36-755: Advanced Statistics **Fall 2016**

Lecture 16: September 7

Lecturer: Alessandro Rinaldo *Scribes: Kayla Frisoli*

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

16.1 Recap

Last time we looked at the combinatorial properties of the VC-dimension. We looked at some classes of subsets in $\mathbb{R}^d$ that have finite VC-dimension. There are many other classes that don't have finite VC-dimension.

Examples:

- In $\mathbb{R}^2$ the VC-dimension of **polygons** is ∞
- The VC-dimension of **convex sets** are infinite (shown in homework 5)
- In $\mathbb{R}^d$ the VC-dimension of a **polytope** (high dimensional version of a polygon) with an unbounded number of facets is also infinite
  - Any subset of points can be picked out by a polytope

Note: our definition of VC-dimension matches that of the book *A Probabilistic Theory of Pattern Recognition*, but there are other places it is defined differently. We define it to be the largest n such that $\mathcal{S}_A(n) = 2^n$, but it can also be defined as the smallest n such that $\mathcal{S}_A(n) < 2^n$. $\mathcal{S}_A(n) = 2^n$ means that every possible combination (2^n) of points can be picked out. In our definition of VC-dimension, we want the largest n such that all combinations of $X_1, \cdots, X_n$ can be picked out.

The derivation of the VC (Vapnik Chervnonkis) inequality in the book is non-traditional. Other references use the traditional derivation.

16.2 VC dimension continued

16.2.1 Traditional derivation

If we have a class of subsets in R^d, then

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \geq \epsilon\right) \leq c_1 \mathcal{S}_A(n) e^{-c_2 n\epsilon^2} \qquad \text{for } c_1, c_2 > 0$$

$\mathcal{S}_A(n) = \max_{x_i^n} |A(x_i^n)| \to \{x_1^n \cap A, A \in \mathcal{A}\}$

Proof:

Step 1: Symmetrization by ghost sample

Assume $n\epsilon^2 \geq 2$

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 2\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P_n'(A)| > \epsilon/2\right)$$

Where P_n' is the empirical distribution based on the ghost sample (we never get to see) $(X_1', \cdots, X_n')$ which are iid and independent of $(X_1, \cdots, X_n)$

Step 2: Symmetrization by random signs

Note: because we are dealing with the Rademacher complexity, we have randomness in X and ϵ. Using Rademacher is useful because we can condition on the X and make them independent of the rademacher complexity.

$$\leq 2\mathbb{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n}|\sum_{i=1}^n \epsilon_i \mathbb{1}(A_i \in \mathcal{A})| > \epsilon/4\right) \qquad \epsilon_1 \cdots \epsilon_n \overset{iid}{\sim} \text{Rademacher}, \perp\!\!\!\perp x_1 \cdots x_n \perp\!\!\!\perp x_1' \cdots x_n'$$

$$= 2\mathbb{E}_x\mathbb{P}_{\epsilon,x}\left(\sup_{A\in\mathcal{A}} \frac{1}{n}|\sum_{i=1}^n \epsilon_i f_A(x)| > \epsilon/4 \,\middle|\, x_1 \cdots x_n\right) \qquad \text{rewrite, utilize usefulness of Rademacher}$$

Now, we fix $(x_i \cdots x_n) = x_1^n$, where $f_A = \mathbb{1}(x_i \in \mathcal{A})$. Because we have the supremum in there still, we can't use hoeffding. We need just one A to use hoeffding. But, we can bound.

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} \frac{1}{n}|\sum_{i=1}^n \epsilon_i f_A(x)| > \epsilon/4 \,\middle|\, X_1^n = x_1^n)\right)$$
$$\leq \mathcal{S}_A(n) \sup_{A\in\mathcal{A}} \underbrace{\mathbb{P}_\epsilon\left(\frac{1}{n}|\sum_{i=1}^n \epsilon_i f_A(x) - \mathbb{E}_\epsilon[\epsilon_i f(x_i)]| > \epsilon/4\right)}_{(*)}$$

Step 3: Use hoeffding inequality to bound the above probability, (*)

The terms are uniformly bounded so we can use hoeffding.

$$\lesssim 8\mathcal{S}_A(n)e^{-n\epsilon^2/32} \qquad \mathcal{S}_A(n) \leq (n+1)^v \text{ if A has VC-dim v}$$

■

16.2.2 Sharpening of the VC-inequality

16.2.2.1 Inequality for relative deviation

As we just saw, the VC inequality depends on the hoeffding inequality. We know hoeffding isn't always the sharpest. So, going from hoeffding to Bernstein is a way to improve the inequality when the variances are small.

We might want to distinguish between when P(A) is small or large. For the binomial case, if the probabilities are very small or very large, hoeffding is bad. Currently, the VC inequality will not give separate behavior for different values of P(A). So, we divide by $\sqrt{P_n(A)}$ and $\sqrt{P_(A)}$.

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}}|\frac{P_n(A)-P(A)}{\sqrt{P_n(A)}}|\geq\epsilon\right)\leq 4\mathcal{S}_A(2n)e^{-n\epsilon^2/4} \iff$$
$$\mathbb{P}\left(\sup_{A\in\mathcal{A}}|\frac{P(A)-P_n(A)}{\sqrt{P(A)}}|\geq\epsilon\right)\leq 4\mathcal{S}_A(2n)e^{-n\epsilon^2/4}$$

Then, $\forall t\in(0,1]$,

$$\mathbb{P}\left(\exists A\in\mathcal{A}\text{ st }P(A)\geq t,\mathbb{P}P_n(A)\leq(1-t)P(A)\right)\leq 4\mathcal{S}_A(2n)e^{-n\epsilon^2/4}$$

An important conservancy of relative deviation is

$$P(A)\leq P_n(A)+2\sqrt{\frac{P_n(A)\log\mathcal{S}_A(2n)+\log(4/\delta)}{n}}+\frac{4\log\mathcal{S}_A(2n)+\log(4/\delta)}{n}$$

Simultaneously over all $A\in\mathcal{A}$ and for any $\delta\in(0,1)$ this follows from the following inequality

$$A\leq B+C\sqrt{A}\implies A\leq B+C^2+\sqrt{B}C\text{ for }A,B,C>0$$

We think about the relative deviation as a better VC inequality, when the probabilities are very small.

We want to extend this to functions, because currently it only applies to binary f. What we've done so far is good for probability, but we want to make our findings broader. We have VC for classes of sets only, or, equivalently, for classes of functions that are binary.

So, if we let $\mathcal{F}$ be a collection of real valued functions from $\mathbb{R}^d$ into $[0,1]$ and we set $Pf=\mathbb{E}[f(x)]$ and $P_nf=\frac{1}{n}\sum f(x_i)$. Note that we can replace $[0,1]$ with any bounded set.

Proposition: $\sup_{f\in\mathcal{F}}|P_nf-Pf|=||P_n-P||_{\mathcal{F}}\leq\sup_{\substack{f\in\mathcal{F}\\t\in[0,1]}}\frac{1}{n}\sum_{i=1}^n[\mathbb{1}(f(x_i)>t)-\mathbb{P}(f(x_i)>t)]$

Proof:

We use the fact that if $x\geq 0$, $x=\int_0^\infty\mathbb{1}(x\geq t)dt$

For any f treat $\sup_{f\in\mathcal{F}}|P_nf-Pf|=||P_n-P||_{\mathcal{F}}$ as a random variable.

If $Z\geq 0,\mathbb{E}[Z]=\int_0^\infty\mathbb{1}(x\geq\eta)d\eta$

Then, let $\mathcal{A}=\{\{x:f(x)>t\},f\in\mathbb{F},t\in[0,1]\}$ so that $||P_n-P||_{\mathcal{F}}\leq\sup_{A\in\mathcal{A}}|P_n(A)-P(A)|$

We can use VC theory if $\mathcal{A}$ has a finite VC-dimension and therefore we can say that $\mathcal{F}$ is a VC-class.

■

In reality and practice, this isn't the best tool. It's hard to find the VC dimension of function classes. There are more manageable tools when dealing with function classes. In dealing with function classes, we need to use covering numbers and metric entropies. Reminder: the δ-covering number of a metric space is the smallest number of points such that balls centered there will cover the space.

On the space $\mathcal{F}$ use this metric: $d_{1,p_n}(f,g) = \frac{1}{n}\sum_i^n |f(x_i) - g(x_i)|$ where $x_1 \cdots x_n \overset{iid}{\sim} P$ random metric. This is a random metric because it depends on our sample.

Let $N(\mathcal{F}, \frac{\epsilon}{\delta}, P_{2n}) \to \frac{\epsilon}{\delta}$ covering number of $\mathcal{F}$ with respect to d_1, P_{2n}. Here, N is a random covering.

Then,

$$\mathbb{P}(||P_n - P||_{\mathcal{F}} > \epsilon) \leq \mathbb{E}[N(\mathcal{F}, \frac{\epsilon}{\delta}, d_1, 2n)]e^{-n\epsilon^2/32} \qquad \text{for } n \geq \epsilon^2/2$$

Instead of using d_{1,p_n} we can use $d_\infty(f,g) = \sup_x |f(x) - g(x)|$ because $d_\infty(f,g)$ is often easier to compute.

Then we replace the random covering with an appropriate deterministic (not random) L_∞ covering of $\mathcal{F}$. This is a weak result, but it is very useful. You will prove it in the homework (it is very simple, just use Hoeffding).

16.3 Talagrand inequality for empirical processes

What is the sharpest concentration inequality that man has ever known? The talagrand inequality. The tools to prove this, even the simplest proof, is still very complicated.

Given a class of functions on $\mathbb{R}^d$ into $[0,1]$ (or taking value into a bounded set) then

$$\mathbb{P}\left(||P_n - P||_{\mathcal{F}} \leq \mathbb{E}[||P_n - P||_F] - \sqrt{\frac{2t}{n} + 2\mathbb{E}[||P_n - P||_F] - \frac{t}{n} + \sigma^2(F)}\right) \leq e^{-t}$$

Where $\sigma^2(\mathcal{F}) = \sup_f V[f(x)]$. We need to bound $\mathbb{E}[||P_n - P||_F]$.

Remark: Since we are taking the sup over infinite classes, we are assuming $[||P_n - P||_F]$ is measurable. In general, we don't know this.

The fact that we assume separability is not a stretch. We say set $\mathcal{F}$ is separable if there exists a dense subset $\mathcal{F}_0 \subset \mathcal{F}$ such that $\forall f \in \mathcal{F}, \exists f_n \subset \mathcal{F}_0$ such that $\lim_n f_n(x) = f(x)\ \forall x$. Even the French papers avoid measurability issues by assuming a countable, dense subset.

Main idea: assume a countable, dense subset.

Useful links

`http://www.szit.bme.hu/~gyorfi/pbook.pdf`
`https://en.wikipedia.org/wiki/Polytope`