

Introduction to Statistical Learning

Exercise set #1

Exercise 1 - Consider the binary classification model where the random pair (X, Y) has distribution P over $\mathbb{R} \times \{0, 1\}$ and:

$$\begin{aligned}\mathcal{L}(X | Y = 0) &= \mathcal{U}([0, \theta]), \\ \mathcal{L}(X | Y = 1) &= \mathcal{U}([0, 1]), \\ p &= \mathbb{P}(Y = 1).\end{aligned}$$

with $p, \theta \in (0, 1)$ fixed.

Compute the posterior probability

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

for any $x \in \mathbb{R}$, as a function of p, θ . What if $\theta = 1/2$?

Correction 1 - Let $f_X^+(x) := f_X(x | Y = 1)$, $f_X^-(x) := f_X(x | Y = 0)$. We have

$$\begin{aligned}\eta(x) &= \mathbb{P}(Y = 1 | X = x) \\ &= \frac{f_X^+(x)\mathbb{P}(Y = 1)}{f_X(x)} \\ &= \frac{pf_X^+(x)}{pf_X^+(x) + (1-p)f_X^-(x)}\end{aligned}$$

If $\theta < x$, then $f_X^-(x) = 0$. If $x < 0$ or $x > 1$, $f_X^-(x) = f_X^+(x) = 0$. For any $0 < x < \theta$, we have $f_X^-(x) = \frac{1}{\theta}$. For any $0 < x < 1$, $f_X^+(x) = 1$.

Conclusion. Overall,

$$\eta(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{\theta p}{\theta p + (1-p)} & \text{if } 0 < x < \theta, \\ 1 & \text{if } \theta < x < 1, \\ 0 & \text{if } x > 1. \end{cases}$$

If $\theta = 1/2$, we have for all $0 < x < \theta$, $\eta(x) = \frac{p}{p+2(1-p)} = \frac{p}{2-p}$.

Exercise 2 - Consider the binary classification model where the random pair (X, Y) has distribution P over $\mathbb{R}_+ \times \{0, 1\}$ and :

- the marginal distribution of X over $\mathbb{R}_+$ is denoted P_X ,
- the conditional distribution of Y given $X = x$ is a Bernoulli distribution with parameter $\eta(x) = \frac{x}{x+\theta}$ for any $x \in \mathbb{R}_+$ and for fixed $\theta > 0$.

Find the Bayes classifier for this model (i.e., the minimizer of $L(g) = \mathbb{P}(Y \neq g(X))$) over all measurable classifiers $g : \mathbb{R}_+ \rightarrow \{0, 1\}$. Give the expression of the Bayes error $L^* = L(g^*)$ in the case where $P_X = \mathcal{U}([0, \alpha\theta])$ with $\alpha > 1$. What is the value of α that maximizes L^* ?

Correction 2 - Let $g : \mathbb{R}_+ \rightarrow \{0, 1\}$ a classifier. We have $\mathbb{P}(Y = 1 \mid X = x) = \eta(x) = \frac{x}{x+\theta}$, for any $x \in \mathbb{R}_+$. The error of a classifier is given by

$$\begin{aligned}\mathbb{P}(Y \neq g(X)) &= \mathbb{E}(\mathbb{1}_{g(X) \neq Y}(x)) \\ &= \mathbb{E}(\mathbb{1}_{g(X)=0}(x)\mathbb{P}(Y=1 \mid X=x) + \mathbb{1}_{g(X)=1}(x)\mathbb{P}(Y=0 \mid X=x)) \\ &= \mathbb{E}(\eta(x)\mathbb{1}_{g(X)=0}(x) + (1-\eta(x))\mathbb{1}_{g(X)=1}(x))\end{aligned}$$

The Bayes classifier g^* is given by

$$\begin{aligned}g^*(x) &= \begin{cases} 1 & \text{if } \eta(x) > 1/2 \iff x > \theta, \\ 0 & \text{if } \eta(x) < 1/2 \iff x < \theta, \end{cases} \\ &= \mathbb{1}_{x > \theta}(x)\end{aligned}$$

Conclusion.

$$g^*(x) = \mathbb{1}_{x > \theta}(x)$$

Moreover, the Bayes error is given by

$$\begin{aligned}L(g^*) &= \mathbb{P}(Y \neq g(X)) \\ &= \mathbb{E}(\eta(x)\mathbb{1}_{g(X)=0}(x) + (1-\eta(x))\mathbb{1}_{g(X)=1}(x)) \\ &= \int_0^\theta \eta(x)dP_X(x) + \int_\theta^1 (1-\eta(x))dP_X(x)\end{aligned}$$

In the case where $P_X = \mathcal{U}([0, \alpha\theta])$, this yields

$$\begin{aligned}L(g^*) &= \int_0^\theta \eta(x)dP_X(x) + \int_\theta^1 (1-\eta(x))dP_X(x) \\ &= \int_0^\theta \frac{x}{(x+\theta)\alpha\theta}dx + \int_\theta^{\alpha\theta} \frac{\theta}{(x+\theta)}dx \\ &= \frac{1}{\alpha\theta} \int_0^\theta \left(1 - \frac{\theta}{(x+\theta)}\right)dx + \int_\theta^{\alpha\theta} \frac{\theta}{(x+\theta)}dx \\ &= \frac{1}{\alpha\theta} (\theta - (\theta \ln(2\theta) - \theta \ln(\theta))) + (\theta \ln(\alpha\theta + \theta) - \theta \ln(2\theta)) \\ &= \frac{\theta \ln(\theta) + \ln((\alpha+1)\theta) - 2\theta \ln(2\theta) + \theta}{\theta\alpha} \\ &= \frac{2 \ln(\theta) + \ln(1+\alpha) - 2 \ln(2\theta) + 1}{\alpha} \\ &= \frac{1 + 2 \ln(\frac{\theta}{2\theta}) + \ln(1+\alpha)}{\alpha} \\ &= \frac{1 - 2 \ln(2) + \ln(1+\alpha)}{\alpha}\end{aligned}$$

Conclusion.

$$L(g^*) = \frac{1 - 2 \ln(2) + \ln(1+\alpha)}{\alpha}$$

We can obtain the value of α maximizing $L(g^*)$ by computing the first order optimality condition:

$$\begin{aligned}\frac{dL^*}{d\alpha} &= 0 \\ \iff \frac{\alpha}{(1+\alpha)\alpha^2} - \frac{(\alpha+1)(1-2\ln(2) + \ln(1+\alpha))}{(\alpha+1)\alpha^2} &= 0 \\ \iff \alpha - (1+\alpha)(1-2\ln(2) + \ln(1+\alpha)) &= 0 \\ \iff \alpha &\approx 1.80\end{aligned}$$

Conclusion.

$$\max_{\alpha \in \mathbb{R}^+} L(g^*) \approx 1.80$$

Exercise 3 - Let $X = (T, U, V)^\top$ where T, U, V IID real-valued random variables with exponential distribution $\mathcal{E}(1)$. Define $Y = \mathbb{1}\{T + U + V < \theta\}$ with fixed $\theta > 0$.

1. Find the Bayes classifier $g^*(T, U)$ when V is not observed. Give the expression of the classification error of g^* (also called Bayes error). Compute it for $\theta = 9$.
2. Now assume that only T is observed, and address the same questions as above.
3. Propose a classifier for X when none of T, U, V are observed. What is its classification error ?

Correction 3 -

1. The Bayes classifier is defined by

$$g^*(t, u) = \begin{cases} 1 & \text{if } \eta(t, u) > \frac{1}{2}, \\ 0 & \text{if } \eta(t, u) < \frac{1}{2}. \end{cases}$$

where $\eta(t, u) := \mathbb{P}(Y = 1 \mid T = t, U = u)$. Moreover,

$$\begin{aligned} \eta(t, u) &= \mathbb{E}(\mathbb{1}_{V < \theta - T - U}) \\ &= \begin{cases} \int_0^{\theta - u - t} e^{-v} dv = 1 - e^{-\theta + u + t} & \text{if } t + u < \theta, \\ 0 & \text{if } t + u \geq \theta, \end{cases} \end{aligned}$$

Conclusion. The Bayes classifier is given by

$$g^*(T, U) = \mathbb{1}_{t+u \leq \theta - \ln 2}(t, u) = \begin{cases} 1 & \text{if } t + u \leq \theta - \ln 2, \\ 0 & \text{if } t + u > \theta - \ln 2, \end{cases}$$

Introduce $Z = T + U$. The density of Z is given by

$$\begin{aligned} f_Z(z) &= \int_{t=0}^z f_T(t) f_U(z-t) dt \\ &= \int_{t=0}^z e^{-t} e^{-(z-t)} dt \\ &= z e^{-z} \end{aligned}$$

Conclusion. With $\eta(z) = (1 - e^{-\theta+z}) \mathbb{1}_{z < \theta}(z)$, the Bayes error is then given by

$$\begin{aligned} L(g^*) &= \mathbb{E}(\mathbb{1}_{z \leq \theta - \ln 2}(z)(1 - \eta(z)) + \mathbb{1}_{z > \theta - \ln 2}(z)\eta(z)) \\ &= \int_{z=0}^{\theta - \ln 2} (1 - \eta(z)) df_Z(z) + \int_{z=\theta - \ln 2}^{\theta} \eta(z) df_Z(z) \\ &= \dots \\ &= e^{-\theta} \left(1 + \theta + \frac{\theta^2}{2} - 2 \ln 2 (1 + \theta) + \ln^2 2 \right) \end{aligned}$$

With $\theta = 9$, $L(g^*) \approx 0.0046$.

2. Now, $\eta(t) := \mathbb{P}(Y = 1 \mid T = t)$ (only T is observed). Let $Z = U + V$, and $f_Z(z) = ze^{-z}$. Hence,

$$\begin{aligned}\eta(t) &= \mathbb{1}_{t < \theta}(t) \int_{z=0}^{\theta-t} ze^{-z} dz \\ &= \mathbb{1}_{t < \theta}(t)(1 - (\theta - t + 1)e^{-\theta-t})\end{aligned}$$

Conclusion. The Bayes classifier is given by

$$g^*(T) = \mathbb{1}_{(\theta-t+1)e^{-\theta-t} < 1/2} = \begin{cases} 1 & \text{if } \theta - t - \ln(\theta - t + 1) > \ln 2, \\ 0 & \text{if } \theta - t - \ln(\theta - t + 1) < \ln 2, \end{cases}$$

The Bayes error is then given by

$$\begin{aligned}L(g^*) &= \mathbb{E}(\min(\eta(t), 1 - \eta(t))) \\ &= \int_{t=0}^{\theta} (\mathbb{1}_{\theta-t-\ln(\theta-t+1) > \ln 2}(t)(1 - \eta(t)) + \mathbb{1}_{\theta-t-\ln(\theta-t+1) < \ln 2}(t)(\eta(t)))e^{-t} dt \\ &\approx 0.0059 \quad \text{if } \theta = 9.\end{aligned}$$

Remark

Because we have access to less information, the Bayes error is higher than in question 1.

3. With $S = T + U + V$, we have $f_S(s) = \frac{s^2 e^{-s}}{2}$. Then $\eta(x) = \mathbb{P}(S < \theta)$, so that

$$\begin{aligned}\eta(x) &= \int_0^{\theta} \frac{s^2 e^{-s}}{2} ds \\ &= 1 - \frac{e^{-\theta}}{2}(\theta^2 + 2\theta + 2)\end{aligned}$$

Conclusion. The Bayes classifier is given by

$$g^* = \mathbb{1}_{e^{-\theta}(\theta^2 + 2\theta + 2) < 1}.$$

Its Bayes error is

$$\begin{aligned}L(g^*) &= \int_{s=0}^{\theta} \left[\mathbb{1}_{e^{-\theta}(\theta^2 + 2\theta + 2) < 1} \left(1 - \frac{e^{-\theta}}{2}(\theta^2 + 2\theta + 2)\right) + \mathbb{1}_{e^{-\theta}(\theta^2 + 2\theta + 2) > 1} \left(\frac{e^{-\theta}}{2}(\theta^2 + 2\theta + 2)\right) \right] ds \\ &\approx 0.0062 \quad \text{if } \theta = 9.\end{aligned}$$

Remark

Because we have access to even less information, the Bayes error is even higher than in question 2.

Exercise 4 - Find the expressions of f_+ , f_- and η in the following probabilistic models :

- Discriminant Analysis : find η

$$f_+ = \mathcal{N}_d(\mu_+, \Sigma_+), \quad f_- = \mathcal{N}_d(\mu_-, \Sigma_-)$$

- Logistic Regression : find f_+ , f_-

$$\log \left(\frac{\eta_{\theta}(x)}{1 - \eta_{\theta}(x)} \right) = h(x, \theta), \quad \text{typically } h(x, \theta) = \theta^{\top} x.$$

Correction 4 - Discriminant Analysis:

$$f_+(x) = \frac{1}{(2\pi)^{d/2} \|\Sigma_+\|^{1/2}} e^{-\frac{1}{2}(x-\mu_+)^T \Sigma_+^{-1} (x-\mu_+)}$$

$$f_-(x) = \frac{1}{(2\pi)^{d/2} \|\Sigma_-\|^{1/2}} e^{-\frac{1}{2}(x-\mu_-)^T \Sigma_-^{-1} (x-\mu_-)}$$

We need to compute $\eta(x) = \frac{pf_+(x)}{pf_+(x)+(1-p)f_-(x)}$. After heavy computations, one can get

$$\eta(x) = \frac{1}{1 + \frac{1-p}{p} \left(\frac{\|\Sigma_+\|}{\|\Sigma_-\|} \right)^{1/2} e^{-\frac{1}{2}((x-\mu_-)^T \Sigma_-^{-1} (x-\mu_-) - (x-\mu_+)^T \Sigma_+^{-1} (x-\mu_+))}}$$

Logistic Regression: We assume $\log\left(\frac{\eta_\theta(x)}{1-\eta_\theta(x)}\right) = \theta^\top x$. Hence,

$$\begin{aligned} \eta_\theta(x)(1 + e^{\theta^\top x}) &= \eta_\theta(x) + \frac{\eta_\theta(x)^2}{1 - \eta_\theta(x)} \\ &= \frac{\eta_\theta(x) - \eta_\theta(x)^2 + \eta_\theta(x)^2}{1 - \eta_\theta(x)} = e^{\theta^\top x} \\ \implies \eta_\theta(x) &= \frac{1}{1 + e^{-\theta^\top x}} \end{aligned}$$

But by definition, $\eta_\theta(x) = \frac{pf_+(x)}{f_X(x)}$, so that $f_+(x) = \frac{f_X(x)}{p(1+e^{-\theta^\top x})}$.

Exercise 5 - Find the optimal elements in the following cases of error measures with binary classification data :

1. Asymmetric cost - set $\omega \in (0, 1)$,

$$L_\omega(g) = 2\mathbb{E}((1 - \omega)\mathbb{1}_{Y=1}\mathbb{1}_{g(X)=-1} + \omega\mathbb{1}_{Y=-1}\mathbb{1}_{g(X)=1})$$

2. Classification with mass constraint - set $u \in (0, 1)$

$$\min_g \mathbb{P}(Y \neq g(X)) \quad \text{subject to} \quad \mathbb{P}(g(X) = 1) = u$$

3. Classification with reject option - set $\gamma \in (0, 1/2)$

$$L_d^R(g) = \mathbb{P}(Y \neq g(X), g(X) \neq R) + \gamma\mathbb{P}(g(X) = R)$$

Correction 5 -

1. Asymmetric cost. Using the law of total expectation, one gets

$$\begin{aligned} L_\omega(g) &= 2\mathbb{E}_{X,Y}((1 - \omega)\mathbb{1}_{Y=1}\mathbb{1}_{g(X)=-1} + \omega\mathbb{1}_{Y=-1}\mathbb{1}_{g(X)=1}) \\ &= 2\mathbb{E}_{X,Y}((1 - \omega)\mathbb{E}_Y(\mathbb{1}_{Y=1} \mid X)\mathbb{1}_{g(X)=-1} + \omega\mathbb{E}_Y(\mathbb{1}_{Y=-1} \mid X)\mathbb{1}_{g(X)=1}) \\ &= 2\mathbb{E}((1 - \omega)\eta(X)\mathbb{1}_{g(X)=-1} + \omega(1 - \eta(X))\mathbb{1}_{g(X)=1}) \end{aligned}$$

so that $g^*(x) = 1 \iff \omega(1 - \eta(x)) < (1 - \omega)\eta(x) \iff \omega < \eta(x)$.

Conclusion.

$$g^*(x) = 2\mathbb{1}_{\omega < \eta(x)}(x) - 1.$$

The Bayes loss is then given by

$$L_\omega(g^*) = 2\mathbb{E}(\min((1 - \omega)\eta(X), \omega(1 - \eta(X)))).$$

2. Classification with mass constraint. We aim to find the solution of

$$\begin{cases} \min_g L_u(g) = \mathbb{E}(\eta(X)\mathbb{1}_{g(X)=-1} + (1 - \eta(X))\mathbb{1}_{g(X)=1}) \\ \text{s.t. } \mathbb{P}(g(X) = 1) = u \end{cases}$$

The constraint $\mathbb{P}(g(X) = 1) = u$ is equivalent to $\mathbb{P}(\eta(X) > q) = u$ for some $q \in (0, 1)$ (i.e., there necessarily exists q that verifies $\mathbb{P}(\eta(X) > q) = u$). Using that $\eta(X) = \eta(X) - q + q$, and $1 - \eta(X) = 1 - q + q - \eta(X)$, we have

$$\begin{aligned} L_u(g) &= \mathbb{E}((\eta(X) - q)\mathbb{1}_{g(X)=-1} + q\mathbb{1}_{g(X)=-1} + (1 - q)\mathbb{1}_{g(X)=1} + (q - \eta(X))\mathbb{1}_{g(X)=1}) \\ &= q(1 - u) + (1 - q)u + \mathbb{E}((\eta(X) - q)\mathbb{1}_{g(X)=-1} + (q - \eta(X))\mathbb{1}_{g(X)=1}) \end{aligned}$$

g^* then verifies $g^*(x) = 1 \iff q - \eta(x) < \eta(x) - q \iff q < \eta(x)$.

Conclusion. The optimal classifier is

$$g^*(x) = 2\mathbb{1}_{q < \eta(x)} - 1,$$

and its loss is given by

$$L_u(g^*) = \mathbb{E}(\min(\eta(X) - q, q - \eta(X)) + q(1 - u) + (1 - q)u).$$

3. Classification with reject option. By definition,

$$L_d^R(g) = \mathbb{P}(Y \neq g(X), g(X) \neq R) + \gamma \mathbb{P}(g(X) = R).$$

Hence,

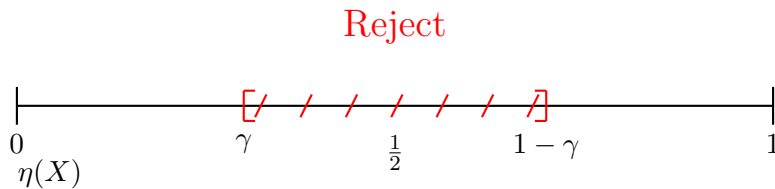
$$L_d^R(g) = \mathbb{E}(\mathbb{1}_{g(X)=-1}\eta(X) + \mathbb{1}_{g(X)=1}(1 - \eta(X)) + \gamma \mathbb{1}_{g(X)=R})$$

There are three values possible to take for the classifier :

$$\begin{cases} g^*(X) = -1 \iff \eta(X) < 1 - \eta(X) \quad \text{and} \quad \eta(X) < \gamma, \\ g^*(X) = 1 \iff 1 - \eta(X) < \eta(X) \quad \text{and} \quad 1 - \eta(X) < \gamma, \\ g^*(X) = R \iff \gamma < \eta(X) \quad \text{and} \quad \gamma < 1 - \eta(X). \end{cases}$$

Conclusion. This is equivalent to

$$\begin{cases} g^*(X) = -1 \iff \eta(X) < \gamma, \\ g^*(X) = 1 \iff 1 - \eta(X) < \gamma, \\ g^*(X) = R \iff \gamma < \eta(X) < 1 - \gamma. \end{cases}$$



Exercise 6 - Consider (X, Y) a random pair that models classification data with labels in $\{0, 1\}$. Define the following classification error

$$L_\omega(g) = \mathbb{E}(2\omega(Y) \cdot \mathbb{1}_{Y \neq g(X)})$$

where $\omega(0) + \omega(1) = 1$. Consider the unit square in $\mathbb{R}^2$.

1. Plot the curves defined by $g \mapsto (\mathbb{P}\{g(X) = 1 \mid Y = 0\}, \mathbb{P}\{g(X) = 1 \mid Y = 1\})$ when g varies such that $L_\omega(g) = C$ with C fixed, for different values of C .
2. Same question but assuming now that $\mathbb{P}\{g(X) = 1\} = C$ with C fixed.

Correction 6 -

1. Assume $L_\omega(g) = C$. If $\omega(0) = 0$, this means that $2\mathbb{P}(g(X) = 0 \mid Y = 1) = C \iff \mathbb{P}(g(X) = 1 \mid Y = 1) = 1 - C/2$. Else,

$$C = L_\omega(g)$$

$$\iff C = 2\omega(0)\mathbb{P}(g(X) = 1 \mid Y = 0)\mathbb{P}(Y = 0) + 2\omega(1)\mathbb{P}(g(X) = 0 \mid Y = 1)\mathbb{P}(Y = 1)$$

$$\iff C = 2\omega(0)\mathbb{P}(g(X) = 1 \mid Y = 0)\mathbb{P}(Y = 0) + 2(1 - \omega(0))(1 - \mathbb{P}(g(X) = 1 \mid Y = 1))(1 - \mathbb{P}(Y = 0))$$

$$\iff C = 2(1 - \omega(0))(1 - \mathbb{P}(Y = 0)) + 2\omega(0)\mathbb{P}(g(X) = 1 \mid Y = 0)\mathbb{P}(Y = 0) - 2\mathbb{P}(g(X) = 1 \mid Y = 1)(1 - \omega(0))$$

$$\iff \mathbb{P}(g(X) = 1 \mid Y = 0) = \frac{C - 2(1 - \omega(0))(1 - \mathbb{P}(Y = 0))}{2\omega(0)\mathbb{P}(Y = 0)} + \frac{\mathbb{P}(g(X) = 1 \mid Y = 1)(1 - \omega(0))(1 - \mathbb{P}(Y = 0))}{\omega(0)\mathbb{P}(Y = 0)}$$

Hence, a classifier g verifying $C = L_\omega(g)$ must verify the linear relationship

$$\mathbb{P}(g(X) = 1 \mid Y = 0) = \frac{(1 - \omega(0))(1 - \mathbb{P}(Y = 0))}{\omega(0)\mathbb{P}(Y = 0)}\mathbb{P}(g(X) = 1 \mid Y = 1) + \frac{C - 2(1 - \omega(0))(1 - \mathbb{P}(Y = 0))}{2\omega(0)\mathbb{P}(Y = 0)}$$

The curve is thus a linear function.

2. Now if $\mathbb{P}(g(X) = 1) = C$, this means that

$$C = \mathbb{P}(g(X) = 1)$$

$$\iff C = \mathbb{P}(g(X) = 1 \mid Y = 0)\mathbb{P}(Y = 0) + \mathbb{P}(g(X) = 1 \mid Y = 1)\mathbb{P}(Y = 1)$$

$$\iff \mathbb{P}(g(X) = 1 \mid Y = 0) = -\frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = 0)}\mathbb{P}(g(X) = 1 \mid Y = 1) + \frac{C}{\mathbb{P}(Y = 0)}$$

Again, the curve is a linear function.

Exercise 7 - We consider the model for classification data where X is a random vector on $\mathbb{R}^d$ and Y is a random variable taking values in $\{-1, +1\}$. We denote $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$ the posterior probability. We consider the following problems for which the question is to compute the optimal decision rule g^* or f^* - please also provide the main proof arguments.

1. Criterion to minimize : $R(g) = \mathbb{E}((Y - g(X))^2)$ where $g : \mathbb{R}^d \rightarrow \{-1, +1\}$
2. Criterion to minimize : $R(f) = \mathbb{E}((Y - f(X))^2)$ where $f : \mathbb{R}^d \rightarrow \mathbb{R}$
3. Criterion to minimize : $A(f) = \mathbb{E}(\exp(-Yg(X)))$ where $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{-\infty, +\infty\}$
4. Criterion to minimize : $A(f) = \mathbb{E}(\log(1 + \exp(-Yg(X))))$ where $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{-\infty, +\infty\}$
5. Criterion to minimize : $A_2(f) = \mathbb{E}(\max(0, 1 - Yf(X)))$ where $f : \mathbb{R}^d \rightarrow \mathbb{R}$

Explain why such criteria are relevant for the binary classification problem.

Correction 7 - Let $\eta(X) := \mathbb{P}(Y = 1 \mid X = x)$.

1. We aim to find the classifier minimizing the quadratic risk:

$$\begin{aligned} R(g) &= \mathbb{E}((Y - g(X))^2) \\ &= \mathbb{E}(2^2 \mathbb{1}_{Y \neq g(X)}) \\ &= 4\mathbb{E}(\eta(X) \mathbb{1}_{g(X)=-1} + (1 - \eta(X)) \mathbb{1}_{g(X)=1}) \end{aligned}$$

which is defined by

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2, \\ -1 & \text{if } \eta(x) < 1/2. \end{cases}$$

This is the Bayes optimal classifier.

2. We aim to find the regressor minimizing the quadratic risk:

$$\begin{aligned} R(f) &= \mathbb{E}((Y - f(X))^2) \\ &= \int_{-\infty}^{\infty} ((1 - f(x))^2 \eta(x) + (-1 - f(x))^2 (1 - \eta(x))) d\mathbb{P}_X(x) \\ &= \int_{-\infty}^{\infty} (1 + 2f(x) - 4f(x)\eta(x) + f^2(x)) d\mathbb{P}_X(x) \\ &= \int_{-\infty}^{\infty} ((f(x) + 1)^2 - 4f(x)\eta(x)) d\mathbb{P}_X(x) \end{aligned}$$

which is defined by

$$\begin{aligned} f^*(x) &= \arg \min_{y \in \mathbb{R}} (y + 1)^2 - 4y\eta(x) \\ &= \arg \min y^2 + (2 - 4\eta(x))y + 1 \\ &= -\frac{2 - 4\eta(x)}{2} \\ &= 2\eta(x) - 1 \end{aligned}$$

This optimal regressor is relevant for giving confidence measures of the output of the Bayes classifier.

3. We aim to find the regressor minimizing the exponential risk:

$$\begin{aligned} A(f) &= \mathbb{E}(\exp(-Y f(X))) \\ &= \int_{-\infty}^{\infty} (\exp(-f(x))\eta(x) + \exp(f(x))(1 - \eta(x))) d\mathbb{P}_X(x) \end{aligned}$$

which is defined by

$$f^*(x) = \arg \min_{y \in \mathbb{R}} \underbrace{\exp(-y)\eta(x) + (1 - \eta(x))\exp(y)}_{=: F(y)}$$

Differentiating yields $F'(y) = -\eta(x)\exp(-y) + (1 - \eta(x))\exp(y) = 0 \iff \exp(2y) = \frac{1 - \eta(x)}{\eta(x)}$. Hence,

$$f^*(x) = \frac{1}{2} \log \frac{\eta(x)}{1 - \eta(x)}.$$

Remark

This regressor is the minimizer of the exponential loss. This loss is used for instance in AdaBoost. It induces the same Bayes classifier as the square loss: $\text{sgn}(f^*(x)) = \text{sgn}(\eta(x) - 1/2)$.

4. We aim to find the regressor minimizing the logit risk:

$$\begin{aligned} A(f) &= \mathbb{E}(\log(1 + \exp(-Y f(X)))) \\ &= \int_{-\infty}^{\infty} (\log(1 + \exp(-f(x))\eta(x) + \log(1 + \exp(f(x))(1 - \eta(x))))d\mathbb{P}_X(x) \end{aligned}$$

which is defined by

$$f^*(x) = \arg \min_{y \in \mathbb{R}} \underbrace{\log(1 + \exp(-y)\eta(x) + \log(1 + \exp(y))(1 - \eta(x)))}_{=: F(y)}$$

Differentiating yields

$$\begin{aligned} 0 &= F'(y) \\ \iff 0 &= -\eta(x) \frac{\exp(-y)}{1 + \exp(-y)} + (1 - \eta(x)) \frac{\exp(y)}{1 + \exp(y)} = 0 \\ \iff 0 &= -\eta(x) \exp(-y)(1 + \exp(y)) + (1 - \eta(x)) \exp(y)(1 + \exp(-y)) \\ \iff 0 &= -\eta(x)(\exp(-y) + 1) + (1 - \eta(x))(\exp(y) + 1) \\ \iff 0 &= (1 - \eta(x)) \exp(y) + (1 - 2\eta(x)) - \eta(x) \exp(-y) \\ \iff 0 &= (1 - \eta(x)) \exp(y)^2 + (1 - 2\eta(x)) \exp(y) - \eta(x) \\ \iff e^y &= \frac{-(1 - 2\eta(x)) + 1}{2(1 - \eta(x))} = \frac{\eta(x)}{1 - \eta(x)} \end{aligned}$$

The optimal regressor is then given by

$$f^*(x) = \log \frac{\eta(x)}{1 - \eta(x)}.$$

Remark

The logit loss is well-calibrated: we have that $\text{sgn}(f^*(x)) = \text{sgn}(\eta(x) - 1/2)$. Moreover, we recover the link between the regressor and the posterior probability via the sigmoid function: $\sigma(f^*(x)) = \eta(x)$.

5. We aim to find the regressor minimizing the risk:

$$\begin{aligned} A_2(f) &= \mathbb{E}(\max(0, 1 - Y f(X))) \\ &= \int_{-\infty}^{\infty} (\max(0, 1 - f(x))\eta(x) + \max(0, 1 + f(x))(1 - \eta(x)))d\mathbb{P}_X(x) \end{aligned}$$

which is defined by

$$f^*(x) = \arg \min_{y \in \mathbb{R}} \underbrace{\max(0, 1 - y)\eta(x) + \max(0, 1 + y)(1 - \eta(x))}_{=: F(y)}$$

We have that

$$F(y) = \begin{cases} (1 + y)(1 - \eta(x)) & \text{if } y > 1, \\ (1 - y)\eta(x) & \text{if } y < -1, \\ (1 + y)(1 - \eta(x)) + (1 - y)\eta(x) & \text{if } -1 \leq y \leq 1. \end{cases}$$

But, for all $y > 1$, $F(y) = (1 + y)(1 - \eta(x)) > 2(1 - \eta(x)) = F(1)$. Similarly, for all $y < -1$, $F(y) > F(-1)$. Hence, we can differentiate F on $[-1, 1]$ to find the minimum and this yields

$$F'(y) = 1 - 2\eta(x) > 0 \iff \eta(x) < 1/2,$$

so that the minimum is reached at

$$f^*(x) = \begin{cases} -1 & \text{if } \eta(x) < 1/2, \\ [-1, 1] & \text{if } \eta(x) = 1/2, \\ 1 & \text{if } \eta(x) > 1/2. \end{cases}$$

Remark

Here, f^* only takes discrete values. The loss is used by SVMs. The fact that f^* is bounded on $[-1, 1]$ is directly linked with the essence of SVMs: the classifier only takes into account the points near the decision bound and does not decrease its loss on points far from it.

Remark

All the losses from questions 2 to 5 are convex. It ensures that the problems are well-posed: in practice, those are empirically solvable and aligned with the Bayes optimal classifier from question 1.

Exercise 8 - Consider IID random pairs (X, Y) and (X', Y') over $\mathbb{R}^d \times \mathcal{Y}$. Set the following posterior probabilities :

$$\begin{aligned} \forall x, x' \in \mathbb{R}^d, \quad \rho_+(x, x') &= \mathbb{P}(Y - Y' > 0 \mid X = x, X' = x'), \\ \rho_-(x, x') &= \mathbb{P}(Y - Y' < 0 \mid X = x, X' = x'), \end{aligned}$$

and for any preference rule $\pi : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \{-1, 0, 1\}$, consider the pairwise error measure

$$L(\pi) = \mathbb{P}((Y - Y') \cdot \pi(X, X') < 0).$$

1. Find the Bayes rule π^* and the Bayes error $L^* = L(\pi^*)$ for this problem, as well as the excess of risk $L(\pi) - L^*$ for any preference rule π (will involve ρ_+ and ρ_-).
2. Assume $\mathcal{Y} = \{-1, +1\}$ and denote by $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$. Provide the expressions for $\rho_+(x, x')$ and $\rho_-(x, x')$ and discuss how the behavior of η could lead to difficult situations for the learning process to be efficient.
3. Assume now that $\mathcal{Y} = \mathbb{R}$ and that $Y = m(X) + \sigma(X) \cdot N$ where m and σ are $\mathbb{P}_X$ measurable functions, N a random noise with normal distribution $\mathcal{N}(0, 1)$, while N and X are independent variables. Provide the expressions for $\rho_+(x, x')$ and $\rho_-(x, x')$ in this case and discuss the relation between properties of the model and the learning process.

Correction 8 -

1. π^* minimizes the pairwise error :

$$\begin{aligned} L(\pi) &= \mathbb{P}((Y - Y')\pi(X, X') < 0) \\ &= \mathbb{E}(\mathbb{1}_{(Y-Y')\pi(X, X') < 0}) \\ &= \mathbb{E}(\mathbb{1}_{Y-Y' < 0} \mathbb{1}_{\pi(X, X') > 0} + \mathbb{1}_{Y-Y' > 0} \mathbb{1}_{\pi(X, X') < 0}) \\ &= \mathbb{E}(\mathbb{E}(\mathbb{1}_{Y-Y' < 0} \mathbb{1}_{\pi(X, X') > 0} \mid X = x, X' = x') + \mathbb{E}(\mathbb{1}_{Y-Y' > 0} \mathbb{1}_{\pi(X, X') < 0} \mid X = x, X' = x')) \\ &= \mathbb{E}(\rho_-(X, X') \mathbb{1}_{\pi(X, X') = 1} + \rho_+(X, X') \mathbb{1}_{\pi(X, X') = -1}) \end{aligned}$$

Remark

There is probably a mistake in the text: $\pi : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \{-1, 1\}$ rather than $\{-1, 0, 1\}$: otherwise, $\pi^* \equiv 0$ is the Bayes rule.

Hence, the Bayes rule is given by

$$\pi^*(x, x') = \begin{cases} 1 & \text{if } \rho_-(x, x') < \rho_+(x, x'), \\ -1 & \text{if } \rho_-(x, x') > \rho_+(x, x'). \end{cases}$$

For any rule π , the excess risk is given by

$$\begin{aligned} L(\pi) - L^* &= \mathbb{E}((\rho_+(X, X') - \rho_-(X, X'))\mathbb{1}_{\pi(X, X')=1}\mathbb{1}_{\pi^*(X, X')=-1} \\ &\quad + (\rho_-(X, X') - \rho_+(X, X'))\mathbb{1}_{\pi(X, X')=-1}\mathbb{1}_{\pi^*(X, X')=1}) \\ &= \mathbb{E}((\max(\rho_-(X, X'), \rho_+(X, X')) - \min(\rho_-(X, X'), \rho_+(X, X')))\mathbb{1}_{\pi(X, X') \neq \pi^*(X, X')}) \end{aligned}$$

2. If $\mathcal{Y} = \{-1, +1\}$, we have that

$$\begin{aligned} \rho_+(x, x') &= \mathbb{P}(Y - Y' > 0 \mid X = x, X' = x') \\ &= \mathbb{P}(Y = 1, Y' = -1 \mid X = x, X' = x') \\ &= \mathbb{P}(Y = 1 \mid X = x)\mathbb{P}(Y' = -1 \mid X' = x') \\ &= \eta(x)(1 - \eta(x')) \end{aligned}$$

Similarly,

$$\begin{aligned} \rho_-(x, x') &= \mathbb{P}(Y - Y' < 0 \mid X = x, X' = x') \\ &= \mathbb{P}(Y = -1, Y' = 1 \mid X = x, X' = x') \\ &= \mathbb{P}(Y = -1 \mid X = x)\mathbb{P}(Y' = 1 \mid X' = x') \\ &= \eta(x')(1 - \eta(x)) \end{aligned}$$

If η is heavily imbalanced, the learning process can be unefficient because of ρ_+ and ρ_- will be close to 0.

3. In this case,

$$\begin{aligned} \rho_+(x, x') &= \mathbb{P}(Y - Y' > 0 \mid X = x, X' = x') \\ &= \mathbb{P}(s(x, x'), Z > m(x') - m(x)) \\ &= \mathbb{P}\left(Z > \frac{m(x') - m(x)}{s(x, x')}\right) \\ &= 1 - \Phi\left(\frac{m(x') - m(x)}{s(x, x')}\right) \\ &= \Phi\left(\frac{m(x) - m(x')}{s(x, x')}\right) \end{aligned}$$

where Φ is the distribution function of $\mathcal{N}(0, 1)$. Similarly,

$$\rho_-(x, x') = \Phi\left(\frac{m(x') - m(x)}{s(x, x')}\right)$$

Remark

The Bayes optimal rule only depends on $\text{sgn}(m(x) - m(x'))$. Moreover, the difficulty of the learning process will be driven by the ratio between the signal (means m, m') and the noise (standard deviations σ, σ'). If the gaussian distributions are intertwined (i.e., similar means or large stds), learning becomes challenging.