CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 14**

Glivenko-Cantelli Classes

Lecturer: Peter Bartlett *Scribe: Michelle Bensi*

1 Introduction

This lecture will cover Glivenko-Cantelli (GC) classes and introduce Rademacher averages. We are interested in GC classes because, for these classes, we get uniform convergence of the empirical average to the true expectation. Rademacher averages provide a measure of complexity. In this lecture, the primary focus will be on introducing the GC classes of functions and proving the GC Theorem. It will end with the definition of Rademacher averages.

Recall from previous lectures that:

We can choose:

$$\hat{f} = \operatorname*{argmin}_{f \in F} \hat{R}(f)$$

And we want:

$$\hat{R}(f) \simeq \inf_{f \in F} R(f)$$

And it suffices to show:

$$\sup_{f \in F} |R(f) - \hat{R}(f)| \text{ is small}$$

For GC class functions this sufficient condition is satisfied as n gets large.

2 GC Classes

We begin with a definition of the GC class of functions.

Definition. F is a GC Class if, for all $\epsilon > 0$:

$$\lim_{n \to \infty} \sup_{P} P^n \left(\sup_{f \in F} \left| \mathbb{E}f - \hat{\mathbb{E}_n} f \right| > \epsilon \right) = 0$$

Note: P^n means n independent draws from a distribution.

2.1 The Glivenko-Cantelli Theorem

Let:

- $x_1, ..., x_n$ be i.i.d. data points from a distribution F.
- $F_n(x)$ be the empirical distribution function
- $F(t)$ be the true distribution function f^n of P

We have the following expressions for the CDF's:

$$F(t) = \mathbb{E}[1[x \leq t]]$$

$$F_n(t) = \mathbb{E}_n[1[x \leq t]]$$

Now define:

$$G = \{x \mapsto 1[x \leq 0] : \theta \in \mathbb{R}\}$$

That is, there is a one-to-one mapping between G and $\mathbb{R}$. Therefore:

$$\text{Glivenko-Cantelli Theorem} \iff \forall P, \sup_{g \in G} |\mathbb{E}g - \mathbb{E}_n g| \to 0$$

Thus, we can interpret this classical result as a result about uniform convergence over this class of subsets of the reals.

2.2 GC Theorem

We'll now formally present the GC Theorem, and give a proof that is suggestive of an approach that applies much more generally (which we'll meet in the next lecture).

Theorem 2.1. Define:

- $F_n(t) = P_n((-\infty, t])$ (the empirical distribution function)
- $F(t) = P((-\infty, t])$ (the true distribution function f^n of P)

For all probability distributions P on $\mathbb{R}$, $F_n \xrightarrow{\text{a.s.}} F$ uniformly on $\mathbb{R}$

Or, symbolically we can write:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0$$

Note: the law of large numbers ensures pointwise convergence of distribution functions, however, with the GC class of functions we obtain something stronger, namely uniform convergence.

The proof of the Glivenko-Cantelli Theorem involves three parts:

1. Use of the McDiarmid concentration equality
2. Use of symmetrization
3. Application of "simple" restrictions

Proof.

1. Through application of the McDiarmid concentration inequality we know that with probability at least $1 - \exp(-2\epsilon^2 n)$,

$$\sup_{g \in G} |\mathbb{E}_n g - \mathbb{E} g| \leq \mathbb{E}\left(\sup_{g \in G} |\mathbb{E}_n g - \mathbb{E} g|\right) + \epsilon$$

That is, the deviations are concentrated around their expectation.

2. Next we apply symmetrization.
Recall that we ultimately would like to prove:

$$\sup_{g \in g} |\mathbb{E}_n g - \mathbb{E} g| \xrightarrow{\text{a.s.}} 0$$

Also, note that we can write:

$$|\mathbb{E}_n g - \mathbb{E} g| = \left| \frac{1}{n} \sum_{i=1}^{n} g(x_i) - \mathbb{E} \frac{1}{n} \sum_{i=1}^{n} g(x_i) \right|$$

Let: $x'_1, ..., x'_n$ be i.i.d. copies of $x_1, ..., x_n$.

$$\begin{aligned}
\sup_{g \in g} |\mathbb{E}_n g - \mathbb{E} g| &= \mathbb{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} g(x_i) - \mathbb{E} g \right| \quad \text{(expanding on definition of } \mathbb{E}_n\text{)} \\
&= \mathbb{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} (g(x_i) - \mathbb{E} g(x'_i)) \right| \\
&= \mathbb{E} \sup_{g \in G} \left| \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} (g(x_i) - g(x'_i)) | x_1, ..., x_n \right] \right| \quad \text{(properties of conditional expectation)} \\
&\leq \mathbb{E}\mathbb{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} (g(x_i) - g(x'_i)) \right| \quad \text{(bringing the } \mathbb{E} \text{ out front)} \\
&= \mathbb{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (g(x_i) - g(x'_i)) \right|
\end{aligned}$$

Where ϵ_i is a Rademacher variable (uniform on $\{\pm 1\}$). So we have the following upperbound on the previous expression:

$$\leq \mathbb{E} \sup_{g \in G} \left(\left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \, g(x_i) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(x'_i) \right| \right) \leq 2 \underbrace{\mathbb{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(x_i) \right|}_{\text{Rademacher averages of G}}$$

3. Next, we consider "simple" restrictions on G.
We can write:

$$2\,\mathbb{E} \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(x_i) \right| = 2\,\mathbb{E}\,\mathbb{E}\left[\sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \, g(x_i) \right| | x_1, ..., x_n \right]$$

But,$|\{(g(x_1),...,g(x_n)) : g \in G\}| = |\{(g(x_{(1)}),...,g(x_{(n)})) : g \in G\}| \leq n+1$
Where we have ordered the data: $\{x_1,...x_n\} = \{x_{(1)},...,x_{(n)}\}$ and $x_{(1)} \leq ... \leq x_{(n)}$

Next we apply the follow lemma to seek the bound of the expression from above:

$$2\ \mathbb{E}\mathbb{E}\ \left[\sup_{g\in G}\left|\sum_{i=1}^{n}\epsilon_i g(x_i)\right| | x_1,...,x_n\right]$$

□

Lemma 2.2. For $A \subseteq \mathbb{R}^n$ with $R = \max_{a\in A}(\sum_i a_i^2)^{\frac{1}{2}}$, we have:

$$\mathbb{E}\sup_{a\in A}\underbrace{\left(\sum_{i=1}^{n}\epsilon_i a_i\right)}_{Z_a} \leq R\sqrt{2\log|A|}$$

Proof.

$$\begin{aligned}
\exp[\underbrace{s}_{s>0}\mathbb{E}\sup_{a\in A}Z_a] &\leq \mathbb{E}\ \exp\left[s\sup_{a\in A}Z_a\right] \quad \text{(because exponential function is convex)}\\
&= \mathbb{E}\sup_{a\in A}(\exp(sZ_a))\\
&\leq \sum_{a\in A}\mathbb{E}\ \exp[sZ_a]\\
&\leq \sum_{a\in A}\exp\left(\frac{s^2}{2}\underbrace{\sum_{i=1}^{n}a_i^2}_{\leq R^2}\right) \quad \text{(by Hoeffding's inequality)}\\
&\leq |A|\exp(s^2\frac{R^2}{s})
\end{aligned}$$

So,

$$\mathbb{E}\ \sup_{a\in A}\left(\sum_{i=1}^{n}\epsilon_i a_i\right) \leq \inf_{s>0}\left(\frac{\log|A|}{s}+\frac{sR^2}{2}\right) = R\sqrt{2\log|A|}$$

□

Note: For our application, $R \leq \frac{1}{\sqrt{n}}$ and $|A| \leq n+1$
Hence:

$$Pr\left(\sup_{g\in G}|\hat{\mathbb{E}}(g)-\mathbb{E}g|\right) \geq \epsilon + 2\sqrt{\frac{2\log(n+1)}{n}} \leq \exp(-2\epsilon^2 n),$$

which completes the proof.

3 Rademacher averages

Definition. For a class F of real-valued functions defined on X, for i.i.d. $x_1,...,x_n \in X$, and for independent Rademacher random variables $\epsilon_1,...,\epsilon_n$, define:

- $R_n(F) = \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i)$ (Rademacher averages on F)
- $\hat{R}_n(F) = \mathbb{E}[\sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) | x_1, ..., x_n]$ (empirical Rademacher averages)