Lecture 18

Last time

- Matrix Calculus
- Matrix Chernoff
- Matrix Hoeffding

Today

- Intro to estimation
- Maximum likelihood estimation

Intro to estimation

So far we have developed foundational tools, in the time we have left we leverage these tools for two important statistical tasks: **estimation** and **learning**.

We start talking about estimation. Suppose we are given a sample

$$Z_1, \dots, Z_n \sim P_\theta$$ ← supported in $\mathbb{R}^d$

where P_θ is a distribution that depends on some object θ

Goal: Estimate θ accurately.

Examples: Mean, Covariance, CDF.

There are two types of estimation:

▷ **Parametric:** Suppose that $\theta \in \mathbb{R}^k$ and we can describe P_θ via a finite-dimensional parameterization involving θ.

Example: We wish to estimate $\theta \in \mathbb{R}^d$ and assume

$$Z_i \sim N(\theta, I).$$

▷ **Nonparametric:** P_θ does not belong to some "nice" finite dimensional family.

Example: The object we want to estimate θ is the density of the distribution and it can be any continuous function.

We will focus on parametric problems.

Def: An estimator of θ is a (measurable) function of the data

$$\hat{\theta}_n = g(X_1, \ldots, X_n).$$

We say that an estimator is unbiased if

$$\mathbb{E}\,\hat{\theta}_n = \theta.$$

How do we measure accuracy? One popular way is via the Mean Squared Error:

$$MSE(\hat{\theta}_n) := \mathbb{E}\,\|\theta - \theta_n\|_2^2.$$

In turn, this can be decompsed as

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}\,\|\theta - \mathbb{E}\hat{\theta}_n + \mathbb{E}\hat{\theta}_n - \hat{\theta}_n\|_2^2 \\ &= \mathbb{E}\,\|\theta - \mathbb{E}\hat{\theta}_n\|_2^2 \\ &\quad + \mathbb{E}\langle \theta - \mathbb{E}\hat{\theta}_n,\ \mathbb{E}\hat{\theta}_n - \hat{\theta}_n\rangle \\ &\quad + \mathbb{E}\,\|\mathbb{E}\hat{\theta}_n - \hat{\theta}\|_2^2 \\ &= \underbrace{\mathbb{E}\,\|\theta - \mathbb{E}\hat{\theta}_n\|_2^2}_{Bias(\theta_n)^2 :=} + \underbrace{\mathbb{E}\,\|\hat{\theta}_n - \mathbb{E}\hat{\theta}_n\|_2^2}_{Var(\theta_n) :=}. \end{aligned}$$

Often, it pays to incur in some bias to reduce the overall MSE (see Stein's Paradox).

Maximum Likelihood Estimation

MLE gives a systematic way to come up with estimators.

Suppose we have a parametric family

$$\{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^k\}.$$

Density for P_θ (Probability of x for discrete dist)

Given that we observed $Z_1, \ldots, Z_n$, the likelihood function is

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

The log-likelihood is given by

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta).$$

The MLE is given by

$$\hat{\theta}_n \in \text{often equal } \operatorname*{arg\,max}_{\theta \in \Theta} \mathcal{L}_n(\theta).$$

Intuitively, this corresponds to estimating the parameter that makes the data more likely.

Note that we can equivalently have

$$\hat{\theta}_{MSE} \in \operatorname{argmax}_{\theta \in \Theta} \ell_n(\theta).$$

Let's work out a couple of examples.

Example: Suppose $Z_1, \dots, Z_n \sim \text{Ber}(p)$

Then

$$f(z; p) = p^z (1-p)^{1-z}.$$

Therefore

$$\ell_n(p) = \sum_{i=1}^{n} z_i \log p + \sum_{i=1}^{n} (1 - z_i) \log(1-p)$$

($S_n = \sum_{i=1}^{n} z_i$)

$$= S_n \log p + (n - S_n) \log(1-p)$$

This function is concave in p and $\nabla \ell_n(p) = 0$ reveals that

$$\hat{p}_{MLE} = \frac{1}{n} S_n. \quad \dashv$$

Example: Suppose that $\theta \in \mathbb{R}^d$ and we observe iid samples $Z_i = (X_i, y_i)$ with

$$y_i = \theta^T X_i + \varepsilon_i.$$

($\varepsilon_i \sim N(0, \sigma^2)$; $X_i \sim D$ or deterministic)

Equivalently,

$$y_i \sim N(\theta^T x_i, \sigma^2).$$

Our goal is to estimate θ. Notice that there is another parameter that we don't care about, i.e., the variance σ^2. These is known as a nuisance parameter. Often (somewhat magically) the MSE doesn't require to know nuisance parameters, e.g.,

$$\mathcal{L}_n(\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} e^{-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum (y_i - \theta^T x_i)^2\right)$$

$$\Rightarrow \ell_n(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2.$$

Notice that maximizing $\ell_n(\theta)$ is equivalent to

$$\hat{\theta}_{MLE} \overset{\text{Why?}}{=} \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \theta_i^T x_i)^2,$$

which recovers least squares.

4

Beyond just estimating a parameter we might want to get confidence intervals or use the sample/estimate for hypothesis testing; all of these are dubbed as "statistical inference" problems; take Statistical Theory I & II for more.

MSE via Stochastic Optimization

Notice that we can always write

$$\hat{\theta}_{MSE} \in \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log\left(f(z_i;\theta)\right).$$

Stochastic optimization aims to solve problems of the form

$$(P) \qquad \min_{\theta \in \Theta} \mathbb{E}_{z \sim D} F(\theta, z)$$

$F: \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}$

having access only to an iid sample $z_1, \ldots, z_n \sim D$. Since we don't have access to D, a natural approach is to solve

$$(P_n) \quad \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} F(\theta, z_i).$$

This is known as Empirical risk minimization. Clearly, MSE is an instance of ERM.

There are a number of interesting questions:

- How do we solve (P_n)?
  Take Nonlinear I and II.
- Say θ_n^* is a solution of (P_n) and θ^* is a solution of (P). Do we have that $\theta_n^* \xrightarrow{a.s.} \theta^*$? How about convergence rates?
  We will study this question for linear least squares.
- How about confidence intervals? High prob. guarantees?
  This is a potential project!