid Y = 0) = \Phi(-d(\mu_1, \mu_0)/2)
\]

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

6. When \(\Sigma_1 \neq \Sigma_0\), what is the nature of the frontier between \(\{h^*_s = 1\}\) and \(\{h_s = 0\}\)?
2 Logistic Regression

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 $P[Y = 1 | X = x]$. The most popular model in $\mathbb{R}^d$ is probably the logistic model, where

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

with $\beta^* \in \mathbb{R}^d$. In this case, we have $P[Y = 1 | X = x] > 1/2$ 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 $\beta^*$.

We can estimate the parameter $\beta^*$ by maximizing the conditional likelihood of $Y$ given $X$

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

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

Our goal below is to compute some confidence bounds for $\beta^*$.

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

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

are given by

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

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

2. Prove that there exists $\hat{\beta}$ such that $\|\hat{\beta} - \beta^*\| \leq \|\hat{\beta} - \beta^*\|$ and

$$\hat{\beta} - \beta^* = -H_n(\hat{\beta})^{-1} \nabla \ell_n(\beta^*).$$

In the following we assume that the $x_i$ are uniformly bounded, $\hat{\beta} \rightarrow \beta^*$ a.s. and that there exists a continuous and non-singular $H(\beta)$ such that $n^{-1}H_n(\beta)$ converges to $H(\beta)$, uniformly in a ball around $\beta^*$.

3. (optional) We set $p_i(\beta) = e^{\langle x_i, \beta \rangle} / (1 + e^{\langle x_i, \beta \rangle})$. Check that

$$\mathbb{E} e^{-n^{-1/2}t, \nabla \ell_n(\beta^*)} = \prod_{i=1}^n \left( 1 - p_i(\beta^*) + p_i(\beta^*) e^{\langle t, x_i \rangle} / \sqrt{n} \right) e^{-p_i(\beta^*) \langle t, x_i \rangle} / \sqrt{n}$$

$$= \exp \left( \frac{1}{2} t^T (n^{-1}H_n(\beta^*)) t + O(n^{-1/2}) \right)$$

4. What is the asymptotic distribution of $-n^{-1/2} \nabla \ell_n(\beta^*)$? of $\sqrt{n}(\hat{\beta} - \beta^*)$?

5. Propose a confidence interval $I_{n, \alpha}$ such that $\beta^*_\alpha \in I_{n, \alpha}$ with asymptotic probability $1 - \alpha$.

6. Propose a confidence ellipsoid $E_{n, \alpha}$ such that the probability that $\beta^* \in E_{n, \alpha}$ is asymptotically $1 - \alpha$.

7. Propose two schemes to select the coordinates of $x$ which are useful for predicting the class of a new data point $x$. 