Lecture 21: November 14

Lecturer: Alessandro Rinaldo *Scribes: Keith Shannon*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

21.1 Spectral Clustering

Last time, the eigengap condition was:

$$min\left\{q, \frac{p-q}{2}\right\}n$$

however one would expect community detection to be easy if $q \to 0$. In that case, you should compute the two leading eigenvectors of A: $[\hat{v_1}, \hat{v_2}]$ and view this as n points in $\mathbb{R}^2$. Then peroform k-means clustering on these n vectors.

See (Lei & Rinaldo 2016) and "Tutorial on Spectral Clustering" by Ulrike von Luxburg.

21.2 Empirical Process Theory

21.2.1 Uniform Law of Large Numbers

Example: $X_1 ... X_n \sim F_x$ *iid* have empirical cdf:

$$x \mapsto \hat{F_n}(x) = \frac{1}{n}\sum^{n} \mathbf{1}\{x_i \leq x\}$$

For each fixed x, $|F_x(x) - \hat{F_n}(x)|$ is easy to bound, as the indicator function is a binomial random variable.

However in general, it is hard to bound the supremum $sup_x|F_x(x) - \hat{F_n}(x)|$. This leads us to:

21.2.2 Empirical Process Theory

If $X_1 ... X_n \sim P$ *iid* on $(\mathscr{X}, \mathscr{B})$, and $\mathscr{F}$ is a collection of real-valued function on $\mathscr{X}$: based on sample $(X_1 ... X_n)$ construct empirical measure P_n (a random probability measure on $(\mathscr{X}, \mathscr{B})$) such that:

$$\forall A \in \mathscr{B}, P_n(A) \to P(A) = \frac{1}{n}\sum_{i}^{n} \mathbf{1}\{X_i \in A\}$$

For each $f \in \mathscr{F}$, let $Pf = E_P[f(x)]$. So,

$$P_n f = E_{P_n}[f(x)] = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

In Empirical Process Theory, we want to compute usable bounds for:

$$||P - P_n||_{\mathscr{F}} = sup_{f\in\mathscr{F}}\,|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - E[f(X_i)])|$$

This is a stochastic function over the class of $\mathscr{F}$. Calculating $E[||P-P_n||_{\mathscr{F}}]$ is hard, see (van der Vaart & Wellner 2001) and (van der Geer 2001). So, the following sections will cover bounding methods.

Examples :

For the example of $(\mathscr{X}, \mathscr{B}) = (\mathbb{R}, \mathscr{B})$:

$\mathscr{F} = \{(-\infty, z], z \in \mathbb{R}\}$. If $f = (-\infty, z]$, then

$$E_{X\sim F_x}[f(X)] = P(X \le z) = F_x(z)$$

$$P_n f = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \le z\} = \hat{F_n}(z)$$

so,

$$||P - P_n||_{\mathscr{F}} = sup_{z\in\mathbb{R}}|F_x(z) - \hat{F_n}(z)|$$

For example, if X is a random vector in $\mathbb{R}^d$, and $||X|| = sup_{z\in S^{d-1}} z^T X$, let

$$x \in \mathbb{R}^d \mapsto f_z(x) = z^T x,\ then\ \mathscr{F} = \{fz, z \in S^{d-1}\}.$$

So we see that sometimes it is natural to express things as the supremum of a stochastic process.

Another example is the operator norm of empirical covariance:

$$||\Sigma - \hat{\Sigma_n}||_{op} = sup_{z\in S^{d-1}}|z^T(\Sigma - \hat{\Sigma_n})z|$$

21.2.3 Empirical Risk

Let $\mathscr{P} = \{P_\theta, \theta \in \Theta\}$ on $(\mathscr{X}, \mathscr{B})$, and $X_1...X_n \sim P_{\theta^*} \in \mathscr{P}$. This is a classic parametric setup.

Definition : Loss function $\mathscr{L} : \Theta \times \mathscr{X} \to \mathbb{R}_T$

Definition : Risk:

$$R(\theta, \theta^*) = E_{X\sim P_{\theta^*}}[\mathscr{L}(\theta, X)],\ \ \theta, \theta^* \in \Theta$$

This is the risk of thinking the parameter is θ when it is actually θ^*.

Definition : Empirical Risk:

$$\hat{R}(\theta, \theta^*) = \frac{1}{n}\sum_{i=1}^{n} \mathscr{L}(\theta, X_i) = P_n[\mathscr{L}(\theta, X)]$$

Let $\hat{\theta_n} \in argmin_{\theta \in \Theta} \hat{R}(\theta, \theta^*)$

Example : **KL Divergence**

$$\mathscr{L}(\theta, X) = ln(\frac{p_{\theta^*}(x)}{p_\theta(x)}), \quad p_\theta = \frac{dP_\theta}{d\mu}$$

$$R(\theta, \theta^*) = KL(P_{\theta^*}, P_\theta), \quad E_{P_{\theta^*}}[ln(\frac{dP_{\theta^*}}{dP_\theta}(x))]$$

p_θ is the density of P_θ. If $\hat{\theta_n}$ is a minimizer of empirical risk $\hat{R}(\theta, \theta^*)$, then $\hat{\theta_n}$ is an MLE of θ^*.

Aside : Usually when dealing with the supremum of an empirical process $||P - P_n||_{\mathscr{F}}$, it concentrates well around the expected value.

Example : **Classification problem to** ± 1

$X_i = (Y_i, Z_i) \in \{-1, 1\} \times \mathbb{R}^d$, $i = 1...d$.

Goal: Estimate $f : \mathbb{R}^d \mapsto \{-1, 1\}$ such that $P(f(Z) \neq Y)$ is smallest. That is, $(Z, Y) \sim P_{f^*}$, where f^* is the regression function.

In that case,

$$\mathscr{L}(f, (Z, Y)) = \begin{cases} 1, \ f(Z) \neq Y \\ 0, \ f(Z) = Y \end{cases}$$

f^* is the Bayes classifier:

$$f^*(Z) = \begin{cases} 1, \ \psi(Z) \geq \frac{1}{2} \\ -1, \ \psi(Z) < \frac{1}{2} \end{cases}$$

Empirical Risk of a classifier f is:

$$\hat{R}(f, f^*) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{f(Z_i) \neq Y_i\}$$

21.2.4 Excess Risk

Let $\hat{\theta_n}$ be the empirical risk minimizer:

$$R(\hat{\theta_n}, \theta^*) = E_{X \sim P_{\theta^*}}[\mathscr{L}(\hat{\theta_n}, X)], \quad X \perp \hat{\theta_n}$$

Definition : Excess Risk is:

$$ER(\hat{\theta_n}, \theta^*) = R(\hat{\theta_n}, \theta^*) - inf_{\theta \in \Theta} R(\theta, \theta^*)$$

The inf can be 0 if $\theta^* \in \Theta$.

The Excess Risk is the difference between risk at the MLE and the smallest possible risk given class Θ. Bounding this will require using a supremum. Letting $\theta_0 = argmin_{\theta \in \Theta} R(\theta, \theta^*)$:

$$ER(\hat{\theta_n}, \theta^*) = R(\hat{\theta_n}, \theta^*) - \hat{R}(\hat{\theta_n}, \theta^*) + \hat{R}(\hat{\theta_n}, \theta^*) - \hat{R}(\theta_0, \theta_x) + \hat{R}(\theta_0, \theta_x) - R(\theta_0, \theta_x)$$

Grouping as follows:

$$\begin{aligned} T_1 &= R(\hat{\theta_n}, \theta^*) - \hat{R}(\hat{\theta_n}, \theta^*) \\ T_2 &= \hat{R}(\hat{\theta_n}, \theta^*) - \hat{R}(\theta_0, \theta_x) \\ T_3 &= \hat{R}(\theta_0, \theta_x) - R(\theta_0, \theta_x) \\ ER(\hat{\theta_n}, \theta^*) &= T_1 + T_2 + T_3 \end{aligned} \tag{21.1}$$

We can bound the terms seperately:

- $T_2 \leq 0$
- The second term in T_3 is the expected value of the first, so we can use LLN
- T_1 is hard to bound. We need to sup out the dependence between $\hat{R}$ and $\hat{\theta}$.

$$T_1 \leq sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} (\mathscr{L}(\theta, X_i) - E[\mathscr{L}(\theta, X_i)] \right| = ||P_n - P||_{\mathscr{F}}$$

where $\mathscr{F}$ is a class of form $\mathscr{F} = \{\mathscr{L}(\theta, \cdot), \theta \in \Theta\}$, i.e. $x \mapsto \mathscr{L}(\theta, x)$.

For example, a discretization argument is a method of suping out.

21.2.5 ULLN: Rademacher Complexities

$\mathscr{F}$ is our target function class. $X_1^n = (x_1 ... x_n) \in \mathscr{X}^n$.

$$\mathscr{F}(X_1^n) = \{(f(x_1)...f(x_n)) \in \mathbb{R}^n, f \in \mathscr{F}\} \subset \mathbb{R}^n$$

So the function class specifies a subset of $\mathbb{R}^n$.

Definition : Empirical Rademacher Complexity of $\mathscr{F}(X_1^n)$ is:

$$R_n(\mathscr{F}(X_1^n)) = E_\epsilon[sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right|]$$

where $\epsilon = (\epsilon_1 ... \epsilon_n) \sim$ Rademacher, $\epsilon_i = \pm 1 \; w.p. \frac{1}{2}$

This is the average maximal "correlation" of vectors in $\mathscr{F}(x_1^n)$ with pure noise.

Definition : Rademacher Complexity of $\mathscr{F}$ is:

$$R_n(\mathscr{F}) = E_{X,\epsilon}[sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right|], \quad X = (x_1 ... x_n) \perp \epsilon = (\epsilon_1 ... \epsilon_n)$$

This is a measure of how well $\mathscr{F}$ can fit pure noise. If $\mathscr{F}$ is large it will fit noise better, so this measures the complexity of the class.

The main point is that as $n \to \infty$, $R_n(\mathscr{F}) \to 0$ iff $||P_n - P||_{\mathscr{F}} \to 0$ in probability. Also, the rates at which they go to 0 will depend on each other.