Vapnik-Chervonenkis theory

Risi Kondor

June 13, 2008

For the purposes of this lecture, we restrict ourselves to the binary supervised batch learning setting. We assume that we have an input space $\mathcal{X}$, and an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \{-1, +1\}$. Given a training set $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$ drawn from $\mathcal{D}$, our learning algorithm tries to find a hypothesis $\widehat{h}\colon \mathcal{X} \to \{-1, +1\}$ that will predict well on future (x, y) examples drawn from $\mathcal{D}$ in the sense that the "true error"

$$\mathcal{E}(\widehat{h}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \mathbb{I}\,[\,\widehat{h}(x) \neq y\,] \tag{1}$$

will not be too big. Of course we cannot measure $\mathcal{E}(\widehat{h})$, because we don't know $\mathcal{D}$. What we have instead is the empirical error measured on the sample S:

$$\mathcal{E}_S(\widehat{h}) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\,[\,\widehat{h}(x_i) \neq y\,]. \tag{2}$$

This is otherwise known as the training error, and typicially it is an overoptimistic estimate of the true error, simply because the way most learning algorithms work is to implicitly or explicitly drive down just this quantity. Some amount of overfitting is then unavoidable.

The job of generalization bounds is to relate these two quantities, so that we can report things like "my algorithm found the hypothesis $\widehat{h}$, on the training data it has error $\mathcal{E}_S$, and on future data the error is not likely to be more than ϵ worse". In other words,

$$\mathcal{E}(\widehat{h}) - \mathcal{E}_S(\widehat{h}) < \epsilon.$$

No matter how we set ϵ, however, a really really bad training set (in the sense of a very unrepresentative sample) can always mislead us even more, so explicit bounds of this form are generally not obtainiable. The "P" part of PAC stands for aiming for guarantees of this form only in the probabilistic sense, i.e., finding (ϵ, δ) pairs, where both are small positive real numbers, such that

$$\mathbb{P}_S[\,\mathcal{E}(\widehat{h}) - \mathcal{E}_S \geq \epsilon\,] \;\leq\; 1 - \delta. \tag{3}$$

Here $\mathbb{P}_S$ stands for "probability over choice of training set".

Concentration

The fundamental idea behind all generalization bounds is that although it is possible that the same quantity, in our case, the error, will turn out to be very different on two different samples from the same distribution, this is not likely to happen. In general, as the sample size grows, empirical quantities tend to concentrate more and more around their mean, in our case, the true error. The simplest inequality capturing this fact, and one that is key to our development, is Hoeffding's inequality, which states that if $Z_1, Z_2, \ldots, Z_m$ are independent draws of a Bernoulli random variable with parameter p and $\gamma > 0$, then

$$\mathbb{P}\left[\ \frac{1}{m}\sum_{i=1}^{m} Z_i > p + \gamma\ \right] \leq e^{-2m\gamma^2}.$$

This fits our problem nicely, because for any $h \in C$ if we take $Z_i = \mathbb{I}\left[h(x_i) \neq y_i\right]$, Hoeffding's inequality says something about the probability of deviations from the true error $p = \mathcal{E}(h)$ of the hypothesis h.

Plugging in the hypothesis $\widehat{h}$ returned by algorithm and blindly applying Hoeffding's inequality gives

$$\mathbb{P}[\ \mathcal{E}(\widehat{h}) - \mathcal{E}_S(\widehat{h}) > \epsilon\ ] \leq e^{-2m\epsilon^2},$$

so setting the right hand side equal to $1 - \delta$, the PAC bound

$$\mathbb{P}[\ \mathcal{E}(\widehat{h}) - \mathcal{E}_S(\widehat{h}) > \epsilon\ ] < \delta \tag{4}$$

is satisfied when

$$\epsilon > \sqrt{\frac{\ln\left(1/\delta\right)}{2m}}.$$

Unfortunately, this simple argument is NOT CORRECT. What goes wrong is that *given the fact that our algorithm returned* $\widehat{h}$, the (x_i, y_i) examples in the sample are *not* IID samples from $\mathcal{D}$, and consequently neither are the Z_i's IID samples from Bernoulli$(\mathcal{E}(\widehat{h}))$.

One way to put this coupling between the sample and the hypothesis in relief is to note that our algorithm is really just a function $A\colon S \mapsto \widehat{h}$. This makes it clear that the empirical error is a function of S in two ways: through the hypothesis $A(S)$ and the individual training examples $(x, y) \in S$. Clearly, we can't hold one constant and regard $\mathcal{E}_S$ as a statistic based on IID draws of the other.

Uniform convergence

Overcoming the problem of the coupling between S and $\widehat{h}$ is a major hurdle in proving generalization bounds. The way to proceed is to instead of focusing on any one particular h, to focus on all of them simulatenously. In particular, if we can find an (ϵ, δ) pair such that

$$\mathbb{P}[\, \mathcal{E}(h) - \mathcal{E}_S(h) \leq \epsilon \;\; \forall h \in C \,] \geq 1 - \delta,$$

or equivalently,

$$\mathbb{P}[\, \exists\, \widehat{h} \in C \;\text{ such that }\; \mathcal{E}(h) - \mathcal{E}_S(h) > \epsilon \,] < \delta,$$

then that (ϵ, δ) pair will certainly satisfy the PAC bound (3). At first sight this seems like a terrible overkill, since C might include some crazy irregular functions that might never be chosen by any reasonable algorithm, but these functions might make our bound very loose. On the other hand, it is worth noting that at least amongst "reasonable" functions the $\widehat{h}$ chosen by our learning algorithm is actually likely to be towards the top of the list in terms of the magnitude of $\mathcal{E}(h) - \mathcal{E}_S(h)$, simply because learning algorithms by their very nature tend to drive down $\mathcal{E}_S(h)$. So bounding $\mathcal{E}(h) - \mathcal{E}_S(h)$ for *all* $h \in C$ might not be such a crazy thing to do after all. How best to do it is not obvious, though.

If C has only a finite number of hypotheses, the simplistic method is to use the union bound:

$$\mathbb{P}[\, \exists\, \widehat{h} \in C \text{ such that } \mathcal{E}(h) - \mathcal{E}_S(h) > \epsilon \,] \;\leq\; \sum_{h \in C} \mathbb{P}[\, \mathcal{E}(h) - \mathcal{E}_S(h) > \epsilon \,]$$

where on the right hand side now we are allowed to use the Hoeffding bound

$$\mathbb{P}[\, \mathcal{E}(h) - \mathcal{E}_S(h) > \epsilon \,] \leq e^{-2m\epsilon^2},$$

leading to $|\, C \,|\, e^{-2m\gamma^2} \leq \delta$ and therefore

$$\epsilon > \sqrt{\frac{\ln |\, C \,| + \ln(1/\delta)}{2m}}\,.$$

The union bound is clearly very loose though, and the explicit appearance of the number of hypotheses in C is also worrying: what if two hypotheses are "almost" identical? Should we still count them as separate? Clearly, there must be some more appropriate way of quantifying the richness of a concept space than just counting the number of hypotheses in C. VC theory is an attempt to do just this.

Symmetrization

Concentration inequalities don't just tell us that on a single sample S, $\mathcal{E}_S(h)$ can't deviate too much from its mean $\mathcal{E}(h)$, they also imply that for a pair of independent samples S_1 and S_2, $\mathcal{E}_{S_1}(h)$ can't be very far from $\mathcal{E}_{S_2}(h)$. The key idea behind Vapnik and Chervonenkis' pioneering work was to use exploit this fact by a process called symmetrization to reduce everything to just looking at finite samples. We start with the following simple application of Hoeffding's inequality.

Proposition 1 *Let S and S' be two independent samples of size $2m$ drawn from a distribution $\mathcal{D}$ on $\mathcal{X} \times \{-1, +1\}$ and let $\mathcal{E}$ and $\mathcal{E}_S$ be defined as in (1) and (2). Then for any $h \in C$,*

$$\mathbb{P}\left[\, \mathcal{E}(h) - \mathcal{E}_S(h) > \epsilon \,\right] \;\leq\; 2\,\mathbb{P}\left[\, \mathcal{E}_{S'}(h) - \mathcal{E}_S(h) > \epsilon/2 \,\right].$$

This result also readily generalizes to the uniform case.

Proposition 2 *Let S and S' be two independent samples of size $2m$ drawn from a distribution $\mathcal{D}$ on $\mathcal{X} \times \{-1, +1\}$ and let $\mathcal{E}$ and $\mathcal{E}_S$ be defined as in (1) and (2). Then for any $h \in C$,*

$$\mathbb{P}\left[\sup_{h \in C} \left[\mathcal{E}(h) - \mathcal{E}_S(h)\right] > \epsilon \right] \;\leq\; 2\,\mathbb{P}\left[\sup_{h \in C} \left[\mathcal{E}_{S'}(h) - \mathcal{E}_S(h)\right] > \epsilon/2 \right].$$

Now let us define $\overline{S} = S \cup S'$ and ask ourselves: what is the probability that the errors incurred by h on S are distributed in such a way that $m\epsilon/2$ more of them fall in S' than in S? For the sake of simplicity here we only consider the case that exactly 0 errors fall in S and $k = m\epsilon/2$ fall in S'.

Proposition 3 *Consider $2m$ balls of which exactly k balls are black. If we randomly split the balls into two sets of size m, then the probability $P_{k,0}$ that all the black balls end up in the first set is at most $1/2^k$*

Proof.

$$P_{k,0} = \binom{m}{k} \Big/ \binom{2m}{k} = \frac{m(m-1)(m-2)\ldots(m-k)}{(2m)(2m-1)\ldots(2m-k)} < 1/2^k.$$

■

For the general case of u and $u + k$ balls in the two sets, a similar combinatorial inequality holds.

The real siginficance of symmetrization is that it allows us to quantify the complexity of C in terms of just the joint sample $\overline{S}$ instead of its behavior on the entire input space. In particular, in bounding $\sup_h \left[\mathcal{E}_{S'}(h) - \mathcal{E}_S(h)\right]$, two hypothesis h and h' only need to be counted as distinct if they differ on $\overline{S}$: how they behave over the rest of $\mathcal{X}$ is immaterial. To be somewhat more explict, we define the restriction of h to $\overline{S}$ as

$$h\!\downarrow_{\overline{S}} \colon \overline{S} \to \{-1, +1\} \qquad\qquad h\!\downarrow_{\overline{S}}(x) = h(x),$$

and the corresponding restricted concept class as $C\downarrow_{\overline{S}} = \{\, h\downarrow_{\overline{S}} \mid h \in C \,\}$. While $|\, C\downarrow_{\overline{S}} \,|$ is of course a property of $\overline{S}$, it is also a characteristic of the entire concept class C in the sense that it is often possible to bound its size independently of $\overline{S}$. The maximal rate at which $|\, C\downarrow_{\overline{S}} \,|$ grows with the size of m,

$$\Pi_C(n) = \max_{U \in \mathcal{X}^n} |\, C\downarrow_U \,|$$

is called the **growth function**.

Using the growth function and Proposition 3, we can now give the finite sample version of the union bound:

$$\mathbb{P}\left[\sup_{h \in C} \big[\mathcal{E}_{S'}(h) - \mathcal{E}_S(h)\big] > \epsilon/2 \right] \;\leq\; \Pi_C(2m)\, 2^{-m\epsilon/2}.$$

The VC dimension is solely a device for computing $\Pi_C(n)$.

The VC dimension

The concept class C is said to **shatter** a set $S \subset \mathcal{X}$ if C can realize all possible labelings of S, i.e., if $C\downarrow_S = 2^{|\, S \,|}$. The Vapnik-Chervonenkis dimension of C is the size of the largest subset of S that C can shatter,

$$\mathrm{VC}(C) = \max_{|\, S \,|} \{\, S \subset \mathcal{X} \text{ and } |\, C\downarrow_S \,| = 2^n \,\}.$$

The following famous result (called the Sauer-Shelah lemma) tells us how to bound the growth function in terms of the VC dimension.

Proposition 4 *(Sauer-Shelah lemma) Let C be a concept class of VC-dimension d, and let $\Pi(m)$ be the corresponding growth function. Then for $m \leq d$, $\Pi(m) = 2^m$; and for $m > d$,*

$$\Pi(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d. \tag{5}$$

Proof. The $m \leq d$ case is just a restatement of the defininition of VC dimension. For $m > d$ we use induction on m. For $m = 2$, it is trivial to show that (5) holds. Assuming that it holds for m and any d, we now show that it also holds for $m+1$ and any d. Let S be any subset of $\mathcal{X}$ of size $m+1$, and fixing some $x \in S$, let us write it as $S = S^{\setminus x} \cup \{x\}$, where $|\, S^{\setminus x} \,| = m$. Now for any $h \in C\downarrow_{S^{\setminus x}}$, consider its two possible extensions

$$\begin{aligned} h^+ &: S \to \{-1, +1\} \quad \text{with} \quad h^+(x') = h(x') \quad \text{for} \quad x' \in S \quad \text{and} \quad h^+(x) = 1 \\ h^- &: S \to \{-1, +1\} \quad \text{with} \quad h^-(x') = h(x') \quad \text{for} \quad x' \in S \quad \text{and} \quad h^-(x) = -1. \end{aligned}$$

Either both of these hypotheses are in $C\downarrow_S$ or only one of them is. Let U be the subset of $C\downarrow_{S^{\setminus x}}$ of hypotheses for which both h^+ and h^- are in $C\downarrow_S$, and let $U\uparrow^S = \bigcup_{h \in U} \{h^+, h^-\}$. Then we have

$$|\, C\downarrow_S \,| = |\, C\downarrow_{S^{\setminus x}} \,| + |\, U \,|.$$

By the inductive hypothesis $|\, C\downarrow_{S\setminus x} \,| \leq \sum_{i=1}^{d} \binom{m}{i}$. As for the second term, consider that if U shatters any set V, then $U\uparrow^S$ will shatter $V \cup \{x\}$, so $\mathrm{VC}(U) \leq \mathrm{VC}(U\uparrow^S) - 1 \leq d-1$, so by the inductive hypothesis $|\, U \,| \leq \sum_{i=1}^{d-1} \binom{m}{i}$. Therefore

$$|\, C\downarrow_S \,| \leq \sum_{i=0}^{d} \binom{m}{i} + \sum_{i=0}^{d-1} \binom{m}{i} = \sum_{i=0}^{d} \binom{m+1}{i},$$

since $\sum_{i=1}^{d} \binom{m+1}{i}$ is just the number of ways of choosing up to d objects from $m+1$, and the sum corresponds to decomposing these choices according to whether a particular object has been chosen or not. Finally,

$$\sum_{i=1}^{d} \binom{m}{i} < \left(\frac{m}{d}\right)^d \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^i < \left(\frac{m}{d}\right)^d \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i = \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m \leq \left(\frac{m}{d}\right)^d \exp(d)$$

■

Putting it all together