PAC Bounds

Carlos C. Rodríguez
http://omega.albany.edu:8008/

October 7, 2004

What do we have so far?

Let's recall what the VC-theory, introduced in the previous lectures, has accomplished so far.

First we played God and assumed we knew the underlying distribution of $Z = (X, Y)$ that is generating the labeled data. In this case we can find the bayes classifier δ^* by simply assigning the label y to an observed vector x that maximizes the posterior probability given the observation. Then, we tried the obvious next thing. Instead of assuming that we knew the distribution of Z we assumed that we had the possibility of collecting a sample $Z_1, \ldots, Z_n$ of independent observations with the same distribution as Z. We can then use these examples (e.g. provided by a trusted teacher) to approximate the necessary probabilities $P\{Z \in A\}$ with the empirical frequencies $P_n\{A\}$ defined by,

$$P_n\{A\} = \frac{1}{n}\sum_{i=1}^{n} 1[Z_i \in A].$$

In particular the VC-theory estimates the risk $R(g) = P\{g(X) \neq Y\}$ of a given classifier g with the observed frequency of errors $R_n(g) = P_n\{A\}$ where,

$$A = \{z = (x, y) : g(x) \neq y\}.$$

Hence, the question becomes: How good is P_nA as an estimate of PA ? In this way the LLNs got naturally into the picture. It is intuitively clear that some kind of constraint on g is needed to avoid over fitting the observations. For otherwise we could always take $g_n(X_i) = Y_i$ with no empirical error (i.e. fitting the observed data perfectly) but making a mistake on all other possible data, i.e. $g_n(X) = 1 - Y$ so that the true error is worst possible, e.g. $R(g_n) = 1$ when P has no atoms. We are then forced to consider $g \in \mathcal{G}$ only. There is still another compelling reason for restricting classifiers to a given class $\mathcal{G}$. That has to do with the issue of consistency. We mentioned (without proof) the existence of *No Free Lunch* theorems stating that there are no universally consistent classification rules with fix rates of convergence. Sure, there are universally consistent rules g_n (e.g. the knn) for which $R(g_n) \to R^*$ (the bayes

risk) no matter what the underlying distribution of (X, Y) is. However, one can show that for any sequence of positive numbers a_n converging to zero, there exist distributions for which $R(g_n) \geq a_n$ for all n. In other words, truth is asymptotically learnable but it may take forever to find out what it actually is!

The need to consider prior information is unavoidable. Data by itself cannot provide the ability to generalize. That may come as a surprise to some empiricists but for the bayesians (specially for those who have read Jaynes) data is another logical proposition and it only has meaning in a given domain of discourse. For example the number 2.5 or the binary string 010011 are not data. They are only data if they are understood as a logical proposition like "*the result of such and such experiment was* 2.5". In other words we cannot isolate data from prior knowledge. Data is a function of the theory, or the way I like to put it:

There is no data in the vacuum.

Yes, the currently accepted interpretations of quantum theory are in conflict with this view. But the currently accepted interpretations of quantum theory are in conflict with themselves, so there is nothing to worry about.

The need for Uniform LLNs

Thus, we need $g \in \mathcal{G}$. What $\mathcal{G}$ s are good $\mathcal{G}$ s? To answer this let $g_n \in \mathcal{G}$ be the classifier minimizing the empirical risk over $\mathcal{G}$, i.e.

$$R_n(g_n) \leq R_n(g) \text{ for all } g \in \mathcal{G}$$

and let $g^* \in \mathcal{G}$ be the best classifier in $\mathcal{G}$, i.e.,

$$R(g^*) \leq R(g) \text{ for all } g \in \mathcal{G} .$$

Finally let R^* be the bayes risk, i.e.,

$$R^* \leq R(g) \text{ for all } g.$$

The estimated error relative to the bayes risk decomposes into two parts,

$$R_n(g_n) - R^* = [R_n(g_n) - R(g^*)] + [R(g^*) - R^*]$$

The first term in square brackets is the *estimation error* and the second term is the *approximation error* . Better and bigger sampling reduces the first error and better and bigger models $\mathcal{G}$ reduce the second. It is important to notice the tug-of-war about model size implicit in the error decomposition. On the one hand, small models allowing few choices $g \in \mathcal{G}$ should make the task of identifying the best $g^* \in G$, with the information provided by the observed data, easier. In other words, the smaller $\mathcal{G}$ is, the more helpful the sample becomes. On the other hand, the bigger the model the closer $g^* \in \mathcal{G}$ is from the bayes rule, and the smaller the approximation error. We'll concentrate primarily on the sampling error.

Three important inequalities to consider:

1. Observed error compared to minimum true risk in $\mathcal{G}$.

$$R_n(g_n) - R(g^*) = [R_n(g_n) - R_n(g^*)] + [R_n(g^*) - R(g^*)]$$

the first term in brackets is non positive, so

$$R_n(g_n) - R(g^*) \leq [R_n(g^*) - R(g^*)]$$

thus,

$$R_n(g_n) - R(g^*) \leq \sup_{\mathcal{G}} |R_n(g) - R(g)|$$

2. True error of estimator compared to best in $\mathcal{G}$.

$$R(g_n) - R(g^*) = [R(g_n) - R_n(g_n)] + [R_n(g_n) - R(g^*)]$$

and using $x \leq |x|$ and the previous inequality we get,

$$R(g_n) - R(g^*) \leq 2 \sup_{\mathcal{G}} |R_n(g) - R(g)|$$

3. True error compared to empirical error.

$$R(g_n) - R_n(g_n) \leq \sup_{\mathcal{G}} |R_n(g) - R(g)|$$

The three inequalities above led us to consider,

$$\sup_{\mathcal{G}} |R_n(g) - R(g)| = \sup_{\mathcal{A}} |P_n A - PA| \ := \|R_n - R\|$$

and the uniform laws of large numbers came to the table.

Probably Approximately Correct Bounds

The VC inequality,

$$E\|R_n - R\| \leq 2 \left(\frac{\log \ 2\mathcal{S} \ (n, \mathcal{A} \ )}{n} \right)^{1/2},$$

and MacDiarmid's concentration inequality,

$$P\{\|R_n - R\| - E\|R_n - R\| > t\} \leq e^{-2nt^2}$$

imply,

$$P\{\|R_n - R\| > \epsilon\} \leq \exp\{-2n(\epsilon - 2 \left(\frac{\log \ 2\mathcal{S} \ (n, \mathcal{A} \ )}{n} \right)^{1/2})^2\}.$$

Inversion

The following simple fact is often useful for obtaining PAC bounds. If

$$P\{X > t\} \leq F(t)$$

then with probability at least $1 - \delta$,

$$X \leq F^{-1}(\delta).$$

PAC Bounds for $R(g_n)$

Thus, for any small $\delta > 0$ with probability at least $1 - \delta$,

$$R(g_n) \leq R(g^*) + 2\Delta_n$$

and

$$R(g_n) \leq R_n(g_n) + \Delta_n$$

where,

$$\Delta_n \leq \left(\frac{1}{2n} \log \frac{1}{\delta}\right)^{1/2} + 2\left(\frac{\log\ 2\mathcal{S}\ (n, \mathcal{A}\ )}{n}\right)^{1/2}$$

The Finite Case

It is highly informative to compare the general bound above with the one obtained when the class $\mathcal{G}$ is finite, i.e., when $|\mathcal{G}\ | = N < \infty$. Using Boole's and Hoeffding's inequalities we can write,

$$\begin{aligned} P\{\sup_{\mathcal{G}} |R_n(g) - R(g)| > t\} &\leq P\left\{\bigcup_{g \in \mathcal{G}} \{|R_n(g) - R(g)| > t\}\right\} \\ &\leq \sum_{\mathcal{G}} P\{|R_n(g) - R(g)| > t\} \\ &\leq N 2e^{-2nt^2} \end{aligned}$$

and applying the inversion method (see above) we get,

$$\Delta_n \leq \left(\frac{1}{2n} \log \frac{1}{\delta}\right)^{1/2} + 2\left(\frac{\log\ 2N}{n}\right)^{1/2}$$

The interpretation of the shatter coefficient $\mathcal{S}\ (n, \mathcal{A}\ )$ as an effective size N when the class is infinite is plain to see. The shatter coefficients (and thus the VC dimension as well) give us the maximum number of points in $\mathcal{G}$ that can be distinguished on the basis of a finite sample. Recalling that N equally likely choices have entropy $\log\ N$ we could naturally interpret the logarithm of the

shatter coefficients as a kind of entropy for the infinite class $\mathcal{G}$. This, I believe, is the most important lesson to be learned from the VC theory. We have now a method to reduce large (infinite dimensional) non-parametric classes to an effective finite number that we can actually tell apart from each other on the basis of only $n < \infty$ observations. The rest is counting. The bottom line is combinatorics. The fruits of this theory are immense!