Topics in Statistical Theory

Course outline

1. Introduction
2. Density Estimation
3. Regression Problems
4. Minimax lower bounds
5. Classification problems
6. ?

Introduction

Parametric vs non-parametric

A statistical model specifies a family of possible data generating mechanisms. For example

1) Let $X_1, \dots X_n \overset{iid}{\sim} \Gamma(m, \theta)$ where m is known, and $\theta \in \Theta = (0, \infty)$ is an unknown parameter

2) Linear model : $Y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1, \dots, n$, where $x_1, \dots, x_n$ are known and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Here the unknown parameter is 3-dimensional $\theta = (\alpha, \beta, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R} \times (0, \infty)$

If the parameter space Θ is finite dimensional, we speak of a parametric model. Often in such cases, we can use MLE $\hat{\theta}_n$ to estimate ~~the~~ θ, and have $\hat{\theta}_n - \theta = O_p(n^{-1/2})$.

This rate assumes that the model contains the true data generating process. If this is not the case, our inferences may be misleading.

We may instead use a non-parametric model. Some examples:

3) Let $X_1, \dots X_n \overset{iid}{\sim} F$, where F is an arbitrary distribution function

4) Let $X_1, \dots X_n \overset{iid}{\sim} f$, where f is a twice differentiable density function

5) Shape constrained estimation

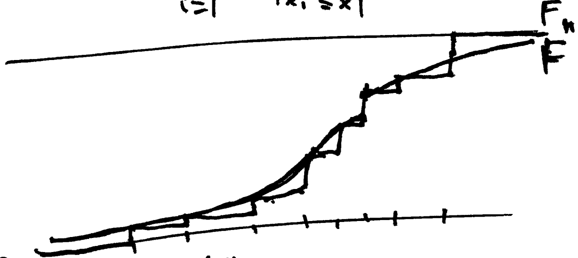
$Y_i = m(x_i) + \varepsilon_i$, $i = 1, \dots, n$, where $x_1, \dots, x_n$ known, $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and m is a monotone increasing function.

Such ∞-dimensional models are less vulnerable to ~~overfitting~~ misspecification, typically however we pay a price in terms of the rate of convergence, e.g. $O_p(n^{-2/5})$ in example 4.

Estimating an arbitrary dist^n f^n

Let $X_1, \dots X_n \overset{iid}{\sim} F$. The empirical distribution function, $\hat{F}_n$ is given by

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}$$



$$\mathbb{E}(\mathbb{1}_{\{X_i \le x\}}) = F(x)$$

Theorem 1 (Glivenko – Cantelli, 1933)
(the Fundamental Theorem of Statistics)

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \to 0 \text{ a.s.}$$

Proof Given $\varepsilon > 0$, choose a partition

$-\infty = x_0 < x_1 < \dots < x_k = \infty$ such that

$F(x_i^-) - F(x_{i-1}) \leq \varepsilon$ where

$$F(x^-) := \lim_{y \uparrow x} F(y)$$

Note that any point at which F jumps by more than ε, must be in the partition.

Now by the SLLN, there exists an event Ω_0 with $\mathbb{P}(\Omega_0) = 1$, such that for all $\omega \in \Omega_0$, $\exists\, n_0 = n_0(\omega) \in \mathbb{N}$ with the property that for all $n \geq n_0$

$$|\hat{F}_n(x_i) - F(x_i)| \leq \varepsilon \quad \text{and}$$

$$|\hat{F}_n(x_i^-) - F(x_i^-)| \leq \varepsilon$$

for all $i = 1, \dots, k-1$

Fix $x \in \mathbb{R}$ and find i such that ~~the~~ $x \in [x_{i-1}, x_i)$. For $\omega \in \Omega_0$ $n \geq n_0(\omega)$

$$\begin{aligned}
\hat{F}_n(x) - F(x) &\leq \hat{F}_n(x_i^-) - F(x_{i-1}) \\
&\leq \hat{F}_n(x_i^-) - F(x_i^-) \\
&\quad + F(x_i^-) - F(x_{i-1}) \\
&\leq \varepsilon + \varepsilon
\end{aligned}$$

In the same way can show

$$F(x) - \hat{F}_n(x) \leq 2\varepsilon$$

for all $n \geq n_0(\omega)$

In fact, much more is true.

Theorem 2

Let $x_1, \dots, x_n \overset{iid}{\sim} F$. Then for all $\varepsilon > 0$

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| > \varepsilon \right) \leq 2 e^{-2n\varepsilon^2}$$