Machine learning
2022-2023

ISUP - Sorbonne Université

---

# Discriminant analysis

---

Linear discriminant analysis assumes that the random variables $(X, Y) \in \mathbb{R}^p \times \{0,1\}$ has the following distribution. For all $A \in \mathcal{B}(\mathbb{R}^p)$ and all $y \in \{0,1\}$,

$$\mathbb{P}\left(X \in A; Y = y\right) = \pi_y \int_A g_y(x)\mathrm{d}x\,,$$

where $\pi_0$ and $\pi_1$ are positive real numbers such that $\pi_0 + \pi_1 = 1$ and $g_0$ (resp. $g_1$) is the probability density of a Gaussian random variable with mean $\mu_0 \in \mathbb{R}^d$ (resp. $\mu_1$) and positive definite covariance matrix $\Sigma_0 \in \mathbb{R}^{d\times d}$ (resp. $\Sigma_1$). The Bayes classifier $h_* : \mathbb{R}^p \to \{0,1\}$ is defined by

$$h_* : x \mapsto \mathbb{1}_{\{\pi_1 g_1(x) > \pi_0 g_0(x)\}}\,.$$

1. Give the distribution of the random variable $X$ and prove that

$$\mathbb{P}(h_*(X) \neq Y) = \min_{h:\mathbb{R}^p\to\{0,1\}} \left\{\mathbb{P}(h(X) \neq Y)\right\}\,.$$

2. Assume that $\mu_0 \neq \mu_1$. Prove that when $\Sigma_0 = \Sigma_1 = \Sigma$, for all $x \in \mathbb{R}^p$,

$$h_*(x) = 1 \Leftrightarrow (\mu_1 - \mu_0)^T\Sigma^{-1}\left(x - \frac{\mu_1 + \mu_0}{2}\right) > \log(\pi_0/\pi_1)\,.$$

   Provide a geometrical interpretation.

3. Prove that when $\pi_1 = \pi_0$,

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

   where $\Phi$ is the cumulative distribution function of a standard Gaussian random variable and

$$d(\mu_1, \mu_0)^2 = (\mu_1 - \mu_0)^T\Sigma^{-1}(\mu_1 - \mu_0)\,.$$

4. Assume now that $\Sigma_1 \neq \Sigma_0$. What is the nature of the frontier between $\{x\,;\, h_*(x) = 1\}$ and $\{x\,;\, h_*(x) = 0\}$?