1 Outside references

This material is contained in M. Mohri's lecture notes (lecture 3)[1], their book [2], and the excellent and comprehensive notes by Boucheron et al. [1].

2 Universal Glivenko-Cantelli classes

Recall from last lecture the definition of a UGC class. Let $(\mathcal{X}, P)$ be a probability space and $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$ be a family of binary-valued functions on $\mathcal{X}$. We say that $\mathcal{F}$ is a Universal Uniform Glivenko-Cantelli (UUGC) class if

$$\forall \varepsilon > 0 : \lim_{n\to\infty} \sup_P \mathbb{P}\left[\sup_{f\in\mathcal{F}} |Pf - P_n f| > \varepsilon\right] = 0,$$

where $\sup_P$ is the supremum over all the distributions on $\mathcal{X}$ and

$$P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \qquad Pf = \mathbb{E} f(X_1)$$

for $X_1, \ldots, X_n$ distributed iid $\sim P$.

3 Rademacher complexity

The UGC condition says that the *maximal deviation* $\sup_{f\in\mathcal{F}} |Pf - P_n f|$ —which is a random variable— converges to zero in probability at a rate not depending on the distribution. To show this, it will suffice to establish that

$$\mathbb{E} \sup_{f\in\mathcal{F}} (Pf - P_n f) \underset{n\to\infty}{\longrightarrow} 0$$

at a distribution-free rate. Our first step was *symmetrization by ghost sample*, whose effect was to make the two terms look more alike. If $X = (X_1, \ldots, X_n)$ is the original sample and $X' = (X'_1, \ldots, X'_n)$ is another sequence of n points drawn iid from the same distribution (independently of anything else), then Jensen's inequality implies that

$$\mathbb{E} \sup_{f\in\mathcal{F}} (Pf - P_n f) \quad \leq \quad \mathbb{E}_X \mathbb{E}_{X'} \sup_{f\in\mathcal{F}} (P'_n f - P_n f),$$

[1] `http://www.cs.nyu.edu/~mohri/mls/lecture_3.pdf`

where $P'_n f = \frac{1}{n}\sum_{i=1}^n f(X'_i)$. [Note that this would also hold for $(P_n f - Pf)$.]

The next step was to *symmetrize by Rademacher sequence*. The latter is $\sigma = (\sigma_1, \ldots, \sigma_n)$, drawn uniformly from $\{-1, 1\}^n$. Convince yourself that

$$\begin{aligned}
\mathbb{E}_X \mathbb{E}_{X'} \sup_{f\in\mathcal{F}} (P'_n f - P_n f) &= \mathbb{E}_X \mathbb{E}_{X'} \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i (f(X_i) - f(X'_i)) \\
&\leq 2\mathbb{E}_{X,\sigma} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i) \\
&=: 2\mathcal{R}_n(\mathcal{F})
\end{aligned}$$

where the last line defines the Rademacher average of $\mathcal{F}$.

4 Rademacher complexity for finite $A \subset \mathbb{R}^n$

We proved the following basic result: if $Y_1, \ldots, Y_N$ are independent random variables with *subgaussian moment* bounded by some $\gamma > 0$,

$$\mathbb{E}e^{tY_i} \leq e^{t^2\gamma^2/2}, \qquad \forall i \in [N], \forall t \geq 0,$$

then:

$$\begin{aligned}
\mathbb{E}\max_{i\in[N]} Y_i &\leq \gamma\sqrt{2\log N}, \\
\mathbb{E}\max_{i\in[N]} |Y_i| &\leq \gamma\sqrt{2\log(2N)}.
\end{aligned}$$

Now in fact Rademacher complexity $\mathcal{R}_n$ is defined for any $A \subset \mathbb{R}^n$:

$$\mathcal{R}_n(A) = \mathbb{E}_\sigma \sup_{a\in A} \frac{1}{n}\sum_{i=1}^n \sigma_i a_i.$$

Recall Hoeffding's lemma: if $\mathbb{E}X = 0$ and $a \leq X \leq b$, then X is subgaussian with $\gamma = (b-a)/2$

$$\mathbb{E}e^{tX} \leq e^{t^2(b-a)^2/8}.$$

This implies that for any fixed $a \in A$, the random variable $X = \sum_{i=1}^n \sigma_i a_i$ has subgaussian moment at most $\|a\| = \sqrt{\sum_{i=1}^n a_i^2}$. It follows that for finite A,

$$\mathcal{R}_n(A) \leq \frac{1}{n}(\max_{a\in A} \|a\|)\sqrt{2\log|A|}.$$

5 Projecting $\mathcal{F}$ on a sample

The next observation is the following simple but profound insight:

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i) \quad = \quad \sup_{f\in\mathcal{F}(X_1,\ldots,X_n)} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f_i,$$

where

$$\mathcal{F}(X_1,\ldots,X_n) = \{(f(X_1), f(X_2),\ldots,f(X_n)) : f\in\mathcal{F}\} \subseteq \{0,1\}^n$$

is the *projection* of $\mathcal{F}$ on (also *restriction* of $\mathcal{F}$ to) the sample $X_1,\ldots,X_n$. Since $\|f\| \le \sqrt{n}$ for all $f \in \mathcal{F}(X_1,\ldots,X_n)$, we have

$$\mathbb{E} \sup_{f\in\mathcal{F}(X_1,\ldots,X_n)} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f_i \le \sqrt{\frac{2\log|\mathcal{F}(X_1,\ldots,X_n)|}{n}}.$$

6 Bounding $|\mathcal{F}(X_1,\ldots,X_n)|$ via the VC-dimension

We say that $\mathcal{F}$ *shatters* a set $S \subset \mathcal{X}$ if $|\mathcal{F}(S)| = 2^{|S|}$. The VC-dimension of $\mathcal{F}$ is the size of the largest set shattered by $\mathcal{F}$ (or ∞ if $\mathcal{F}$ shatters sets of arbitrary size). The celebrated Perles-Sauer-Shelah-Vapnik-Chervnonenkis lemma (which we'll prove next time) states that when VC-dim of $\mathcal{F}$ is d, we have

$$|\mathcal{F}(X_1,\ldots,X_n)| \quad \le \quad \Big(\frac{en}{d}\Big)^d, \qquad n \ge d.$$

We conclude:

$$\mathcal{R}_n(\mathcal{F}) \le \sqrt{\frac{2d\log(en/d)}{n}}.$$

In other words, function classes $\mathcal{F}$ with finite VC-dimension are UUGC.

References

[1] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[2] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations Of Machine Learning*. The MIT Press, 2012.