Statistique en grande dimension

Lecturer : Dalalyan A., Scribe : Thomas F.-X.

Second lecture

1 Previously...

$Z_1, \dots, Z_n$ are iid, their common law is $\mathbb{P}$ (unknown).
There are 2 types of learning :

- Supervised : $Z_i = (X_i, Y_i)$
- Unsupervised : $Z_i = X_i$

Predictions $(X, Y) \in \mathcal{X} \times \mathcal{Y}$: Given an example X, we'd like to predict the value of Y.

- Binary classification : $\mathcal{Y} = \{0, 1\}$ and $\ell(y, y') = \mathbf{1}(y \neq y')$

$$g^*(x) = \mathbf{1}\left(\eta^*(x) > \frac{1}{2}\right) \qquad \text{with} \qquad \eta^*(x) = \mathbb{E}[Y|X = x]$$

- Least-squares regression : $\mathcal{Y} \subset \mathbb{R}$, and $\ell(y, y') = (y - y')^2$

$$g^*(x) = \eta^*(x) = \mathbb{E}[Y| = x]$$

Risk

$$\mathbb{R}[g] = \mathbb{E}_{\mathbb{P}}[\ell(Y, g(X))]$$

Bayes Predictor

$$g^* \in \arg\min_{g:\mathcal{X}\to\mathcal{Y}} \mathbb{R}[g]$$

Excess risk

$$\mathbb{R}[\hat{g}] - \mathbb{R}[g^*] \geqslant 0$$

2 Link between Binary Classification & Regression

2.1 Plug-in rule

- We start by estimating $\eta^*(x)$ by $\hat{\eta}_n(x)$,
- We define $\hat{g}_n(x) = \mathbf{1}\left(\hat{\eta}_n > \frac{1}{2}\right)$.

Question: *How good the plug-in rule $\hat{g}_n$ is ?*

Proposition 1 *Let $\hat{\eta}$ be an estimator of the regression function η^*, and let $\hat{g}(x) = \mathbf{1}\left(\hat{\eta}(x) > \frac{1}{2}\right)$. Then, we have :*

$$\mathcal{R}_{\text{class}}[\hat{g}] - \mathcal{R}_{\text{class}}[g^*] \leqslant 2\sqrt{\mathcal{R}_{\text{reg}}[\hat{\eta}] - \mathcal{R}_{\text{reg}}[\eta^*]}$$

Proof Let $\eta : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$, and $g(x) = \mathbf{1}\left(\eta(x) > \frac{1}{2}\right)$, and let's compute the excess risk of g. By definition:

$$\mathcal{R}_{\text{class}}[g] - \mathcal{R}_{\text{class}}[g^*] = \mathbb{E}\left[\mathbf{1}\left(Y \neq g(X)\right) - \mathbf{1}\left(Y \neq g^*(X)\right)\right]$$

Simple algebra yields:

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}\left(Y \neq g(X)\right)\right] &= \mathbb{E}\left[\mathbf{1}\left(Y \neq g(X)\right)\mathbf{1}\left(Y = 1\right)\right] + \mathbb{E}\left[\mathbf{1}\left(Y \neq g(X)\right)\mathbf{1}\left(Y = 0\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left(\mathbf{1}\left(1 \neq g(x)\right)\mathbf{1}\left(Y = 1\right)|X\right)\right] + \mathbb{E}\left[\mathbb{E}\left(\mathbf{1}\left(0 \neq g(X)\right)\mathbf{1}\left(Y = 0\right)|X\right)\right].\end{aligned}$$

Let us compute now the conditional expectations appearing above:

$$\begin{aligned}\mathbb{E}\Big[\underbrace{\mathbf{1}(1 \neq g(X))}_{=1-g(X)} \cdot \mathbf{1}(Y = 1) \,|\, X\Big] &= \big(1 - g(X)\big)\mathbb{P}\left(Y = 1|X\right) = \big(1 - g(X)\big)\eta^*(X) \\ \mathbb{E}\Big[\underbrace{\mathbf{1}(0 \neq g(X))}_{=g(X)} \cdot \mathbf{1}(Y = 0) \,|\, X\Big] &= g(X)\mathbb{P}(Y = 0|X) = g(X)\left(1 - \eta^*(X)\right)\end{aligned}$$

Combining these relations, we get

$$\mathbb{E}\left[\mathbf{1}\left(Y \neq g(X)\right)\right] = \mathbb{E}[\eta^*(X)] + \mathbb{E}\left[g(X)\left(1 - 2\eta^*(X)\right)\right].$$

Therefore,

$$\mathcal{R}_{\text{class}}[g] - \mathcal{R}_{\text{class}}[g^*] = \mathbb{E}\left[\left(g(X) - g^*(X)\right)\left(1 - 2\eta^*(X)\right)\right].$$

Since g and g^* are both indicator functions and, therefore, take only the values 0 and 1, their difference will be nonzero if and only if one of them is equal to 1 and the other one is equal to 0. This leads to

$$\begin{aligned}\mathcal{R}_{\text{class}}[g] - \mathcal{R}_{\text{class}} &\leq \mathbb{E}\left[\mathbf{1}\left(\eta(X) \leqslant \frac{1}{2} < \eta^*(X)\right)|2\eta^*(X) - 1|\right] \\ &\quad + \mathbb{E}\left[\mathbf{1}\left(\eta^*(X) \leqslant \frac{1}{2} < \eta(X)\right)|2\eta^*(X) - 1|\right] \\ &= 2\mathbb{E}\left[\mathbf{1}\left(\frac{1}{2} \in [\eta^*(X), \eta(X)]\right)\left|\eta^*(X) - \frac{1}{2}\right|\right]\end{aligned}$$

If $\eta(X) \leqslant \frac{1}{2}$ and $\eta^*(X) > \frac{1}{2}$, then $\left|\eta^*(X) - \frac{1}{2}\right| \leqslant |\eta^*(X) - \eta(X)|$, and thus :

$$\begin{aligned}\mathcal{R}_{\text{class}}[g] - \mathcal{R}_{\text{class}}[g^*] &\leqslant 2\mathbb{E}\left[\mathbf{1}\left(\frac{1}{2} \in [\eta(X), \eta^*(X)]\right)|\eta(X) - \eta^*(X)|\right] \\ &\leqslant 2\mathbb{E}\left[|\eta(X) - \eta^*(X)|\right] \\ &\leqslant 2\sqrt{\mathbb{E}\left[\left(\eta(X) - \eta^*(X)\right)^2\right]} = 2\sqrt{\mathcal{R}_{\text{reg}}(\eta) - \mathcal{R}_{\text{reg}}(\eta^*)}.\end{aligned}$$

Since this inequality is true for every deterministic η, we get the desired property. □

2.2 Link between the classification and density estimation

Framework: $(X_0, Y_0), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} \mathbb{P}$.

Assume that, conditionally to Y, $\mathbb{P}$ admits a density on $\mathcal{X}$. Let $f_1(x)$ be the density of $\mathcal{X}$ conditionally to $Y = 1$, and f_0 likewise to $Y = 0$. Let $\pi_1 = \mathbb{P}(Y = 1)$ and $\pi_0 = 1 - \pi_1$. The well-known Bayes formula implies that:

$$\eta^*(x) = \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x) = \frac{f_1(x)\pi_1}{f_1(x)\pi_1 + f_0(x)\pi_0}$$

Henceforth, the Bayes rule can be written as:

$$g^*(x) = \begin{cases} 1, \text{ if } \eta^*(x) > \frac{1}{2} \\ 0, \text{ otherwise} \end{cases} = \begin{cases} 1, \text{ if } f_1(x)\pi_1 > f_0\pi_0 \\ 0, \text{ otherwise} \end{cases}$$

General approach

- Estimate π_0, π_1 by :
$$\hat{\pi}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad \hat{\pi}_0 = 1 - \hat{\pi}_1$$
- Use a non-parametric method for estimating f_1 and f_0.
- Set $\hat{g}(x) = \mathbf{1}\left(\hat{f}_1(x)\hat{\pi}_1 > \hat{f}_0(x)\hat{\pi}_0\right)$

Multi-scale classification In the case where $\mathcal{Y} = \{a_1, \ldots, a_M\}$, similar arguments yield

$$g^*(x) = \arg\max_{a \in \mathcal{Y}} \mathbb{P}(Y = a|X = x).$$

3 Density model

Let now $X_1, \ldots, X_n$ be iid with density f. To start with, we assume that $X_i \in [0, 1]$; the relaxations of this condition will be discussed later.

Let's assume

$$f \in \text{Lip}(L) = \{f : [0, 1] \to \mathbb{R} : |f(x) - f(y)| \leqslant L|x - y|, \forall x, y \in [0, 1]\}.$$

We are going to estimate f by a histogram. The set $\text{Lip}(L)$ can be approximated by a K-dimensional space Θ_K :

$$\Theta_K = \left\{\theta \in \mathbb{R}^K, \theta_k \geq 0, \sum_{k=1}^{K} \theta_k = 1\right\}$$

In fact, the following application is an injection from Θ_K to $\text{Lip}(L)$:

$$\theta \in \Theta_K \mapsto f_\theta(x) = \frac{\theta_k}{h}, \quad \forall x \in [(k-1)h, kh[, \qquad \text{where} \quad h = \frac{1}{K}.$$

Under the assumption that $f = f_\theta$ with $\theta \in \Theta_K$, the estimation of f is equivalent to estimating θ, which we can be done by the method of maximum likelihood:

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta_K} \prod_{i=1}^{n} f_\theta(X_i) = \arg\max_{\theta \in \Theta_K} \sum_{i=1}^{n} \ln f_\theta(X_i)$$

It is easy to check (**good exercise for the students**) that the maximum is attained at:

$$\hat{\theta}_n = \left(\hat{\theta}_n^{(1)}, \ldots, \hat{\theta}_n^{(K)}\right) \quad \text{st} \quad \hat{\theta}_n^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(X_i \in [(k-1)\,h, kh[\right)$$

The following estimator is then called *histogram with bandwidth* $h = \frac{1}{K}$:

$$\hat{f}_{n,h}(x) = \frac{1}{nh}\sum_{k=1}^{K}\left[\mathbf{1}_{C_{k,h}}(x)\sum_{i=1}^{n}\mathbf{1}_{C_{k,h}}(X_i)\right] \quad \text{with} \quad C_{k,h} = [(k-1)\,h, kh[\,.$$

We have 2 cases :

- **If K is too large : *overfitting***
- **If K is too small : *underfitting***

Remark 1 If $x \in C_{k,h}$, then :

$$nh\hat{f}_{n,h}(x) = \sum_{i=1}^{n}\mathbf{1}_{C_{k,h}}(X_i) \sim \mathcal{B}\left(n, \int_{C_k} f(x)\,\mathrm{d}x\right)$$

where $\mathcal{B}(n,p)$ stands for the binomial distribution with parameters (n,p). Note that the integral over C_k in the above formula comes from the fact that:

$$\mathbb{P}\left(\mathbf{1}_{C_{k,h}}(X_i) = 1\right) = \mathbb{P}\left(X_i \in C_{k,h}\right) = \int_{C_k} f(x)\,\mathrm{d}x.$$

Proposition 2 *Let $X_1, \ldots, X_n$ be iid with density $f(x)$ and $K \in \mathbb{N}$. Set* where$p_j = \int_{C_j} f$ *for every $j = 1, \ldots, K$. The* Mean Integrated Squared Error *of the histogram $\hat{f}_{n,h}$ is given by :*

$$\mathrm{MISE}_f(h) \triangleq \mathbb{E}_f\left[\int_0^1 \left(\hat{f}_{n,h}(x) - f(x)\right)^2 \mathrm{d}x\right] = \int_0^1 f^2(x)\mathrm{d}x + \frac{1}{nh} - \frac{n+1}{nh}\sum_{i=1}^{K} p_j^2.$$

Proof Combining Fubini with bias-variance decomposition, we get:

$$\begin{aligned}
\mathbb{E}_f\left[\int_0^1 \left(\hat{f}_{n,h}(x) - f(x)\right)^2 \mathrm{d}x\right] &= \int_0^1 \mathbb{E}_f\left[\left(\hat{f}_{n,h}(x) - f(x)\right)^2\right]\mathrm{d}x \\
&= \sum_{k=1}^{K}\int_{C_k} \mathbb{E}_f\left[\left(\hat{f}_{n,h}(x) - f(x)\right)^2\right]\mathrm{d}x \\
&= \sum_{k=1}^{K}\int_{C_k}\left\{\left(\mathbb{E}_f\left[\hat{f}_{n,h}(x)\right] - f(x)\right)^2 + \mathrm{Var}_f\left[\hat{f}_{n,h}(x)\right]\right\}\mathrm{d}x.
\end{aligned}$$

It is clear that:

$$nh\hat{f}_{n,h}(x) \sim \mathcal{B}(n, p_k) \quad \Rightarrow \quad \mathbb{E}\left[\hat{f}_{n,h}(x)\right] = \frac{np_k}{nh} \quad \text{and} \quad \mathrm{Var}_f\left[\hat{f}_{n,h}(x)\right] = \frac{np_k(1-p_k)}{(nh)^2}.$$

Therefore,

$$\begin{aligned}
\mathrm{MISE}_f(h) &= \sum_{k=1}^{K}\left[\int_{C_k}\left(\frac{p_k}{h} - f(x)\right)^2\mathrm{d}x + \frac{p_k(1-p_k)}{nh}\right] \qquad (1) \\
&= \sum_{k=1}^{K}\left[-\frac{p_k^2}{h} + \int_{C_k} f^2\mathrm{d}x + \frac{p_k - p_k^2}{nh}\right] \qquad \text{(we used that } \textstyle\int_{C_k} f = p_k\text{)} \\
&= \int_0^1 f^2\mathrm{d}x + \frac{\sum_{i=1}^{K} p_k}{nh} - \frac{n+1}{nh}\sum_{k=1}^{K} p_k^2
\end{aligned}$$

The desired result follows from the obvious relation $\sum_k p_k = 1$. □

The result of the previous proposition is very useful, since it provides the explicit form of the risk. However, it is not very easy to see how this expression depends on the bandwidth h, which is the parameter responsible for overfitting or underfitting. Under the assumption that $f \in \mathrm{Lip}\,(L)$, we can get a simpler bound on the risk. Indeed, for every $x \in C_k$, we have :

$$\begin{aligned}
\left| f(x) - \frac{p_k}{h} \right| &= \left| f\,(x) - \frac{1}{h} \int_{C_k} f(y) \mathrm{d}(y) \right| \\
&= \frac{1}{h} \left| \int_{C_k} (f(x) - f\,(y))\, \mathrm{d}y \right| \\
&\leqslant \frac{1}{h} \int_{C_k} |f(x) - f\,(y)|\, \mathrm{d}y \\
&\leqslant \frac{L}{h} \int_{C_k} |x - y|\, \mathrm{d}y \\
&\leqslant Lh \qquad (2)
\end{aligned}$$

This leads to the following result.

Proposition 3 *If $f \in \mathrm{Lip}\,(L)$, then :*

$$\mathrm{MISE}_f(h) \leqslant L^2 h^2 + \frac{1}{nh}$$

Proof Using (1), (2) and $\sum_{k=1}^K p_k\,(1 - p_k) \leqslant \sum_{k=1}^K p_k = 1$, we get the result. □

Remark 2

- $L^2h^2 = \frac{L^2}{K^2}$ is an upper bound on the square of the bias. It describes the error of approximation.
- $\frac{1}{nh} = \frac{K}{n}$ is an upper bound on the variance of the estimator; it describes the estimation error.
- If we minimize $L^2h^2 + \frac{1}{nh}$ w.r.t. $h > 0$, we get

  $$h_{\mathrm{opt}} = \left(2nL^2\right)^{-1/3}, \qquad \text{and} \qquad \mathrm{MISE}\,(h_{\mathrm{opt}}) \leqslant 3(L/2)^{2/3}\, n^{-2/3}.$$

  (strictly worse than the parametric rate n^{-1})
- h_{opt} depends on L, which is unknown ! We can not use it in practice...