Statistique en grande dimension

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

First lecture

1 Introduction

1.1 Statistique classique

Statistique *paramétriques* : $Z_1, \ldots, Z_n$ iid, avec une loi commune $\mathbb{P}_\theta$

On fait l'hypothèse $\theta \in \Theta \subset \mathbb{R}^d$

- Connu : $Z_1, \ldots, Z_n$ et Θ
- Inconnu : θ ou $\mathbb{P}_\theta$

Hypothèse importante : d est fixe et $n \to +\infty$

On sait dans ce cas que l'estimateur du MV est asymptotiquement (le plus) efficace (convergent) : $\hat{\theta}_{\text{MV}}$ vérifie quand $n \to +\infty$:

$$\mathbb{E}_P \left[\left\| \hat{\theta}_{\text{MV}} - \theta \right\|^2 \right] = \frac{C}{n} \left(1 + o\left(1\right)\right)$$

On estime θ à une vitesse $\frac{1}{\sqrt{n}}$ (vitesse paramétrique)

Constat. Si $d = d_n$ t.q. $\lim_{n \to +\infty} d_n = +\infty$, alors toute la théorie paramétrique est inutilisable. De plus, l'estimateur du MV n'est **plus** le meilleur estimateur !

1.2 Statistique non paramétrique

On observe $Z_1, \ldots, Z_n$ iid de loi $\mathbb{P}$, inconnue, telle que $\mathbb{P} \in \{\mathbb{P}_\theta, \theta \in \Theta\}$, mais avec Θ soit de dimension infinie, soit de dimension $d = d_n$ finie mais $\to +\infty$ avec la taille de l'échantillon.

Exemples:

$$\Theta = \left\{ f : [0,1] \to \mathbb{R}, f \text{ Lipschitz de constante L} \right\} \tag{1}$$

$$= \left\{ f : [0,1] \to \mathbb{R}, \forall x, y, \left| f\left(x\right) - f\left(y\right) \right| \leqslant L \left| x - y \right| \right\} \tag{2}$$

$$\Theta = \left\{ \theta = \left(\theta_1, \theta_2, \ldots\right), \sum_{j=1}^{\infty} \theta_j^2 < +\infty \right\} = \ell_2 \tag{3}$$

Démarche générale:
On approche Θ par une suite croissante $\{\Theta_k\}$ de sous-ensembles de Θ telle que Θ_k est de

dimension d_k. En procédant comme si θ appartenait à Θ_k (ce n'est pas nécessairement le cas), on utilise une méthode paramétrique pour définir un estimateur $\widehat{\theta}_k$ de θ. Cela nous donne une famille d'estimateurs $\{\widehat{\theta}_k\}$.

Question principale. Comment choisir k pour minimiser le risque de $\widehat{\theta}_k$?

- Si k est petit, on est face à un phénomène de sous-apprentissage *(underfitting)*
- Inversement, si k est grand, phénomène de sur-apprentissage *(overfitting)*

1.3 Principal models in non-parametric statistics

Density model. We have $X_1, \ldots, X_n$ iid with a density f defined on $\mathbb{R}^p$, and :

$$\mathbb{P}(X_1 \in A) = \int_A f(x)dx$$

The assumptions imposed on f are very weak as opposed to the parametric setting. For instance, a typical assumption in parametric setting is that f is the Gaussian density :

$$f(x) = \frac{\det(\Sigma^{-1})}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right],$$

whereas a common assumption on f in nonparametric framework is : f is smooth, say, twice continuously differentiable with a bounded second derivative.

Regression model. We observe $Z_i = (X_i, Y_i)$, with input X_i, output Y_i and error ε_i :

$$Y_i = f(X_i) + \varepsilon_i.$$

The function f is called the regression function. Here, the goal is to estimate f without assuming any parametric structure on it.

Practical examples.

Marketing.

- Each i represents a consumer
- X_i are the features of the consumer

A typical question is "how do I estimate different relevant groups of consumers". A typical answer is then to use clustering algorithms. We assume that $X_1, \ldots, X_n$ are iid with density f. Then, we estimate f in a non-parametric manner by $\hat{f}$. The clusters are defined as regions around the local maxima of the function $\hat{f}$.

1.4 Machine Learning

- Essentially the same as non-parametric statistics
- The main focus here is on the algorithms (rather than on the models), their statistical performance and their computational complexity.

2 Main concepts and notations

Observations : $Z_1, \ldots, Z_n$ iid, $\mathbb{P}$

- Non-supervised learning : $Z_i = X_i$
- Supervised learning : $Z_i = (X_i, Y_i)$, where X_i is an example or a feature, and Y_i a label.

Aim. To learn the distribution $\mathbb{P}$ *or* some properties of it.

Prediction. We assume that a new feature X (from the same prob. distribution as $X_1, \ldots, X_n$) is observed. The aim is to predict the label associated to X.

To measure the quality of a prediction, we need a loss function $\ell(y, \tilde{y})$ (y is the true label, $\tilde{y}$ is the predicted label). In practice, both y and $\tilde{y}$ are random variables, furthermore y and its distribution are unknown, so ℓ is hard to compute!

Risk function. This is the expectation of the loss.

Definition 1 *Assume that $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function. A predictor, or preduction algorithm, is any mapping :*

$$\hat{g} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}$$

The risk of the prediction function g is :

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

The risk of a predictor $\hat{g}$ is $\mathcal{R}_P[\widehat{g}]$, which is random since $\hat{g}$ depends on the data.

$$\mathcal{R}_P[\widehat{g}] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, \hat{g}(x)) \, \mathrm{d}\mathbb{P}(x, y)$$

Examples:

1. Binary classification: $\mathcal{Y} = \{0, 1\}$, with any $\mathcal{X}$

$$\ell(y, \tilde{y}) = \begin{cases} 0, \text{ if } y = \tilde{y} \\ 1, \text{ otherwise} \end{cases} = \mathbf{1}(y \neq \tilde{y}) = (y - \tilde{y})^2.$$

2. Least-squares regression: $\mathcal{Y} \subset \mathbb{R}$, with any $\mathcal{X}$

$$\ell(y, \tilde{y}) = (y - \tilde{y})^2.$$

3 Excess risk and Bayes predictor

We have $Z_i = (X_i, Y_i)$

$$\mathcal{R}_P[g] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, g(x)) \, \mathbb{P}(\mathrm{d}x, \mathrm{d}y)$$
$$\mathbb{P}(\mathrm{d}x, \mathrm{d}y) = \mathbb{P}_{Y|X}(\mathrm{d}y | X = x) \, \mathbb{P}_X(\mathrm{d}x)$$

Definition 2 *Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the Bayes predictor, or "oracle" is the prediction function minimizing the risk :*

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

Remark 1 In practice, g^* is unavailable, since it depends on $\mathbb{P}$, which is unknown. The ultimate goal is to do almost as well as the oracle.

A predictor $\hat{g}_n$ will be considered as a good one if :

$$\lim_{n\to+\infty} \underbrace{\mathcal{R}_P\left[\hat{g}_n\right] - \mathcal{R}_P\left[g^*\right]}_{\text{excess risk}} = 0$$

Definition 3 *We say that the predictor $\hat{g}_n$ is consistent (universally consistent) if $\forall \mathbb{P}$, we have :*

$$\lim_{n\to+\infty} \mathbb{E}_P\left[\mathcal{R}_P\left[\hat{g}_n\right]\right] - \mathcal{R}_P\left[g^*\right] = 0$$

Theorem 1

1. *Suppose that $\forall x \in \mathcal{X}$,the infimum of $y \mapsto \mathbb{E}_P\left[\ell\left(\mathcal{Y}, y\right)|X = x\right]$ is reached. Then the funcion g^* defined by :*
$$g^*(x) \in \arg\min_{y\in\mathcal{Y}} \mathbb{E}_P\left[\ell\left(Y, y\right)|X = x\right]$$
...is a Bayes predictor.

2. *In the case of the binary classification, $\mathcal{Y} = \{0, 1\}$ and $\ell\left(y, \tilde{y}\right) = \mathbf{1}\left(y \neq \tilde{y}\right)$,*
$$g^*(x) = \mathbf{1}\left(\eta^*\left(x\right) > \frac{1}{2}\right) \quad \text{where} \quad \eta^*\left(x\right) = \mathbb{P}\left[Y = 1|X = x\right].$$
Furthermore, the excess risk can be computed by
$$\mathcal{R}_P\left[g\right] - \mathcal{R}_P\left[g^*\right] \quad = \quad \mathbb{E}_P\left[\left(g(X) - g^*(X)\right)\left(1 - 2\eta^*(X)\right)\right]. \tag{4}$$

3. *In the case of the least squares regression,*
$$g^*(x) = \eta^*(x) \quad \text{where} \quad \eta^*(x) = \mathbb{E}_P\left[Y|X = x\right]$$
Furthermore, for any $\eta : \mathcal{X} \to \mathcal{Y}$, we have :
$$\mathcal{R}_P[\,\eta\,] - \mathcal{R}_P\left[\eta^*\right] = \mathbb{E}_P\left[\left(\eta\left(X\right) - \eta^*\left(X\right)\right)^2\right]$$

Proof

1. Let $g \in \mathcal{Y}^{\mathcal{X}}$ and let :
$$g^*(x) \in \arg\min_{y\in\mathcal{Y}} \mathbb{E}_P\left[\ell\left(Y, y\right)|X = x\right].$$
We have :
$$\begin{aligned}
\mathcal{R}_P[g] &= \mathbb{E}_P\left[\ell\left(Y, g\left(X\right)\right)\right] \\
&= \int \mathbb{E}_P\left[\ell\left(Y, g\left(X\right)\right)|X = x\right]\mathbb{P}_X\left(\mathrm{d}x\right) \\
&\geqslant \int \mathbb{E}_P\left[\ell\left(Y, g^*(x)\right)|X = x\right]\mathbb{P}_X\left(\mathrm{d}x\right) \\
&= \mathcal{R}_P\left[g^*\right].
\end{aligned}$$

2. Using the first assertion,

$$\begin{aligned} g^*(x) &\in \arg\min_{y\in\{0,1\}} \mathbb{E}_P\left[\mathbf{1}\left(Y\neq y\right)|X=x\right] \\ &= \arg\min_{y\in\{0,1\}} \mathbb{P}\left(Y\neq y|X=x\right) \\ &= \arg\max_{y\in\{0,1\}} \mathbb{P}\left(Y=y|X=x\right) \\ &= \arg\max_{y\in\{0,1\}} \Big\{\eta^*(x)\mathbf{1}(y=1)+(1-\eta^*(x))\mathbf{1}(y=0)\Big\}. \end{aligned}$$

Therefore,

$$g^*(x) = \begin{cases} 0, & \text{if } \mathbb{P}\left(Y=1|X=x\right)\leqslant \frac{1}{2} \\ 1, & \text{otherwise.} \end{cases}$$

To check (4), it suffices to remark that

$$\begin{aligned} \mathcal{R}_P\left[g\right] &= \mathbb{E}_P[(g(X)-Y)^2] = \mathbb{E}_P[g(X)^2]+\mathbb{E}_P[Y^2]-2\mathbb{E}_P[Yg(X)] \\ &= \mathbb{E}_P[g(X)]+\mathbb{E}_P[Y]-2\mathbb{E}_P[\mathbb{E}_P(Yg(X)|X)] \\ &= \mathbb{E}_P[g(X)]+\mathbb{E}_P[Y]-2\mathbb{E}_P[g(X)\mathbb{E}_P(Y|X)] \\ &= \mathbb{E}_P[g(X)]+\mathbb{E}_P[Y]-2\mathbb{E}_P[g(X)\eta_P^*(X)] \\ &= \mathbb{E}_P[g(X)(1-2\eta_P^*(X)]+\mathbb{E}_P[Y]. \end{aligned}$$

Writing the same identity for g_P^* and making the difference of these two identities, we get the desired result.

3. In view of the first assertion of the theorem, we have:

$$g^*(x)\in \arg\min_{y\in\mathbb{R}} \mathbb{E}_P\left[\left(Y-y\right)^2|X=x\right] = \arg\min_{y\in\mathbb{R}} \varphi\left(y\right)$$

where $\varphi\left(y\right) = \mathbb{E}_P\left[Y^2|X=x\right] - 2y\mathbb{E}_P\left[Y|X=x\right]+y^2$ is a second order polynomial. The minimization of such a polynomial is straightforward and leads to:

$$\arg\min_{y\in\mathbb{R}} \varphi\left(y\right) \quad = \quad \mathbb{E}_P\left[Y|X=x\right].$$

This shows that the Bayes predictor is equal to the regression function $\eta^*(x)$. The risk of this predictor is:

$$\begin{aligned} \mathcal{R}_P[\,\eta\,] &= \mathbb{E}_P\left[\left(Y-\eta\left(X\right)\right)^2\right] \\ &= \mathbb{E}_P\left(\mathbb{E}_P\left[\left(Y-\eta\left(X\right)\right)^2|X\right]\right) \\ &= \mathbb{E}_P\left(\mathbb{E}_P\left[\left(Y-\eta^*\left(X\right)\right)^2|X\right]+2\mathbb{E}_P\left[\left(Y-\eta^*\left(X\right)\right)\left(\eta^*-\eta\right)\left(X\right)|X\right]+\left(\eta^*-\eta\right)^2\left(X\right)\right) \\ &= \mathcal{R}_P\left[\eta^*\right]+0+\mathbb{E}_P\left[\left(\eta^*-\eta\right)^2\left(X\right)\right], \end{aligned}$$

where the cross-product term vanishes since

$$\mathbb{E}_P\left[\left(Y-\eta^*\left(X\right)\right)\left(\eta^*-\eta\right)\left(X\right)|X\right] = \left(\eta^*-\eta\right)\left(X\right)\mathbb{E}_P\left[\left(Y-\eta^*\left(X\right)\right)|X\right]=0.$$

This completes the proof of the theorem. □

3.1 Link between Binary Classification & Regression

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 :*

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

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. We have,

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

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

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

If $\eta(X) \leqslant 1/2$ and $\eta^*(X) > 1/2$, then $\big|\eta^*(X) - 1/2\big| \leqslant \big|\eta^*(X) - \eta(X)\big|$, and thus :

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

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