36-705: Intermediate Statistics **Fall 2017**

Lecture 9: September 18

Lecturer: Siva Balakrishnan

Perhaps the most compelling motivation for studying uniform convergence is to understand a procedure known as empirical risk minimization. Estimators of this type include maximum likelihood estimators, and many estimators we encounter in machine learning (SVMs, Boosting and so on). We will study this in detail in the next lecture.

Binary Classification: In the typical binary classification setting we observe a training set $\{(X_1, y_1), \ldots, (X_n, y_n)\}$ that we assume are drawn i.i.d from some distribution P. Each $X_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$.

A classifier $f : \mathbb{R}^d \mapsto \{-1, +1\}$ is simply a function that takes an instance (a vector in $\mathbb{R}^d$) and outputs a label.

The broad goal of classification is to try to find a function that has low error on future unseen data, i.e. we want a function that has low mis-classification error: $\mathbb{P}(f(X) \neq y)$.

For a given classifier f we can estimate its mis-classification error (risk) as:

$$\widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(f(X_i) \neq y_i),$$

which is simply its error on the training set. If f is some fixed classifier we know by Hoeffding's bound (why?) that,

$$\mathbb{P}(|\widehat{R}_n(f) - \mathbb{P}(f(X) \neq y)| \geq t) \leq 2 \exp(-2nt^2).$$

If we are trying to pick a good classifier from some set of classifiers $\mathcal{F}$, then a natural way to do this is to find the one that looks best on the training set, i.e. to choose

$$\widehat{f} = \arg \min_{f \in \mathcal{F}} \widehat{R}_n(f).$$

This procedure is known as *empirical risk minimization.* The terminology will be clearer later on in the course. For now though, we would like to understand this procedure better. How do we argue that in some cases this procedure will indeed select a good classifier? This question is intricately tied to uniform convergence.

Let f^* be the best classifier in $\mathcal{F}$. We would like to bound the excess risk of the classifier we chose, i.e.

$$\Delta = \mathbb{P}(\widehat{f}(X) \neq y) - \mathbb{P}(f^*(X) \neq y).$$

The typical way to do this is to consider the decomposition:

$$\Delta = \underbrace{\mathbb{P}(\widehat{f}(X) \neq y) - \widehat{R}_n(\widehat{f})}_{T_1} + \underbrace{\widehat{R}_n(\widehat{f}) - \widehat{R}_n(f^*)}_{T_2} + \underbrace{\widehat{R}_n(f^*) - \mathbb{P}(f^*(X) \neq y)}_{T_3}.$$

Since $\widehat{f}$ minimizes the empirical risk we know that $T_2 \leq 0$. We know that T_3 is small just by the Hoeffding argument from before, since f^* is a fixed classifier (i.e. does not depend on the training data).

The key point, one that you should really think carefully about is that we cannot use Hoeffding for the first term. The reason is that the classifier $\widehat{f}$ is data dependent so its empirical risk is not the sum of independent RVs.

Instead we have to rely on a *uniform* convergence bound, i.e. suppose we can show that with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left[\widehat{R}_n(f) - \widehat{R}_n(f^*)\right] \leq \Theta,$$

then we can conclude that the excess risk with probability at least $1 - \delta$ satisfies

$$\Delta = \mathbb{P}(\widehat{f}(X) \neq y) - \mathbb{P}(f^*(X) \neq y) \leq 2\Theta.$$

Everything boils down to showing uniform convergence of the empirical risk to the true error over the collection of classifiers we are interested in.

One way of bounding Θ is using the VC-dimension of the collection of classifiers. For now let us forget about classifiers, and return to the formulation of the uniform convergence question we discussed earlier. Given a collection of sets $\mathcal{A}$, we would like to understand the uniform convergence of empirical frequencies to probabilities, i.e. we want to bound:

$$\begin{aligned}\Delta(\mathcal{A}) &= \sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)| \\ &= \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \in A) - \mathbb{P}(A) \right|,\end{aligned}$$

where $X_1, \ldots, X_n$ are an i.i.d sample from some distribution P.

9.1 Warm up - Finite Collections

The first case to consider is when the collection of sets $\mathcal{A}$ has finite cardinality $|\mathcal{A}|$. In this case, for a fixed A we know by Hoeffding's inequality that,

$$\mathbb{P}(|\mathbb{P}_n(A) - \mathbb{P}(A)| \geq t) \leq 2\exp(-2nt^2).$$

However, we want something stronger we want that this convergence happens uniformly for all sets in $\mathcal{A}$, so we can use the union bound, i.e.

$$\begin{aligned}\mathbb{P}(\Delta(\mathcal{A}) \geq t) &= \mathbb{P}(\cup_{A \in \mathcal{A}}(|\mathbb{P}_n(A) - \mathbb{P}(A)| \geq t)) \\ &\leq \sum_{A \in \mathcal{A}} \mathbb{P}(|\mathbb{P}_n(A) - \mathbb{P}(A)| \geq t)) \\ &\leq 2|\mathcal{A}| \exp(-2nt^2).\end{aligned}$$

So if we want that with probability $1 - \delta$ the deviation be smaller than t we need to choose

$$t \geq \sqrt{\frac{\ln(2|\mathcal{A}|/\delta)}{2n}}.$$

In other words we have that with probability at least $1 - \delta$,

$$\Delta(\mathcal{A}) \leq \sqrt{\frac{\ln(2|\mathcal{A}|/\delta)}{2n}}.$$

This is already quite a nice result and once again highlights one of the main reasons why Hoeffding type exponential concentration inequalities are much more useful than Chebyshev type concentration inequalities: to obtain uniform convergence over $\mathcal{A}$ we pay a price which is logarithmic in the size of the collection.

9.2 VC dimension

Often we are interested in controlling $\Delta(\mathcal{A})$ for infinite classes of sets. The example from last lecture of uniform convergence of the empirical CDF is a canonical example.

In order to define the VC dimension, and understand the associated uniform convergence we need to understand the concept of shattering.

Shattering: Let $\{z_1, \ldots, z_n\}$ be a finite set of n points. We let $N_{\mathcal{A}}(z_1, \ldots, z_n)$ be the number of distinct sets in the collection of sets

$$\{\{z_1, \ldots, z_n\} \cap A : A \in \mathcal{A}\}.$$

$N_{\mathcal{A}}(z_1, \ldots, z_n)$ is counting the *number of subsets* of $\{z_1, \ldots, z_n\}$ that the collection of sets $\mathcal{A}$ picks out. Note that, $N_{\mathcal{A}}(z_1, \ldots, z_n) \leq 2^n$ (why?).

We now define the n-th shatter coefficient of $\mathcal{A}$ as:

$$s(\mathcal{A}, n) = \max_{\{z_1, \ldots, z_n\}} N_{\mathcal{A}}(z_1, \ldots, z_n).$$

The shatter coefficient is the maximal number of different subsets of n points that can be picked out by the collection $\mathcal{A}$.

Quick Example: Suppose we considered points in 1D and the set system was the collection of left intervals $\mathbb{I}(-\infty, t]$ for all t.

If we have n points on the line then we can pick out any left subset of the points, i.e. $s(\mathcal{A}, n) = n + 1$. To verify this consider for instance the case when $n = 3$, and we place the points at $\{0, 1, 2\}$, then clearly we can pick out the following subsets:

$$\{\phi\}, \{0\}, \{0, 1\}, \{0, 1, 2\},$$

and no others. We will see more examples in a little bit.

VC Theorem: For *any distribution* $\mathbb{P}$, and class of sets $\mathcal{A}$ we have that,

$$\mathbb{P}(\Delta(\mathcal{A}) \geq t) \leq 8s(\mathcal{A}, n) \exp(-nt^2/32).$$

Notes: There are two noteworthy aspects of this theorem.

1. The result is very general and it applies to any distribution on the samples, and such results are often called *distribution free.*
2. The VC theorem essentially reduces the question of uniform convergence to a combinatorial question about the collection of sets, i.e. we now need only to understand the shatter coefficients which are completely independent from probability/statistics.
3. The proof of this result is quite straightforward using some of the machinery (introducing a ghost sample, symmetrization) that we will see in the next lecture. If you are curious ask me about it.

Glivenko-Cantelli: This theorem immediately implies the Glivenko-Cantelli theorem we studied in the last lecture, i.e. that the empirical CDF converges in probability to the true CDF. To see this we note that the shatter coefficients of the left intervals are bounded by $n + 1$ so the VC theorem tells us that,

$$\mathbb{P}\left(\sup_x |\widehat{F}_n(x) - F_X(x)| \geq t\right) \leq 8(n + 1) \exp(-nt^2/32).$$

Now verifying convergence in probability is straightforward by noting that for any $t > 0$, $\lim_{n\to\infty} 8(n + 1) \exp(-nt^2/32) = 0$.

VC dimension: The paper of Vapnik and Chervonenkis is a work of art. They proved the VC theorem but did not stop there. Perhaps their more remarkable contribution was a theorem/observation about the shatter coefficients of any set system.

Let us first define the VC dimension of a set system $\mathcal{A}$. The VC dimension d is the largest integer d for which $s(\mathcal{A}, d) = 2^d$.

So using this definition we know that for any $n > d$, we have that $s(\mathcal{A}, n) < 2^n$. The surprising combinatorial result of Vapnik and Chervonenkis (sometimes called Sauer's lemma) is that there is a phase transition of shattering coefficients: once it is no longer exponential (i.e. once $n > d$) the shattering coefficients become polynomial in n, i.e.

Sauer's Lemma: If $\mathcal{A}$ has finite VC dimension d, then for $n > d$ we have that,

$$s(\mathcal{A}, n) \leq (n+1)^d.$$

We can use Sauer's lemma to conclude that for a system $\mathcal{A}$ of VC dimension d.

$$\mathbb{P}(\Delta(\mathcal{A}) \geq t) \leq 8(n+1)^d \exp(-nt^2/32).$$

Doing the usual thing we see that with probability $1 - \delta$,

$$\Delta(\mathcal{A}) \leq \sqrt{\frac{32}{n}\left[d \log(n+1) + \log(8/\delta)\right]}.$$

There are some important notes:

1. If $d < \infty$ then $\Delta(\mathcal{A}) \xrightarrow{p} 0$, and so we have a uniform LLN for the collection of sets $\mathcal{A}$.
2. There are converses to the VC theorem that say roughly that if the VC dimension is infinite then there exists a distribution over the samples for which we do not have a uniform LLN.
3. Roughly, one should think of the VC result as saying for a class with VC dimension d,

$$\Delta(\mathcal{A}) \approx \sqrt{\frac{d \log n}{n}}.$$

9.3 More examples

There are many examples of collections of sets for which the VC dimension is known. A few popular ones are in Table 9.1.

9.4 Connecting back to binary classification

I will be a bit hand-wavy here and you should take 702 to see this done more clearly. In binary classification, we have a collection of classifiers $\mathcal{F}$. This collection induces a set system:

$$\mathcal{A} = \{\{\{x : f(x) = 1\} \times \{0\}\} \bigcup \{\{x : f(x) = 0\} \times \{1\}\}, f \in \mathcal{F}\}.$$

Class $\mathcal{A}$	VC dimension $V_{\mathcal{A}}$
$\mathcal{A} = \{A_1, \ldots, A_N\}$	$\leq \log_2 N$
Intervals $[a, b]$ on the real line	2
Discs in $\mathbb{R}^2$	3
Closed balls in $\mathbb{R}^d$	$\leq d + 2$
Rectangles in $\mathbb{R}^d$	$2d$
Half-spaces in $\mathbb{R}^d$	$d + 1$
Convex polygons in $\mathcal{R}^2$	∞
Convex polygons with d vertices	$2d + 1$

Table 9.1: The VC dimension of some classes $\mathcal{A}$.

If $\mathcal{A}$ has VC dimension d then we can use the VC theorem in a straightforward way to conclude that with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - \mathbb{P}(f(X) \neq y)| = \Delta(\mathcal{A}) \leq \sqrt{\frac{32}{n} [d \log(n+1) + \log(8/\delta)]}.$$

It is not too hard to verify that the VC dimension is essentially driven by the complexity of the sets $\mathbb{I}(f(x) = 1)$ and their complements for the classifiers in $\mathcal{F}$. This in a straightforward way, for instance, leads to a uniform convergence guarantee for empirical risk minimization over linear classifiers since they induce relatively simple sets (half-spaces) whose VC dimension is well-understood.