PC2. Foundations of Machine Learning. (MDC_51006_EP - 2025-2026)

Exercise 1 (Link Estimation/Prediction) Assume we observe a sample $((X_i, Y_i))_{i=1}^n$ where $Y_i \in \{-1, 1\}$ and $X_i \in \mathbb{R}^d$ are independent random variables following the same law than a generic pair (X, Y). We want to predict Y from X using a predictor $f : \mathbb{R}^d \to \{-1, 1\}$ and we measure the quality of our predictor using the $0/1$ loss:

$$\ell^{0/1}(y, y') = \mathbf{1}_{y \neq y'}.$$

The optimal Bayes classifier is given by

$$f^\star(X) = \operatorname{sign}(2p_1(X) - 1)$$

where $p_1(X) = \mathbb{P}(Y = 1|X)$.

Assume we have an estimate of the conditional law of $Y|X$ and denote $\widehat{p}_1(X) = \mathbb{P}(\widehat{Y = 1}|X)$. We define the plug-in classifier as

$$\widehat{f} = \operatorname{sign}(2\widehat{p}_1 - 1).$$

We want to prove that

$$\begin{aligned}\mathbb{E}\Big[\ell^{0,1}(Y, \widehat{f}(\underline{X}))\Big] - \mathbb{E}\big[\ell^{0,1}(Y, f^\star(\underline{X}))\big] &\leq \mathbb{E}\Big[\|\widehat{Y|\underline{X}} - Y|\underline{X}\|_1\Big] \\ &\leq \Big(\mathbb{E}\Big[2\,\mathrm{KL}(Y|\underline{X}, \widehat{Y|\underline{X}}\Big]\Big)^{1/2}\end{aligned}$$

1. Prove that for any predictor f,

$$\mathbb{E}\Big[\ell^{0/1}(Y, f(\underline{X}))\Big] = \mathbb{E}_{\underline{X}}\big[(1 - p_1(\underline{X})) + (2p_1(\underline{X}) - 1)\mathbf{1}_{f(\underline{X})=-1}\big]$$

2. Deduce that

$$\begin{aligned}\mathbb{E}\Big[\ell^{0/1}(Y, \hat{f}(\underline{X}))\Big] - -\mathbb{E}\big[\ell^{0,1}(Y, f^\star(\underline{X}))\big] \\ \leq 2\mathbb{E}_{\underline{X}}[|p_1(\underline{X}) - \widehat{p}_1(\underline{X})|]) = \mathbb{E}_{\underline{X}}[\|p(\underline{X}) - \widehat{p}(\underline{X})\|_1]\end{aligned}$$

3. Finish the proof using $\|P - Q\|_1 \leq \sqrt{2\,\mathrm{KL}(P, Q)}$

Exercise 2 (Linear Regression and Design) Assume we observe some points $((X_i, Y_i))_{i=1}^n$ with $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ are assumed to be independent copies of a generic (X, Y), regressing those points means estimating $\mathbb{E}[Y|X]$ at least on the X_i. Indeed, there are two classical problems that can be studied

- **Fixed Design** We assume that the X_i are fixed and want to estimate $f(X) = \mathbb{E}[Y|X]$ by a predictor $\hat{f}$ only at the observed points X_i. We measure the quality of $\hat{f}$ as the average error on a replication set $((X_i, Y_i'))_{i=1}^n$.

$$\mathsf{Err}_F = \mathbb{E}_{(Y_i)_{i=1}^n}\left[\mathbb{E}_{(Y_i')_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n \ell(Y_i', \hat{f}(X_i))\right]\right]$$

  where the X_i are the same and $Y_i'|X_i$ has the same law as $Y_i|X_i$.

- **Random Design** We assume that the X_i are i.i.d. and we want to estimate $f(X) = \mathbb{E}[Y|X]$ by a predictor $\hat{f}$ everywhere. We measure the quality as the average error on new observation (X', Y') having the same law as (X_i, Y_i):

$$\mathsf{Err}_R = \mathbb{E}_{((X_i,Y_i))_{i=1}^n}\Big[\mathbb{E}_{(X',Y')}\Big[\ell(Y', \hat{f}(X'))\Big]\Big].$$

Linear least square regression is classically studied in the fixed design setting, while the random design setting corresponds to the statistical learning setting. In this exercise, we are going to look at the performance of the linear least square on those two settings.

Following Rosset and Tibshirani, we introduce an intermediate setting:

- **Repeated Fixed Design** We assume that the X_i are random, possibly of different law and want to estimate $f(X) = \mathbb{E}[Y|X]$ by a predictor $\hat{f}$ only at the observed points X_i. We measure the quality by the Fixed Design error averaged over all possible designs:

$$\mathsf{Err}_{RF} = \mathbb{E}_{((X_i,Y_i))_{i=1}^n}\left[\mathbb{E}_{(Y_i')_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n \ell(Y_i', \hat{f}(X_i))\right]\right].$$

as well as the average empirical estimate of the quality:

$$\mathsf{Err}_{Emp} = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \hat{f}(X_i)).$$

We assume that all the samples are independent conditionally to X, that the X_i are fixed in the Fixed design setting, independent in the Repeated Fixed design setting and i.i.d. in the Random design one. Except, for the warm-up, we will assume a homoscedastic setting in which $Y = f(X) + \epsilon$ with ϵ a centered random noise of variance σ^2 and that the loss is the square one.

1. Warm-up
   (a) Verify that $\mathbb{E}_{(X_i)_{i=1}^n}[Err_F] = Err_{RF}$ and, if the (X_i, Y_i) are i.i.d.,

$$\mathsf{Err}_R = \mathbb{E}_{((X_i,Y_i))_{i=1}^n}\left[\mathbb{E}_{((X_i',Y_i'))_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n \ell(Y_i', \hat{f}(X_i'))\right]\right]$$

   where (X_i', Y_i') are independent copies of (X_i, Y_i).
   (b) Verify that if $\hat{f} = \mathrm{argmin}_g \sum_{i=1}^n \ell(Y_i, f(X_i))$ then $\mathbb{E}_{(Y_i)_{i=1}^n}[\mathsf{Err}_{Emp}] \leq \mathsf{Err}_F$ and $\mathbb{E}_{((X_i,Y_i))_{i=1}^n}[\mathsf{Err}_{emp}] \leq \mathsf{Err}_{RF}$.
2. We will now assume that ℓ is the ℓ^2 loss: $\ell(Y, f(X)) = |Y - f(X)|^2$.
   (a) Prove that in the fixed design setting:

$$extErr_F = \sigma^2 + \frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}_{(Y_i)_{i=1}^n}\left[\hat{f}(X_i)\right])^2 + \frac{1}{n}\sum_{i=1}^n \mathbb{V}\mathrm{ar}_{(Y_i)_{i=1}^n}\left[\hat{f}(X_i)\right]$$

   (b) Show that if we assume that the X_i are i.i.d. then for the repeated fixed design setting:

$$\mathsf{Err}_{RF} = \sigma^2 + \mathbb{E}_{(X_i)i=1^n}\left[(f(X_1) - \mathbb{E}_{(Y_i)_{i=1}^n}\left[\hat{f}(X_1)\middle|(X_i)_{i=1}^n\right])^2\right] + \mathbb{E}_{(X_i)_{i=1}^n}\left[\mathbb{V}\mathrm{ar}_{(Y_i)_{i=1}^n}\left[\hat{f}(X_1)\middle|(X_i)_{i=1}^n\right]\right]$$

   (c) Finally prove that for the random design setting

$$\mathsf{Err}_R = \sigma^2 + \mathbb{E}_{(X_i)_{i=1}^n, X'}\left[(f(X') - \mathbb{E}_{(Y_i)}\left[\hat{f}(X')\middle|(X_i)_{i=1}^n\right])^2\right] + \mathbb{E}_{(X_i)_{i=1}^n, X'}\left[\mathbb{V}\mathrm{ar}_{(Y_i)_{i=1}^n}\left[\hat{f}(X')\middle|(X_i)_{i=1}^n\right]\right]$$

3. We define the optimism Opt_F (respectively Opt_{RF}) as the difference between Err_F (respectively Err_{RF}) and the corresponding expectation of Err_{emp}.
   (a) Justify the name *optimism* when the estimator is the minimizer of the empirical risk.
   (b) Verify that in the fixed design setting:

$$\mathsf{Opt}_F = \mathsf{Err}_F - \mathbb{E}_{(Y_i)_{i=1}^n}[\mathsf{Err}_{emp}] = \mathbb{E}_{(Y_i)_{i=1}^n,(Y_i')_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n (Y_i' - \hat{f}(X_i))^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{f}(X_i))^2\right]$$

(c) Prove that

$$\mathsf{Opt}_F = \frac{2}{n}\sum_{i=1}^{n} \mathbb{C}\mathrm{ov}\left[Y_i, \hat{f}(X_i)\right]$$

(d) Deduce that in the repeated fixed design setting

$$\mathsf{Opt}_{RF} = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{(X_i)_{i=1}^n}\left[\mathbb{C}\mathrm{ov}\left[Y_i, \hat{f}(X_i)\right]\middle|(X_i)_{i=1}^n\right]$$

4. We consider now the random design and define Opt_R in the same way.

(a) Verify that

$$\mathsf{Err}_R = \mathsf{Err}_F + \underbrace{\mathbb{E}_{(X_i)_{i=1}^n, X'}\left[(f(X') - \mathbb{E}_{(Y_i)}\left[\hat{f}(X')\middle|(X_i)_{i=1}^n\right])^2\right] - \mathbb{E}_{(X_i)}\left[(f(X_1) - \mathbb{E}_{(Y_i)}\left[\hat{f}(X_1)\middle|(X_i)\right])^2\right]}_{\Delta_B}$$
$$+ \underbrace{\mathbb{E}_{(X_i)_{i=1}^n, X'}\left[\mathbb{V}\mathrm{ar}_{(Y_i)}\left[\hat{f}(X')\middle|(X_i)_{i=1}^n\right]\right] - \mathbb{E}_{(X_i)_{i=1}^n}\left[\mathbb{V}\mathrm{ar}_{(Y_i)}\left[\hat{f}(X_1)\middle|(X_i)_{i=1}^n\right]\right]}_{\Delta_V}$$

(b) Deduce that

$$\mathsf{Opt}_{RF} = \mathsf{Opt}_F + \Delta_B + \Delta_V$$

(c) Why is this reasonable to expect that both Δ_B and Δ_V are non negative?

5. We consider now the linear model $f_\beta(X') = \langle X', \beta\rangle$ and pick β by minimizing the empirical loss.

(a) Prove that, assume the design matrix $\mathbb{X}$ is such that $\mathbb{X}^\top\mathbb{X}$ is an invertible $d\times d$ matrix, the estimate $\widehat{\beta}$ is given by

$$\widehat{\beta} = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\mathbf{Y}$$

so that $\hat{f}(X_i) = \left(\mathrm{Proj}_{\mathrm{span}\,(X_i)}\,\mathbf{Y}\right)_i$.

(b) Deduce that $\mathsf{Opt}_R = \mathsf{Opt}_{RF} = 2\sigma^2\frac{d}{n}$

(c) Prove that $\Delta_B \geq 0$.

(d) Prove that $\Delta_V \geq 0$ (one can use $\mathbb{E}\left[\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\right] - \left(\mathbb{E}\left[\mathbb{X}^\top\mathbb{X}\right]\right)^{-1}$ is s.d.p.)