Theoretical Statistics. Lecture 8.

Peter Bartlett

1. Uniform laws of large numbers:
   (a) Glivenko-Cantelli Theorem.
   (b) Glivenko-Cantelli classes.

Glivenko-Cantelli Theorem

First example of a uniform law of large numbers.

Theorem: $\|F_n - F\|_\infty \overset{as}{\to} 0$.

Here, F is a cumulative distribution function, F_n is the empirical cumulative distribution function,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1[X_i \leq x],$$

where $X_1, \ldots, X_n$ are i.i.d. with distribution F, and $\|F - G\|_\infty = \sup_t |F(t) - G(t)|$.

Glivenko-Cantelli Theorem

Why *uniform* law of large numbers?

$$\begin{aligned} \|F_n - F\|_\infty &= \sup_x |F_n(x) - F(x)| \\ &= \sup_x |P_n(X \leq x) - P[X \leq x]| \\ &\stackrel{as}{\to} 0, \end{aligned}$$

where P_n is the empirical distribution that assigns mass $1/n$ to each X_i.

The law of large numbers says that, for all x, $P_n(X \leq x) \stackrel{as}{\to} P(X \leq x)$. The GC Theorem says that this happens uniformly over x.

Glivenko-Cantelli Theorem: Applications

Why is it useful? Often, we construct estimators of various parameters of interest $\theta(F)$ by replacing F with F_n. These are called *plug-in* estimators. Examples:

1. $\theta(F) = \mathbf{E}X$, where $X \sim F$. $\theta(\hat{F}) = \frac{1}{n}\sum_{i=1}^{n} X_i = P_n X$.
2. $\theta(F) = \inf\{x : F(x) \geq 1/2\}$, the median. More generally, $\theta(F) = \inf\{x : F(x) \geq \alpha\}$, the α-quantile.

$$\theta(\hat{F}) = \inf\left\{x : \frac{1}{n}\sum_{i=1}^{n} 1[X_i \leq x] \geq \alpha\right\}.$$

If θ is continuous wrt $\|\cdot\|_\infty$, then we immediately get $\theta(\hat{F}) \stackrel{as}{\to} \theta(F)$.

Glivenko-Cantelli Classes

Definition: F is a **Glivenko-Cantelli class** for P if

$$\sup_{f \in F} |P_n f - P f| =: \|P_n - P\|_F \xrightarrow{P} 0.$$

Here, P is a distribution on $\mathcal{X}$, $X_1, \ldots, X_n$ are drawn i.i.d. from P, P_n is the empirical distribution (which assigns mass $1/n$ to each of $X_1, \ldots, X_n$), F is a set of measurable real-valued functions on $\mathcal{X}$ with finite expectation under P, $P_n - P$ is an **empirical process**, that is, a stochastic process indexed by a class of functions F, and $\|P_n - P\|_F := \sup_{f \in F} |P_n f - P f|$.

The GC Theorem is a special case, with $F = \{1[x \leq t] : t \in \mathbb{R}\}$ (and with the stronger conclusion that convergence is almost sure—we say that such an F is a 'strong GC class').

Glivenko-Cantelli Classes

Not all F are Glivenko-Cantelli classes. For instance, consider

$$F = \{1[x \in S] : S \subset \mathbb{R}, |S| < \infty\}.$$

Then for a continuous distribution P, $Pf = 0$ for any $f \in F$, but $\sup_{f \in F} P_n f = 1$ for all n. So although $P_n f \overset{as}{\to} Pf$ for all $f \in F$, this convergence is not uniform over F. F is too large.

Glivenko-Cantelli Classes: Measurability

We need to be careful when we're working with a *supremum of an empirical process* like $\|P_n - P\|_F$. See Pollard for some assumptions on F that ensure that this supremum is measurable. An alternative approach is to work, instead of with $\mathbf{E}\|P_n - P\|_F$, with

$$\sup \{\mathbf{E}\|P_n - P\|_G : G \subseteq F, |G| < \infty\} .$$

Then the max over G is measurable because it is over a finite set.

Empirical Risk Minimization

Why study GC classes? They are important for estimators based on empirical risk minimization.

Let's consider a decision theoretic setting: we define a loss function $\ell(\theta, z)$, which measures how bad it is to choose θ when the outcome is z.

Definition: For $Z \sim P$, the **risk** is $L(\theta) = P\ell(\theta, Z)$.

We aim to choose $\theta \in \Theta$ to minimize the risk.

Empirical Risk Minimization

Examples:

1. Pattern classification: $\theta : \mathcal{X} \to \{0, 1\}$, $z = (x, y) \in \mathcal{X} \times \{0, 1\}$, $\ell(\theta, (x, y)) = 1[\theta(x) \neq y]$. Then we aim to choose $\theta \in \Theta$ to minimize the probability of misclassification.

2. Density estimation: p_θ is a density, $X \sim P$, p_{θ^*}, $\ell(\theta, z) = -\log p_\theta(z)$. Then we aim to choose θ to minimize

$$\mathbf{E} \log \frac{p_{\theta^*}(X)}{p_\theta(X)} = D_{KL}(p_{\theta^*} || p_\theta).$$

3. Regression: $\theta \in \mathbb{R}^p$, $z = (x, y)$, $\ell(\theta, (x, y)) = |\theta' x - y|$. Then we aim to choose θ to minimize expected absolute error.

Empirical Risk Minimization

Suppose $Z_1, \ldots, Z_n$ are i.i.d. according to P.

Definition: Define the **empirical risk** as

$$L_n(\theta) = P_n \ell(\theta, Z) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, Z_i).$$

Empirical risk minimization chooses θ to minimize $L_n(\theta)$.

Empirical Risk Minimization

Examples:

1. Pattern classification: $\theta : \mathcal{X} \to \{0, 1\}$, $z = (x, y) \in \mathcal{X} \times \{0, 1\}$, $\ell(\theta, (x, y)) = 1[\theta(x) \neq y]$. Empirical risk minimization chooses θ to minimize misclassifications on the sample.

2. Density estimation: p_θ is a density, $X \sim P$, p_{θ^*}, $\ell(\theta, z) = -\log p_\theta(z)$. ERM is maximum likelihood.

3. Regression: $\theta \in \mathbb{R}^p$, $z = (x, y)$, $\ell(\theta, (x, y)) = |\theta' x - y|$. ERM chooses θ to minimize the average absolute error on the sample.

Glivenko-Cantelli Classes

Why are uniform laws of large numbers useful for empirical risk minimization?

We are interested in controlling the excess risk,

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

where θ^* minimizes L on Θ. We can decompose it as

$$L(\hat{\theta}) - L(\theta^*) = \left[L(\hat{\theta}) - L_n(\hat{\theta})\right] + \left[L_n(\hat{\theta}) - L_n(\theta^*)\right] + \left[L_n(\theta^*) - L(\theta^*)\right].$$

Glivenko-Cantelli Classes

One of these terms is a difference between a sample average and an expectation for the fixed function $\ell(\theta, \cdot)$:

$$L_n(\theta^*) - L(\theta^*) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta^*, Z_i) - P\ell(\theta^*, Z).$$

The law of large numbers shows that this term converges to zero; and with information about the tails of $\ell(\theta^*, Z)$ (such as boundedness), we can get bounds on its value.

Glivenko-Cantelli Classes

$L_n(\hat{\theta}) - L_n(\theta^*)$ is non-positive, because $\hat{\theta}$ is chosen to minimize L_n.

The other difference, $L(\hat{\theta}) - L_n(\hat{\theta})$, is more interesting. For any fixed θ, this difference goes to zero. But $\hat{\theta}$ is random, since it is chosen using the $X_1, \ldots, X_n$. An easy upper bound is

$$L(\hat{\theta}) - L_n(\hat{\theta}) \leq \sup_{\theta \in \Theta} |L(\theta) - L_n(\theta)| \,,$$

and this motivates the study of uniform laws of large numbers.

Proof of Glivenko-Cantelli Theorem

Theorem: $\|F_n - F\|_\infty \overset{as}{\to} 0$. That is, $\|P - P_n\|_G \overset{as}{\to} 0$, where $G = \{1[x \leq t] : t \in \mathbb{R}\}$.

We'll look at a proof that we'll then extend to a more general sufficient condition for a class to be Glivenko-Cantelli.

The proof involves three steps: A concentration inequality, symmetrization, which leads us to consider restrictions of step functions $g \in G$ to the data, and then exploiting the fact that the set of these restrictions is always simple.

Proof of Glivenko-Cantelli Theorem: Concentration

First, since $g(X_i) \in \{0, 1\}$, we have that the following function of the random variables $X_1, \ldots, X_n$ satisfies the bounded differences property with bound $1/n$:

$$\sup_{g \in G} |Pg - P_n g|$$

The bounded differences inequality implies that, with probability at least $1 - \exp(-2\epsilon^2 n)$,

$$\|P - P_n\|_G \leq \mathbf{E}\|P - P_n\|_G + \epsilon.$$

Proof of Glivenko-Cantelli Theorem: Symmetrization

Second, we symmetrize by replacing Pg by $P'_n g = \frac{1}{n}\sum_{i=1}^n g(X'_i)$. In particular, we have

$$\begin{aligned}
\mathbf{E}\|P - P_n\|_G &= \mathbf{E}\sup_{g\in G}\left|\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^n (g(X'_i) - g(X_i))\,\middle|\, X_1^n\right]\right| \\
&\le \mathbf{E}\mathbf{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^n (g(X'_i) - g(X_i))\right|\,\middle|\, X_1^n\right] \\
&= \mathbf{E}\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^n (g(X'_i) - g(X_i))\right| = \mathbf{E}\|P'_n - P_n\|_G.
\end{aligned}$$

Proof of Glivenko-Cantelli Theorem: Symmetrization

Now we symmetrize again: for any $\epsilon_i \in \{\pm 1\}$,

$$\mathbf{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} (g(X_i') - g(X_i)) \right| = \mathbf{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (g(X_i') - g(X_i)) \right|,$$

This follows from the fact that X_i and X_i' are i.i.d., and so the distribution of the supremum is unchanged when we swap them. And so in particular the expectation of the supremum is unchanged. And since this is true for any ϵ_i, we can take the expectation over any random choice of the ϵ_i. We'll pick them independently and uniformly.

Proof of Glivenko-Cantelli Theorem: Symmetrization

$$\begin{aligned}
&\mathbf{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (g(X_i') - g(X_i)) \right| \\
&\leq \mathbf{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(X_i') \right| + \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(X_i) \right| \\
&\leq \underbrace{2 \, \mathbf{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(X_i) \right|}_{\text{Rademacher complexity}} = 2\mathbf{E} \| R_n \|_G ,
\end{aligned}$$

where we've defined the **Rademacher process**
$R_n(g) = (1/n) \sum_{i=1}^{n} \epsilon_i g(X_i)$. (We'll finish the proof next lecture...)