Discriminant analysis

1 Classification error

Linear discriminant analysis assumes that the random variables $(X,Y) \in \mathbb{R}^d \times \{0,1\}$ have the following distribution. For all $A \in \mathcal{B}(\mathbb{R}^d)$ 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 π_0 and π_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. μ_1) and positive definite covariance matrix $\Sigma_0 \in \mathbb{R}^{d\times d}$ (resp. Σ_1). Define the classifier $h_* : \mathbb{R}^d \to \{0,1\}$ 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

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

For all $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}\mathbb{P}(X \in A) &= \mathbb{P}(Y=0)\mathbb{P}(X \in A|Y=0) + \mathbb{P}(Y=1)\mathbb{P}(X\in A|Y=1)\,,\\ &= \pi_0 \int_A g_0(x)\mathrm{d}x + \pi_1 \int_A g_1(x)\mathrm{d}x\,.\end{aligned}$$

The probability density of the random variable X is given, for all $x \in \mathbb{R}^d$, by

$$g(x) = \pi_0 g_0(x) + \pi_1 g_1(x)\,.$$

Then, note that

$$\eta(X) = \mathbb{P}(Y=1|X) = \frac{\mathbb{P}(X|Y=1)\,\mathbb{P}(Y=1)}{g(X)} = \frac{\pi_1 g_1(X)}{\pi_0 g_0(X) + \pi_1 g_1(X)}\,,$$

and the condition $\eta(x) \leqslant 1/2$ can be rewritten as

$$\frac{\pi_1 g_1(x)}{\pi_0 g_0(x) + \pi_1 g_1(x)} \leqslant 1/2\,,$$

that is $\pi_1 g_1(x) \leqslant \pi_0 g_0(x)$.

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

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

Provide a geometrical interpretation.

For all $x \in \mathbb{R}^d$,

$$
\begin{aligned}
\pi_1 g_1(x) > \pi_0 g_0(x)& \\
&\Leftrightarrow \log(\pi_1 g_1(x)) > \log(\pi_0 g_0(x))\,, \\
&\Leftrightarrow -\frac{1}{2}(x-\mu_1)^\top\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^\top\Sigma^{-1}(x-\mu_0) > \log(\pi_0/\pi_1)\,, \\
&\Leftrightarrow -\frac{1}{2}\Big(-\mu_1^\top\Sigma^{-1}x + \mu_1^\top\Sigma^{-1}\mu_1 - x^\top\Sigma^{-1}\mu_1 + \mu_0^\top\Sigma^{-1}x - \mu_0^\top\Sigma^{-1}\mu_0 + x^\top\Sigma^{-1}\mu_0\Big) > \log(\pi_0/\pi_1)\,, \\
&\Leftrightarrow x^\top\Sigma^{-1}\mu_1 - x^\top\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^\top\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^\top\Sigma^{-1}\mu_0 > \log(\pi_0/\pi_1)\,, \\
&\Leftrightarrow (\mu_1-\mu_0)^\top\Sigma^{-1}\Big(x - \frac{\mu_1+\mu_0}{2}\Big) > \log(\pi_0/\pi_1)\,.
\end{aligned}
$$

Therefore, all $x \in \mathbb{R}^d$ is classified according to its position with respect to an affine hyperplane orthogonal to $\Sigma^{-1}(\mu_1 - \mu_0)$.

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

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

where Φ is the cumulative distribution function of a standard Gaussian random variable and

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

Let Z_0 be a Gaussian random variable with mean μ_0 and variance Σ. Note that

$$\mathbb{P}(h_*(X) = 1|Y = 0) = \mathbb{P}\Big(\underbrace{(\mu_1-\mu_0)^\top\Sigma^{-1}(Z_0 - \frac{\mu_1+\mu_0}{2})}_{Z} > 0\Big)\,,$$

where, using $\delta = d(\mu_1,\mu_0)$,

$$\mathbb{E}[Z] = (\mu_1-\mu_0)^\top\Sigma^{-1}(\frac{\mu_0-\mu_1}{2}) = -\frac{\delta^2}{2}$$

and

$$\mathbb{V}[Z] = \mathbb{V}\Big[(\mu_1-\mu_0)^\top\Sigma^{-1}X\Big] = \Big((\mu_1-\mu_0)^\top\Sigma^{-1}\Big)\Sigma\left(\Sigma^{-1}(\mu_1-\mu_0)\right) = \delta^2\,.$$

Hence,

$$\mathbb{P}(h_*(X) = 1|Y = 0) = \mathbb{P}\Big(-\frac{\delta^2}{2} + \delta\varepsilon > 0\Big) = \mathbb{P}\Big(\varepsilon > \frac{\delta}{2}\Big) = \Phi\Big(-\frac{\delta}{2}\Big).$$

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

In this case, for all $x \in \mathbb{R}^d$,

$$
\begin{aligned}
\pi_1 g_1(x) > \pi_0 g_0(x)& \\
&\Leftrightarrow \log(\pi_1 g_1(x)) > \log(\pi_0 g_0(x))\,, \\
&\Leftrightarrow -\frac{1}{2}(x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^\top\Sigma_0^{-1}(x-\mu_0) > \log(\pi_0/\pi_1)\,, \\
&\Leftrightarrow \frac{1}{2}x^\top\Sigma_0^{-1}x - \frac{1}{2}x^\top\Sigma_1^{-1}x + x^\top\Sigma_1^{-1}\mu_1 - x^\top\Sigma_0^{-1}\mu_0 - \frac{1}{2}\mu_1^\top\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^\top\Sigma^{-1}\mu_0 > \log(\pi_0/\pi_1)\,.
\end{aligned}
$$

As the quadratic term does not vanish anymore, the frontier between $\{x\,;\,h_*(x) = 1\}$ and $\{x\,;\,h_*(x) = 0\}$ is a quadric.

2 Maximum likelihood estimation

We assume that the joint distribution of (X, Y) belongs to a family of distributions parametrized by a vector θ with real components. For $k \in \{-1, 1\}$, write $\pi_k = \mathbb{P}(Y = k)$. Assume that conditionally on the event $\{Y = k\}$, X has a Gaussian distribution with mean $\mu_k \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, whose density is denoted g_k. In this case, the parameter $\theta = (\pi_1, \mu_1, \mu_{-1}, \Sigma)$ belongs to the set $\Theta = [0, 1] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$. The parameter π_{-1} is not part of the components of θ since $\pi_{-1} = 1 - \pi_1$. In this case, the parameter $\theta = (\pi_1, \mu_1, \mu_{-1}, \Sigma)$ belongs to the set $\Theta = [0, 1] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$. The parameter π_{-1} is not part of the components of θ since $\pi_{-1} = 1 - \pi_1$.

When Σ and μ_1 and μ_{-1} are unknown, the discriminant analysis classifier cannot be computed explicitely. Assume that $(X_i, Y_i)_{1\leqslant i\leqslant n}$ are independent observations with the same distribution as (X, Y).

1. Write the joint loglikelihood of the observations.

   The loglikelihood of these observations is given by

$$
\begin{aligned}
&\log \mathbb{P}_\theta\left(X_{1:n}, Y_{1:n}\right) \\
&\quad= \sum_{i=1}^{n} \log \mathbb{P}_\theta\left(X_i, Y_i\right) , \\
&\quad= -\frac{nd}{2}\log(2\pi) + \sum_{i=1}^{n}\sum_{k\in\{-1,1\}} \mathbb{1}_{Y_i=k}\left(\log \pi_k - \frac{\log\det\Sigma}{2} - \frac{1}{2}\left(X_i-\mu_k\right)^\top \Sigma^{-1}\left(X_i-\mu_k\right)\right) , \\
&\quad= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\det\Sigma + \left(\sum_{i=1}^{n}\mathbb{1}_{Y_i=1}\right)\log\pi_1 + \left(\sum_{i=1}^{n}\mathbb{1}_{Y_i=-1}\right)\log(1-\pi_1) \\
&\qquad\quad - \frac{1}{2}\sum_{i=1}^{n}\mathbb{1}_{Y_i=1}\left(X_i-\mu_1\right)^\top\Sigma^{-1}\left(X_i-\mu_1\right) - \frac{1}{2}\sum_{i=1}^{n}\mathbb{1}_{Y_i=-1}\left(X_i-\mu_{-1}\right)^\top\Sigma^{-1}\left(X_i-\mu_{-1}\right) .
\end{aligned}
$$

2. Let M_d be the space of real-valued $d\times d$ symmetric positive matrices. Show that the function $\Sigma \mapsto \log\det\Sigma$ is concave on M_d.

   Let $\Sigma, \Gamma \in M_d$ and $\lambda \in [0, 1]$. Since $\Sigma^{-1/2}\Gamma\Sigma^{-1/2} \in M_d$, it is diagonalisable in some orthonormal basis and write $\mu_1, \ldots, \mu_d$ the (possibly repeated) entries of the diagonal. Note in particular that $\det(\Sigma^{-1/2}\Gamma\Sigma^{-1/2}) = \prod_{i=1}^{d}\mu_i$. *Then,*

$$
\begin{aligned}
\log\det[(1-\lambda)\Sigma + \lambda\Gamma] &= \log\det[\Sigma^{1/2}\{(1-\lambda)I + \lambda\Sigma^{-1/2}\Gamma\Sigma^{-1/2}\}\Sigma^{1/2}] \\
&= \log\det\Sigma + \log\det[(1-\lambda)I + \lambda\Sigma^{-1/2}\Gamma\Sigma^{-1/2}] \\
&= \log\det\Sigma + \sum_{i=1}^{d}\log(1-\lambda+\lambda\mu_i) \\
&\geq \log\det\Sigma + \sum_{i=1}^{d}(1-\lambda)\underbrace{\log(1)}_{=0} + \lambda\log(\mu_i) := D
\end{aligned}
$$

   where the last inequality follows from the concavity of the log. *Now, rewrite the rhs D as:*

$$
\begin{aligned}
D &= (1-\lambda)\log\det\Sigma + \lambda[\log\det\Sigma^{1/2} + \log\det\Sigma^{-1/2}\Gamma\Sigma^{-1/2} + \log\det\Sigma^{1/2}] \\
&= (1-\lambda)\log\det\Sigma + \lambda\log\det\Gamma
\end{aligned}
$$

   which completes the proof.

3. Show that the derivative of the real valued function $\Sigma \mapsto \log\det(\Sigma)$ defined on $\mathbb{R}^{d\times d}$ is given by:

$$\partial_\Sigma\{\log\det(\Sigma)\} = \Sigma^{-1}\,,$$

where, for all real valued function f defined on $\mathbb{R}^{d\times d}$, $\partial_\Sigma f(\Sigma)$ denotes the $\mathbb{R}^{d\times d}$ matrix such that for all $1 \leqslant i,j \leqslant d$, $\{\partial_\Sigma f(\Sigma)\}_{i,j}$ is the partial derivative of f with respect to $\Sigma_{i,j}$.

Recall that for all $i \in \{1,\dots,d\}$ we have $\det(\Sigma) = \sum_{k=1}^d \Sigma_{i,k}\Delta_{i,k}$ where $\Delta_{i,j}$ is the (i,j)-cofactor associated with Σ. For any fixed i,j, the component $\Sigma_{i,j}$ does not appear anywhere in the decomposition $\sum_{k=1}^d \Sigma_{i,k}\Delta_{i,k}$, except for the term $k=j$. This implies

$$\frac{\partial \log\det(\Sigma)}{\partial \Sigma_{i,j}} = \frac{1}{\det\Sigma}\frac{\partial\det(\Sigma)}{\partial\Sigma_{i,j}} = \frac{\Delta_{i,j}}{\det\Sigma}\,.$$

Recalling the identity $\Sigma\ [\Delta_{j,i}]_{1\le i,j\le d} = (\det\Sigma)\ I_d$ so that $\Sigma^{-1} = [\Delta_{j,i}]^\top_{1\le i,j\le d}/\det\Sigma$, we finally get

$$\left(\frac{\partial\log\det(\Sigma)}{\partial\Sigma_{i,j}}\right)_{1\le i,j\le d} = (\Sigma^{-1})^\top = \Sigma^{-1}\,,$$

where the last equality follows from the fact that Σ is symmetric.

4. Provide the maximum likehood estimator of θ.

The gradient of $\log\mathbb{P}_\theta(X_{1:n},Y_{1:n})$ with respect to θ is therefore given by

$$\begin{aligned}
\frac{\partial\log\mathbb{P}_\theta(X_{1:n},Y_{1:n})}{\partial\pi_1} &= \left(\sum_{i=1}^n \mathbb{1}_{Y_i=1}\right)\frac{1}{\pi_1} - \left(\sum_{i=1}^n \mathbb{1}_{Y_i=-1}\right)\frac{1}{1-\pi_1}\,,\\
\frac{\partial\log\mathbb{P}_\theta(X_{1:n},Y_{1:n})}{\partial\mu_1} &= \sum_{i=1}^n \mathbb{1}_{Y_i=1}\left(2\Sigma^{-1}X_i - 2\Sigma^{-1}\mu_1\right)\,,\\
\frac{\partial\log\mathbb{P}_\theta(X_{1:n},Y_{1:n})}{\partial\mu_{-1}} &= \sum_{i=1}^n \mathbb{1}_{Y_i=-1}\left(2\Sigma^{-1}X_i - 2\Sigma^{-1}\mu_{-1}\right)\,,\\
\frac{\partial\log\mathbb{P}_\theta(X_{1:n},Y_{1:n})}{\partial\Sigma^{-1}} &= \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{Y_i=1}(X_i-\mu_1)(X_i-\mu_1)^\top - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{Y_i=-1}(X_i-\mu_{-1})(X_i-\mu_{-1})^\top\,.
\end{aligned}$$

The maximum likelihood estimator is defined as the only parameter $\widehat{\theta}^n$ such that all these equations are set to 0. For $k \in \{-1,1\}$, it is given by

$$\begin{aligned}
\widehat{\pi}_k^n &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i=k}\,,\\
\widehat{\mu}_k^n &= \frac{1}{\sum_{i=1}^n \mathbb{1}_{Y_i=k}}\sum_{i=1}^n \mathbb{1}_{Y_i=k}\,X_i\,,\\
\widehat{\Sigma}^n &= \frac{1}{n}\sum_{i=1}^n \left(X_i - \widehat{\mu}_{Y_i}^n\right)\left(X_i - \widehat{\mu}_{Y_i}^n\right)^\top\,.
\end{aligned}$$