Introduction to machine learning

1 Warm-up: Bayes classifier for scalar Gaussian mixtures

Let $(X_i, Y_i)_{1\leqslant i\leqslant n}$ be independent variables in $\mathbb{R}\times\{0,1\}$. Assume that $\mathbb{P}(Y_1 = 0) = 1/2$. Assume also that the distribution of X_1 given $\{Y_1 = 0\}$ (resp. $\{Y_1 = 1\}$) is Gaussian with mean μ_0 (resp. μ_1) and variance 1. The probability density function of X_1 is written g. Write

$$g_0 : x \mapsto (2\pi)^{-1/2}\exp(-(x-\mu_0)^2/2) \quad \text{and} \quad g_1 : x \mapsto (2\pi)^{-1/2}\exp(-(x-\mu_1)^2/2)\,.$$

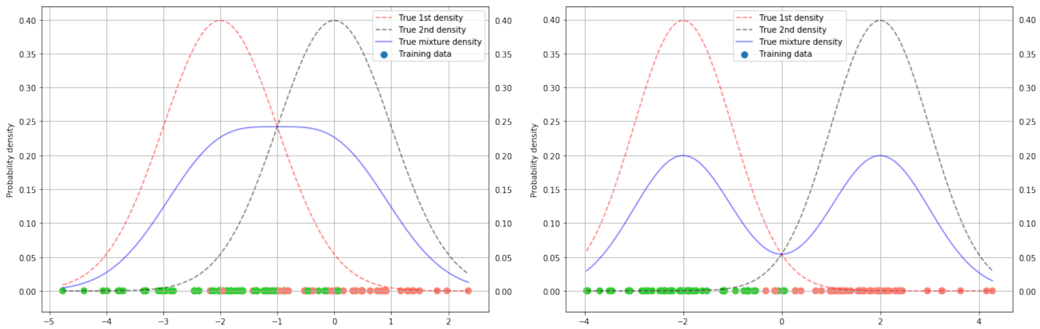

Figure 1: Samples and density when $\mu_0 = -2$ et $\mu_1 = 0$ (left) and $\mu_0 = -2$ and $\mu_1 = 2$ (right).

1. Provide an expression of a classifier h_* minimizing $h \mapsto \mathbb{P}(h(X) \neq Y)$.

   The classifier h_* such that $h_*(X) = 1$ if and only if $\mathbb{P}(Y = 1|X) > \mathbb{P}(Y = 0|X)$ minimizes the missclassification error:

   $$h_* \in \mathrm{Argmin}_{h:\mathbb{R}\to\{0,1\}}\left\{\mathbb{P}(h(X) \neq Y)\right\}\,.$$

2. Using Bayes rule, show that h_* depends only on g_1/g_0.

   By Bayes formula, $\mathbb{P}(Y = 1|X) = \mathbb{P}(Y = 1)g_1(X)/g(X)$, which yields

   $$\frac{\mathbb{P}(Y = 1|X)}{\mathbb{P}(Y = 0|X)} = \frac{g_1(X)}{g_0(X)}\,.$$

   Then, $h_*(X) = 1$ if and only if $g_1(X)/g_0(X) > 1$.

3. Show that the Bayes classifier uses the mean between μ_0 and μ_1 to classify samples.

$h_*(X) = 1$ if and only if $\log g_1(X) - \log g_0(X) > 0$, so that, assuming without loss of generality that $\mu_1 > \mu_0$:

$$\begin{aligned}
h_*(X) = 1 &\Leftrightarrow (X-\mu_0)^2 - (X-\mu_1)^2 > 0\,,\\
&\Leftrightarrow 2(\mu_1-\mu_0)X + \mu_0^2 - \mu_1^2 > 0\,,\\
&\Leftrightarrow X > \frac{\mu_1^2-\mu_0^2}{2(\mu_1-\mu_0)}\,,\\
&\Leftrightarrow X > \frac{\mu_1+\mu_0}{2}\,.
\end{aligned}$$

This criterion can lead to very poor performance if means are close (see Figure 1).

2 Bayes classifier

2.1 Uniform distributions

Assume that $(X, Y) \in \mathbb{R} \times \{0, 1\}$ is defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}(Y = 1) = \pi \in (0, 1)$. Assume that conditionally on $\{Y = 0\}$ (resp. $\{Y = 1\}$) X has a uniform distribution on $[0, \theta]$ with $\theta \in (0, 1)$ (resp. on $[0, 1]$). Compute $\eta(X) = \mathbb{P}(Y = 1|X)$.

Let g be the probability density function of X. For any measurable set A,

$$\begin{aligned}
\mathbb{P}(X \in A) &= \mathbb{P}(Y=0)\mathbb{P}(X \in A|Y=0) + \mathbb{P}(Y=1)\mathbb{P}(X \in A|Y=1)\,,\\
&= (1-\pi)\theta^{-1}\int \mathbb{1}_A(x)\mathbb{1}_{[0,\theta]}(x)\mathrm{d}x + \pi\int \mathbb{1}_A(x)\mathbb{1}_{[0,1]}(x)\mathrm{d}x\,,\\
&= \int \mathbb{1}_A(x)\left\{(1-\pi)\theta^{-1}\mathbb{1}_{[0,\theta]}(x) + \pi\mathbb{1}_{[0,1]}(x)\right\}\mathrm{d}x\,.
\end{aligned}$$

Therefore, $g : x \mapsto (1-\pi)\theta^{-1}\mathbb{1}_{[0,\theta]}(x) + \pi\mathbb{1}_{[0,1]}(x)$. Then, using Bayes rules and writing g_1 the probability density of the distribution of X given $\{Y = 1\}$,

$$\eta(X) = \mathbb{P}(Y=1|X) = \frac{\mathbb{P}(Y=1)g_1(X)}{g(X)} = \frac{\pi\mathbb{1}_{[0,1]}(X)}{(1-\pi)\theta^{-1}\mathbb{1}_{[0,\theta]}(X) + \pi\mathbb{1}_{[0,1]}(X)}\,.$$

2.2 Weighted risk

Assume that $(X, Y) \in \mathbb{R} \times \{0, 1\}$ is defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Using $\omega_0, \omega_1 > 0$, with $\omega_0 + \omega_1 = 1$, we consider the weighted risk:

$$\mathsf{R}(h) = \mathbb{E}[2\omega_Y\mathbb{1}_{Y\neq h(X)}]\,.$$

Compute a classifier h_* minimizing $h \mapsto \mathsf{R}(h)$ and $\mathsf{R}(h_*)$.

For all classifiers h, writing $\eta(X) = \mathbb{P}(Y = 1|X)$,

$$\begin{aligned}
\mathsf{R}(h) = \mathbb{E}[2\omega_Y\mathbb{1}_{Y\neq h(X)}] &= \mathbb{E}[2\omega_Y\mathbb{1}_{Y=1}\mathbb{1}_{h(X)=0} + 2\omega_Y\mathbb{1}_{Y=0}\mathbb{1}_{h(X)=1}]\,,\\
&= \mathbb{E}[2\omega_1\mathbb{1}_{Y=1}\mathbb{1}_{h(X)=0} + 2\omega_0\mathbb{1}_{Y=0}\mathbb{1}_{h(X)=1}]\,,\\
&= \mathbb{E}[2\omega_1\eta(X)\mathbb{1}_{h(X)=0} + 2\omega_0(1-\eta(X))\mathbb{1}_{h(X)=1}]\,,
\end{aligned}$$

Therefore, choosing $h_\star : x \mapsto \mathbb{1}_{\omega_1\eta(X)\geqslant\omega_0(1-\eta(X))}$ yields,

$$\mathsf{R}(h) \geqslant \mathsf{R}(h_*)\,.$$

Then, by definition, for all $x \in \mathbb{R}^d$,

$$h_\star(x) = 1 \Leftrightarrow \omega_1\eta(x) \geqslant \omega_0(1-\eta(x))$$

and

$$2\omega_1\eta(x)\mathbb{1}_{h_*(x)=0} + 2\omega_0(1-\eta(x))\mathbb{1}_{h_*(x)=1} = 2\left(\omega_1\eta(x)\right) \wedge \left(\omega_0(1-\eta(x))\right) .$$

This yields

$$\mathsf{R}(h_*) = 2\mathbb{E}[(\omega_1\eta(X)) \wedge (\omega_0(1-\eta(X)))] .$$

3 Additional exercises

3.1 Bayes classifier: excess risk

Let $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ be random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For any classifier $h : \mathcal{X} \to \{0, 1\}$, define its classification error by

$$\mathsf{R}(h) = \mathbb{P}(Y \neq h(X)) .$$

The classifier h_* defined by:

$$h_*(x) = \mathrm{sign}(\eta(x) - 1/2) ,$$

where

$$\eta(X) = \mathbb{P}(Y = 1|X) ,$$

minimizes $h \mapsto \mathsf{R}(h)$.

1. Prove that

$$\mathsf{R}(h_*) = \mathbb{E}\left[\eta(X) \wedge (1-\eta(X))\right] \leqslant \frac{1}{2} .$$

For all classifiers h, as h and Y take values in $\{0, 1\}$,

$$\mathsf{R}(h) = \mathbb{E}\left[\mathbb{1}_{h(X)\neq Y}\right] = \mathbb{E}\left[h(X)(1-Y) + (1-h(X))Y\right] .$$

As $\mathbb{E}[Y|X] = \eta(X)$ this yields,

$$\mathsf{R}(h) = \mathbb{E}\left[h(X)(1-\eta(X)) + (1-h(X))\eta(X)\right]$$

and

$$\mathsf{R}(h_*) = \mathbb{E}\left[h_*(X)(1-\eta(X)) + (1-h_*(X))\eta(X)\right] = \mathbb{E}\left[\eta(X) \wedge (1-\eta(X))\right] .$$

2. Prove that for all classifiers h, the excess risk is given by

$$\mathsf{R}(h) - \mathsf{R}(h_*) = \mathbb{E}\left[|1-2\eta(X)|\,|h(X) - h_*(X)|\right] .$$

By the previous question, for all classifiers h,

$$\begin{aligned}\mathsf{R}(h) - \mathsf{R}(h_*) &= \mathbb{E}\left[(h(X) - h_*(X))(1-\eta(X)) + (h_*(X) - h(X))\eta(X)\right] ,\\ &= \mathbb{E}\left[(h(X) - h_*(X))(1-2\eta(X))\right] .\end{aligned}$$

By definition of h_*, $h(X) - h_*(X)$ and $1 - 2\eta(X)$ have the same sign so that

$$\mathsf{R}(h) - \mathsf{R}(h_*) = \mathbb{E}\left[|1-2\eta(X)|\,|h(X) - h_*(X)|\right] .$$

3.2 Plug-in classifier

Let $(X,Y) \in \mathbb{R}^d \times \{-1,1\}$ be random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For any classifier $h : \mathcal{X} \to \{-1,1\}$, define its classification error by

$$\mathsf{R}(h) = \mathbb{P}(Y \neq h(X))\,.$$

The classifier h_* defined by:

$$h_*(x) = \mathrm{sign}(\eta(x) - 1/2)\,,$$

where

$$\eta(X) = \mathbb{P}(Y = 1|X)\,,$$

minimizes $h \mapsto \mathsf{R}(h)$. Given n independent couples $\{(X_i,Y_i)\}_{1\leqslant i\leqslant n}$ with the same distribution as (X,Y), an empirical surrogate for h_* is obtained from a possibly nonparametric estimator $\widehat{\eta}_n$ of η:

$$\widehat{h}_n : x \mapsto \mathrm{sign}(\widehat{\eta}_n(x) - 1/2)\,.$$

1. Prove that for any classifier $h : \mathcal{X} \to \{-1,1\}$,

$$\mathbb{P}(Y \neq h(X)|X) = (2\eta(X) - 1)\mathbb{1}_{h(X)=-1} + 1 - \eta(X)$$

and

$$\mathsf{R}(h) - \mathsf{R}(h_*) = 2\mathbb{E}\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}_{h(X)\neq h_*(X)}\right]\,.$$

For all classifiers h,

$$\begin{aligned}\mathbb{P}\left(Y \neq h(X)|X\right) &= \mathbb{P}\left(Y = -1, h(X) = 1|X\right) + \mathbb{P}\left(Y = 1, h(X) = -1|X\right)\,,\\ &= \mathbb{1}_{h(X)=1}\mathbb{P}\left(Y = -1|X\right) + \mathbb{1}_{h(X)=-1}\mathbb{P}\left(Y = 1|X\right)\,,\\ &= \mathbb{1}_{h(X)=-1}(2\eta(X) - 1) + 1 - \eta(X)\,.\end{aligned}$$

Then,

$$\mathsf{R}(h) - \mathsf{R}(h_*) = \mathbb{E}\left[\left(\mathbb{1}_{h(X)=-1} - \mathbb{1}_{h_*(X)=-1}\right)(2\eta(X) - 1)\right] = 2\mathbb{E}\left[\left|\eta(X) - \frac{1}{2}\right| \mathbb{1}_{h(X)\neq h_*(X)}\right]\,.$$

2. Prove that

$$|\eta(x) - 1/2|\mathbb{1}_{\widehat{h}_n(x)\neq h_*(x)} \leqslant |\eta(x) - \widehat{\eta}_n(x)|\mathbb{1}_{\widehat{h}_n(x)\neq h_*(x)}\,,$$

where

$$\widehat{h}_n : x \mapsto \mathrm{sign}(\widehat{\eta}_n(x) - 1/2)\,.$$

Deduce that

$$\mathsf{R}(\widehat{h}_n) - \mathsf{R}(h_*) \leqslant 2\mathbb{E}[|\eta(X) - \widehat{\eta}_n(X)|^2]^{1/2}\,.$$

Note that, for all $x \in \mathbb{R}^d$, $\widehat{h}_n(x) \neq h_*(x)$ if and only if i) $\eta(x) > 1/2$ and $\widehat{\eta}_n(x) \leqslant 1/2$ or ii) $\eta(x) \leqslant 1/2$ and $\widehat{\eta}_n(x) > 1/2$. If $\eta(x) > 1/2$ and $\widehat{\eta}_n(x) \leqslant 1/2$, then $|\eta(x) - \widehat{\eta}_n(x)| = \eta(x) - \widehat{\eta}_n(x) \geqslant \eta(x) - 1/2$. On the other hand, if $\eta(x) \leqslant 1/2$ and $\widehat{\eta}_n(x) > 1/2\}$, $|\eta(x) - \widehat{\eta}_n(x)| = \widehat{\eta}_n(x) - \eta(x) \geqslant 1/2 - \eta(x)$. Therefore, for all $x \in \mathbb{R}^d$,

$$|\eta(x) - 1/2|\mathbb{1}_{\widehat{h}_n(x)\neq h_*(x)} \leq |\eta(x) - \widehat{\eta}_n(x)|\mathbb{1}_{\widehat{h}_n(x)\neq h_*(x)}\,.$$

By the first question and Cauchy-Schwarz inequality,

$$\begin{aligned}\mathsf{R}(\widehat{h}_n) - \mathsf{R}(h_*) &= 2\mathbb{E}\left[|\eta(X) - 1/2|\,\mathbb{1}_{h_*(X)=\widehat{h}_n(X)}\right]\,,\\ &\leqslant 2\mathbb{E}\left[|\eta(X) - \widehat{\eta}_n(X)|\,\mathbb{1}_{\widehat{h}_n(X)\neq h_*(X)}\right]\,,\\ &\leqslant 2\mathbb{E}[|\eta(X) - \widehat{\eta}_n(X)|^2]^{1/2}\,.\end{aligned}$$