Theoretical Statistics. Lecture 9.

Peter Bartlett

Uniform laws of large numbers.

1. Proof of Glivenko-Cantelli Theorem
2. Glivenko-Cantelli classes
3. Bounding Rademacher complexity.

Recall: Glivenko-Cantelli Theorem

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

Uniform law of large numbers because it's uniform over G.

Recall: Proof of Glivenko-Cantelli Theorem

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:

1. Concentration: with probability at least $1 - \exp(-2\epsilon^2 n)$,
$$\|P - P_n\|_G \leq \mathbf{E}\|P - P_n\|_G + \epsilon.$$
2. Symmetrization: $\mathbf{E}\|P - P_n\|_G \leq 2\mathbf{E}\|R_n\|_G$, where we've defined the **Rademacher process** $R_n(g) = (1/n)\sum_{i=1}^n \epsilon_i g(X_i)$ (and this leads us to consider restrictions of step functions $g \in G$ to the data), and
3. Simple restrictions.

Proof of Glivenko-Cantelli Theorem: Restrictions

We consider the set of restrictions
$G(X_1^n) = \{(g(X_1), \ldots, g(X_n)) : g \in G\}$:

$$2\mathbf{E}\|R_n\|_G = 2\mathbf{E}\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i)\right| = 2\mathbf{E}\mathbf{E}\left[\sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i)\right|\middle| X_1^n\right].$$

But notice that the cardinality of $G(X_1^n)$ does not change if we order the data. That is,

$$\begin{aligned}|G((X_1,\ldots,X_n))| &= \left|G((X_{(1)},\ldots,X_{(n)}))\right| \\ &= \left|\left\{(1[X_{(1)} \le t],\ldots,1[X_{(n)} \le t]) : t \in \mathbb{R}\right\}\right| \le n+1,\end{aligned}$$

where $X_{(1)} \le \cdots \le X_{(n)}$ is the data in sorted order (and so $X_{(i)} \le t$ implies $X_{(i-1)} \le t$).

Proof of Glivenko-Cantelli Theorem: Rademacher Averages

Finally, we use the following result.

Lemma: [Finite Classes] For $A \subseteq \mathbb{R}^n$ with $R = \max_{a \in A} \|a\|_2$,

$$\mathbf{E} \sup_{a \in A} \langle \epsilon, a \rangle \leq \sqrt{2R^2 \log |A|}.$$

Hence

$$\mathbf{E} \sup_{a \in A} |\langle \epsilon, a \rangle| = \mathbf{E} \sup_{a \in A \cup -A} \langle \epsilon, a \rangle \leq \sqrt{2R^2 \log(2|A|)}.$$

Proof of Rademacher Averages Result

Proof:

$$\begin{aligned}
\exp\left(\lambda \mathbf{E} \sup_a \langle \epsilon, a\rangle\right) &\leq \mathbf{E} \exp\left(\lambda \sup_a \langle \epsilon, a\rangle\right) \\
&= \mathbf{E} \sup_a \exp\left(\lambda \langle \epsilon, a\rangle\right) \\
&\leq \sum_a \mathbf{E} \exp\left(\lambda \langle \epsilon, a\rangle\right) \\
&\leq \sum_a \exp\left(\frac{\lambda^2}{2} \|a\|_2^2\right) \\
&\leq |A| \exp\left(\frac{\lambda^2 R^2}{2}\right),
\end{aligned}$$

using the fact that $\epsilon_i a_i$ is bounded, hence sub-gaussian. Picking $\lambda^2 = 2 \log |A| / R^2$ gives the result.

Proof of Glivenko-Cantelli Theorem

For the class G of step functions, $R \leq 1/\sqrt{n}$ and $|A| \leq n+1$. Thus, with probability at least $1 - \exp(-2\epsilon^2 n)$,

$$\|P - P_n\|_G \leq \sqrt{\frac{8\log(2(n+1))}{n}} + \epsilon.$$

By Borel-Cantelli, $\|P - P_n\|_G \stackrel{as}{\to} 0$.

Recall: Glivenko-Cantelli Classes

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

$$\|P_n - P\|_F \xrightarrow{P} 0.$$

Recall: Glivenko-Cantelli Classes and ERM

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

We are interested in controlling the excess risk,

$$P\ell_{\hat{\theta}} - \inf_{\theta \in \Theta} P\ell_\theta = P\ell_{\hat{\theta}} - P\ell_{\theta^*}$$

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

$$P\ell_{\hat{\theta}} - P\ell_{\theta^*} = \left[P\ell_{\hat{\theta}} - P_n\ell_{\hat{\theta}}\right] + \left[P_n\ell_{\hat{\theta}} - P_n\ell_{\theta^*}\right] + \left[P_n\ell_{\theta^*} - P\ell_{\theta^*}\right].$$

The last difference is controlled by a LNN, the second is non-positive by the definition of $\hat{\theta}$, and the first term is controlled via a ULNN: $P\ell_{\hat{\theta}} - P_n\ell_{\hat{\theta}} \leq \sup_\theta |P\ell_\theta - P_n\ell_\theta|$.

Recall: Glivenko-Cantelli Classes and ERM

Note that the inequality $P\ell_{\hat{\theta}} - P_n\ell_{\hat{\theta}} \leq \sup_\theta |P\ell_\theta - P_n\ell_\theta|$ might be loose. But there are important examples where it is tight enough to give optimal rates (such as two-class classification and regression with absolute loss, in minimax settings, that is, with a worst-case choice of probability distribution).

Uniform laws and Rademacher complexity

The proof of the Glivenko-Cantelli Theorem involved three steps:

1. Concentration of $\|P - P_n\|_F$ about its expectation.
2. Symmetrization, which bounds $\mathbf{E}\|P - P_n\|_F$ in terms of the Rademacher complexity of F, $\mathbf{E}\|R_n\|_F$.
3. A combinatorial argument showing that the set of restrictions of F to X_1^n is small, and a bound on the **Rademacher complexity** using this fact.

We'll follow a similar path to prove a more general uniform law of large numbers.

Uniform laws and Rademacher complexity

Definition: The **Rademacher complexity** of F is $\mathbf{E}\|R_n\|_F$, where the empirical process R_n is defined as

$$R_n(f) = \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right|,$$

where the $\epsilon_1, \ldots, \epsilon_n$ are Rademacher random variables: i.i.d. uniform on $\{\pm 1\}$.

Note that this is the expected supremum of the alignment between the random $\{\pm 1\}$-vector ϵ and $F(X_1^n)$, the set of n-vectors obtained by restricting F to the sample $X_1, \ldots, X_n$.

Uniform laws and Rademacher complexity

Theorem: For any F, $\mathbf{E}\|P - P_n\|_F \leq 2\mathbf{E}\|R_n\|_F$.
If $F \subset [0,1]^{\mathcal{X}}$,

$$\frac{1}{2}\mathbf{E}\|R_n\|_F - \sqrt{\frac{\log 2}{2n}} \leq \mathbf{E}\|P - P_n\|_F \leq 2\mathbf{E}\|R_n\|_F,$$

and, with probability at least $1 - 2\exp(-2\epsilon^2 n)$,

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

Thus, $\mathbf{E}\|R_n\|_F \to 0$ iff $\|P - P_n\|_F \stackrel{as}{\to} 0$.

That is, the sup of the empirical process $P - P_n$ is concentrated about its expectation, and its expectation is about the same as the expected sup of the Rademacher process R_n.

Uniform laws and Rademacher complexity

The first result is the symmetrization that we saw earlier:

$$\begin{aligned}\mathbf{E}\|P - P_n\|_F &\leq \mathbf{E}\|P'_n - P_n\|_F \\ &= \mathbf{E}\left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(X'_i) - f(X_i)) \right\|_F \\ &\leq 2\mathbf{E}\|R_n\|_F.\end{aligned}$$

where R_n is the Rademacher process $R_n(f) = (1/n) \sum_{i=1}^{n} \epsilon_i f(X_i)$.

Uniform laws and Rademacher complexity

The second inequality (*desymmetrization*) follows from:

$$\begin{aligned}
\mathbf{E}\|R_n\|_F &\leq \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(X_i) - \mathbf{E}f(X_i)\right)\right\|_F + \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{E}f(X_i)\right\|_F \\
&\leq \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(X_i) - f(X_i')\right)\right\|_F + \|P\|_F\,\mathbf{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \\
&= \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - \mathbf{E}f(X_i) + \mathbf{E}f(X_i') - f(X_i')\right)\right\|_F \\
&\qquad + \|P\|_F\,\mathbf{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \\
&\leq 2\mathbf{E}\,\|P_n - P\|_F + \sqrt{\frac{2\log 2}{n}}.
\end{aligned}$$

Uniform laws and Rademacher complexity

And this shows that $\|P - P_n\|_F \stackrel{as}{\to} 0$ implies $\mathbf{E}\|R_n\|_F \to 0$.

The last inequality follows from the triangle inequality and the Finite Classes Lemma.

And Borel-Cantelli implies that $\mathbf{E}\|R_n\|_F \to 0$ implies $\|P - P_n\|_F \stackrel{as}{\to} 0$.

Controlling Rademacher complexity

So how do we control $\mathbf{E}\|R_n\|_F$? We'll look at several approaches:

1. $|F(X_1^n)|$ small. ($\max |F(x_1^n)|$ is the **growth function**)
2. For binary-valued functions: Vapnik-Chervonenkis dimension. Bounds rate of growth function. Can be bounded for parameterized families.
3. Structural results on Rademacher complexity: Obtaining bounds for function classes constructed from other function classes.
4. Covering numbers. Dudley entropy integral, Sudakov lower bound.
5. For real-valued functions: scale-sensitive dimensions.

Controlling Rademacher complexity: Growth function

For the class of distribution functions, $G = \{x \mapsto 1[x \leq \alpha] : \alpha \in \mathbb{R}\}$, we saw that the set of restrictions,

$$G(x_1^n) = \{(g(x_1), \ldots, g(x_n)) : g \in G\}$$

is always small: $|G(x_1^n)| \leq \Pi_G(n) = n + 1$.

Definition: For a class $F \subseteq \{0,1\}^{\mathcal{X}}$, the **growth function** is

$$\Pi_F(n) = \max\{|F(x_1^n)| : x_1, \ldots, x_n \in \mathcal{X}\}.$$