

ELEN6887

Lecture 19: The proof of the Vapnik-Chervonenkis (VC) Inequality

R. Castro

4/27/2009

In the last lecture we have seen that it is possible to derive generalization bounds even if the involved classes of models are uncountable. The main intuition is that although the model class is infinite, using a finite set of training data to select a good rule effectively reduces the number of different models we need to consider. We can measure the effective size of a class \mathcal{F} using the *shatter coefficient* $\mathcal{S}(\mathcal{F}, n)$ introduced before. The behavior of $\mathcal{S}(\mathcal{F}, n)$ with n is quite interesting: for n smaller or equal to some value $VC(\mathcal{F})$ we have $\mathcal{S}(\mathcal{F}, n) = 2^n$ (\mathcal{F} is said to “shatter” n points in this case), and for $n > VC(\mathcal{F})$ we have $\mathcal{S}(\mathcal{F}, n) < 2^n$. The number $VC(\mathcal{F})$ is called the Vapnik-Chervonenkis (VC) dimension.

At this point it is important to review the notation and setting used in last class. Let \mathcal{X} denote the feature space (e.g., $\mathcal{X} = \mathbf{R}^d$), and $\mathcal{Y} = \{0, 1\}$ denote the label space. Let $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a classifier (also called predictor or model). Let $(X, Y) \sim P_{XY}$ where the joint probability distribution. Note that P_{XY} is generally unknown to us. We measure the classification/prediction error using the 0/1 loss function $\ell(f(X), Y) = \mathbf{1}\{f(X) \neq Y\}$, which gives rise to the risk $R(f) = E[\ell(f(X), Y)] = P(f(X) \neq Y)$.

Let \mathcal{F} be a class of candidate models (classification rules). We would like to choose a good model (that is, a model with small probability of error). We don't have access to P_{XY} but only to a training sample D_n ,

$$D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\},$$

where $(X_i, Y_i) \stackrel{i.i.d.}{\sim} P_{XY}$. Given this sample we can define the empirical risk

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{f(X_i) \neq Y_i\}.$$

At the end of lecture 18 we stated the following result.

Theorem 1. (Vapnik-Chervonenkis '71): *For binary classification and the 0/1 loss function we have the following generalization bounds.*

$$P \left(\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| > \epsilon \right) \leq 8\mathcal{S}(\mathcal{F}, n)e^{-n\epsilon^2/32},$$

and

$$E \left[\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \right] \leq 2\sqrt{\frac{\log \mathcal{S}(\mathcal{F}, n) + \log 2}{n}}.$$

We will prove the first inequality. A slightly weaker version of the second inequality can be easily derived from the first one, but a more careful and direct proof gives rise to the one stated in the theorem.

Proof: We will follow closely the approach presented in the book by Devroye, Györfi and Lugosi.

The proof of this result is rather involved, but it can be broken into several important steps.

The main idea is to get to a point where we can take advantage of the effective size of the class induced by the training data. To do this we will need the introduction of a “ghost sample”, that is another sequence of data in all identical to the training data D_n . This ghost sample helps us developing the proof, but it doesn't play a role in the end result (that is, you don't need a ghost sample to apply the result). Let $D'_n = \{(X'_1, Y'_1), \dots, (X'_n, Y'_n)\}$ be a set of random variables independent of D_n , such that $(X'_i, Y'_i) \stackrel{i.i.d.}{\sim} P_{XY}$. This is called the ghost sample. Define also the empirical risk under this sample by

$$\hat{R}'_n(f) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{f(X'_i) \neq Y'_i\} .$$

For the rest of the proof we will assume that $n\epsilon^2 \geq 2$ without loss of generality, since otherwise the bound stated in the theorem is trivial.

Step 1: First symmetrization by a ghost sample:

We are going to show that

$$P \left(\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| > \epsilon \right) \leq 2P \left(\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - \hat{R}'_n(f) \right| > \frac{\epsilon}{2} \right) . \quad (1)$$

Notice that the absolute value term on the right-hand-side is now symmetric, involving two different empirical risks. Begin by defining $f(D_n) \equiv \tilde{f}$ to be an element of \mathcal{F} such that $\left| \hat{R}_n(f) - R(f) \right| > \epsilon$ if such element exists, otherwise \tilde{f} is an arbitrary element of \mathcal{F} . You can think of \tilde{f} as

$$\tilde{f} \approx \arg \max_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| ,$$

although this is generally not well defined there might not be an element in \mathcal{F} attaining the maximum (because \mathcal{F} is typically infinite). The formal definition is an attempt of circumventing this technicality that suffices for our purposes. Notice that \tilde{f} is a function of D_n (we dropped the explicit dependence to make the presentation cleaner).

Now let's look at the right-hand-side of (1).

$$\begin{aligned} & P \left(\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - \hat{R}'_n(f) \right| > \frac{\epsilon}{2} \right) \geq P \left(\left| \hat{R}_n(\tilde{f}) - \hat{R}'_n(\tilde{f}) \right| > \frac{\epsilon}{2} \right) \\ & \geq P \left(\left| \hat{R}_n(\tilde{f}) - R(\tilde{f}) \right| > \epsilon \text{ and } \left| \hat{R}'_n(\tilde{f}) - R(\tilde{f}) \right| < \frac{\epsilon}{2} \right) \\ & = E \left[\mathbf{1} \left\{ \left| \hat{R}_n(\tilde{f}) - R(\tilde{f}) \right| > \epsilon \right\} \mathbf{1} \left\{ \left| \hat{R}'_n(\tilde{f}) - R(\tilde{f}) \right| < \frac{\epsilon}{2} \right\} \right] \\ & = E \left[\mathbf{1} \left\{ \left| \hat{R}_n(\tilde{f}) - R(\tilde{f}) \right| > \epsilon \right\} E \left[\mathbf{1} \left\{ \left| \hat{R}'_n(\tilde{f}) - R(\tilde{f}) \right| < \frac{\epsilon}{2} \right\} \middle| D_n \right] \right] \\ & = E \left[\mathbf{1} \left\{ \left| \hat{R}_n(\tilde{f}) - R(\tilde{f}) \right| > \epsilon \right\} P \left(\left| \hat{R}'_n(\tilde{f}) - R(\tilde{f}) \right| < \frac{\epsilon}{2} \middle| D_n \right) \right] , \end{aligned}$$

where the second inequality follows from the fact that for any reals x , y and z

$$|x - z| > \epsilon \text{ and } |y - z| \leq \epsilon/2 \quad \Rightarrow \quad |x - y| \geq \epsilon/2 .$$

Now, conditionally on D_n we see that

$$\hat{R}'_n(\tilde{f}) - R(\tilde{f}) = \frac{1}{n} \sum_{i=1}^n U_i ,$$

where $U_i = \mathbf{1}\{\tilde{f}(X'_i) \neq Y'_i\} - E[\mathbf{1}\{\tilde{f}(X'_i) \neq Y'_i\} | D_n]$ are zero-mean i.i.d. random variables. We are in good shape to use a concentration inequality here, although for our purposes Chebyshev's inequality suffices (for a non-negative random variable Z this inequality says that $P(Z > t) \leq \text{Var}(Z)/t^2$).

$$\begin{aligned}
P\left(\left|\hat{R}'_n(\tilde{f}) - R(\tilde{f})\right| < \frac{\epsilon}{2} \mid D_n\right) &= P\left(\left|\frac{1}{n} \sum_{i=1}^n U_i\right| < \frac{\epsilon}{2} \mid D_n\right) \\
&= P\left(\left|\sum_{i=1}^n U_i\right| < \frac{n\epsilon}{2} \mid D_n\right) \\
&\geq 1 - \frac{4}{n^2\epsilon^2} \text{Var}\left(\left|\sum_{i=1}^n U_i\right| \mid D_n\right) \\
&= 1 - \frac{4}{n^2\epsilon^2} n \text{Var}(U_i \mid D_n) \\
&\geq 1 - \frac{4}{n\epsilon^2} \frac{1}{4} = 1 - \frac{1}{n\epsilon^2} \geq \frac{1}{2},
\end{aligned}$$

since we assumed that $n\epsilon^2 \geq 2$. Finally

$$\begin{aligned}
&P\left(\sup_{f \in \mathcal{F}} \left|\hat{R}_n(f) - \hat{R}'_n(f)\right| > \frac{\epsilon}{2}\right) \\
&\geq E\left[\mathbf{1}\left\{\left|\hat{R}_n(\tilde{f}) - R(\tilde{f})\right| > \epsilon\right\} P\left(\left|\hat{R}'_n(\tilde{f}) - R(\tilde{f})\right| < \frac{\epsilon}{2} \mid D_n\right)\right] \\
&\geq \frac{1}{2} E\left[\mathbf{1}\left\{\left|\hat{R}_n(\tilde{f}) - R(\tilde{f})\right| > \epsilon\right\}\right] \\
&= \frac{1}{2} P\left(\left|\hat{R}_n(\tilde{f}) - R(\tilde{f})\right| > \epsilon\right) \\
&\geq \frac{1}{2} P\left(\sup_{f \in \mathcal{F}} \left|\hat{R}_n(f) - R(f)\right| > \epsilon\right),
\end{aligned}$$

concluding the proof of (1), where the last step follows from the definition of \tilde{f} .

Step 2: Symmetrization by random signs.

Let's rewrite the right-hand-side of (1).

$$P\left(\sup_{f \in \mathcal{F}} \left|\hat{R}_n(f) - \hat{R}'_n(f)\right| > \frac{\epsilon}{2}\right) = P\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left|\sum_{i=1}^n \mathbf{1}\{f(X_i) \neq Y_i\} - \mathbf{1}\{f(X'_i) \neq Y'_i\}\right| > \frac{\epsilon}{2}\right).$$

Note that $\mathbf{1}\{f(X_i) \neq Y_i\}$ and $\mathbf{1}\{f(X'_i) \neq Y'_i\}$ have the same distribution, and therefore $\mathbf{1}\{f(X_i) \neq Y_i\} - \mathbf{1}\{f(X'_i) \neq Y'_i\}$ has zero mean and a symmetric distribution¹. So if we randomly permute the signs inside the absolute value term we won't change the probability. Let's introduce another "ghost sample"-like sequence.

Let $\sigma_1, \dots, \sigma_n$ be i.i.d. random variables, independent of D_n and D'_n , such that $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$. These are called Rademacher random variables. In light of our remarks

¹A zero-mean random variable Z has a symmetric distribution if Z and $-Z$ have the same distribution.

above

$$\begin{aligned}
& P \left(\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - \hat{R}'_n(f) \right| > \frac{\epsilon}{2} \right) \\
&= P \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \mathbf{1}\{f(X_i) \neq Y_i\} - \mathbf{1}\{f(X'_i) \neq Y'_i\} \right| > \frac{\epsilon}{2} \right) \\
&= P \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (\mathbf{1}\{f(X_i) \neq Y_i\} - \mathbf{1}\{f(X'_i) \neq Y'_i\}) \right| > \frac{\epsilon}{2} \right) \\
&\leq P \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X_i) \neq Y_i\} \right| > \frac{\epsilon}{4} \text{ or} \right. \\
&\quad \left. \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X'_i) \neq Y'_i\} \right| > \frac{\epsilon}{4} \right) \\
&\leq 2P \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X_i) \neq Y_i\} \right| > \frac{\epsilon}{4} \right),
\end{aligned}$$

where the last inequality follows simply from a union bound over the two events of the previous line. So in these two steps we have shown that

$$P \left(\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| > \epsilon \right) \leq 4P \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X_i) \neq Y_i\} \right| > \frac{\epsilon}{4} \right), \quad (2)$$

Note that we manage to reduce the expression in the left-hand-side of the above to a bound on the sum of i.i.d. zero-mean random variables. We also eliminated the dependence of the ghost-sample D'_n we had at the end of step 1. We are now ready to take advantage of the effective size of the \mathcal{F} with respect to D_n , which will be the next step.

Step 3: Conditioning on D_n .

This step is conceptually the same we used in all the generalization bounds we derived in the course so far. We are going to perform a union bound over all the models under consideration. The difference here is that this set is no longer the entire class \mathcal{F} , but instead just a finite subset of it.

Let $x_1, \dots, x_n \in \mathcal{X}$ and $y_1, \dots, y_n \in \mathcal{Y}$ be arbitrary sequences. Let's examine the quantity

$$\frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(x_i) \neq y_i\} \right|,$$

where the randomness is solely on the random signs σ_i . We have observed in the previous lecture that the sequence $(f(x_1), \dots, f(x_n))$ can take at most $\mathcal{S}(\mathcal{F}, n)$ different values, therefore

$$(\mathbf{1}\{f(x_1) \neq y_1\}, \dots, \mathbf{1}\{f(x_n) \neq y_n\}),$$

can take at most $\mathcal{S}(\mathcal{F}, n)$ different values. Let $\mathcal{F}(x_1, \dots, x_n) \subseteq \mathcal{F}$ be the smallest subset of \mathcal{F} such that

$$N_{\mathcal{F}}(x_1, \dots, x_n) = N_{\mathcal{F}(x_1, \dots, x_n)}(x_1, \dots, x_n)$$

where as before $N_{\mathcal{F}}(x_1, \dots, x_n) = \{(f(x_1), \dots, f(x_n)) \in \{0, 1\}^n, f \in \mathcal{F}\}$. In words $\mathcal{F}(x_1, \dots, x_n)$ is the smallest subset of \mathcal{F} that gives rise to all the different prediction rules for the data $(x_1, y_1), \dots, (x_n, y_n)$, therefore $|\mathcal{F}(x_1, \dots, x_n)| \leq \mathcal{S}(\mathcal{F}, n)$.

We are essentially ready to apply our union bound.

$$\begin{aligned}
& P\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(x_i) \neq y_i\} \right| > \frac{\epsilon}{4}\right) \\
&= P\left(\max_{f \in \mathcal{F}(x_1, \dots, x_n)} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(x_i) \neq y_i\} \right| > \frac{\epsilon}{4}\right) \\
&= P\left(\bigcup_{f \in \mathcal{F}(x_1, \dots, x_n)} \left\{ \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(x_i) \neq y_i\} \right| > \frac{\epsilon}{4} \right\}\right) \\
&\leq \sum_{f \in \mathcal{F}(x_1, \dots, x_n)} P\left(\frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(x_i) \neq y_i\} \right| > \frac{\epsilon}{4}\right) \\
&\leq |\mathcal{F}(x_1, \dots, x_n)| \sup_{f \in \mathcal{F}(x_1, \dots, x_n)} P\left(\frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(x_i) \neq y_i\} \right| > \frac{\epsilon}{4}\right) \\
&\leq \mathcal{S}(\mathcal{F}, n) \sup_{f \in \mathcal{F}(x_1, \dots, x_n)} P\left(\frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(x_i) \neq y_i\} \right| > \frac{\epsilon}{4}\right) \\
&\leq \mathcal{S}(\mathcal{F}, n) \sup_{f \in \mathcal{F}} P\left(\frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(x_i) \neq y_i\} \right| > \frac{\epsilon}{4}\right).
\end{aligned}$$

Therefore we managed to “pull” the supremum outside the probability, and take advantage of the effective size of the class \mathcal{F} . We are one step away from concluding the proof.

Step 4: Hoeffding’s inequality:

Notice that

$$\frac{1}{n} \left| \sum_{i=1}^n \underbrace{\sigma_i \mathbf{1}\{f(x_i) \neq y_i\}}_{A_i} \right|$$

is the sum of the absolute value of n independent random variables A_i , with zero mean and bounded between -1 and 1 . We can therefore apply Hoeffding’s inequality.

$$\begin{aligned}
P\left(\frac{1}{n} \left| \sum_{i=1}^n A_i \right| > \epsilon/4\right) &\leq P\left(\left| \sum_{i=1}^n A_i \right| > n\epsilon/4\right) \\
&\leq 2e^{-\frac{2(n\epsilon/4)^2}{\sum_{i=1}^n (\max_i A_i - \min_i A_i)^2}} \\
&\leq 2e^{-\frac{n^2 \epsilon^2 / 8}{4n}} \\
&\leq 2e^{-\frac{n\epsilon^2}{32}}.
\end{aligned}$$

It's time to revisit (2). Let's look at the right-hand-side

$$\begin{aligned}
& P\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X_i) \neq Y_i\} \right| > \frac{\epsilon}{4}\right) \\
&= E\left[\mathbf{1}\left\{\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X_i) \neq Y_i\} \right| > \frac{\epsilon}{4}\right\}\right] \\
&= E\left[E\left[\mathbf{1}\left\{\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X_i) \neq Y_i\} \right| > \frac{\epsilon}{4}\right\} \middle| D_n\right]\right] \\
&= E\left[P\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X_i) \neq Y_i\} \right| > \frac{\epsilon}{4} \middle| D_n\right)\right] \\
&\leq E\left[\mathcal{S}(\mathcal{F}, n) \sup_{f \in \mathcal{F}} P\left(\frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X_i) \neq Y_i\} \right| > \frac{\epsilon}{4} \middle| D_n\right)\right] \\
&\leq \mathcal{S}(\mathcal{F}, n) E\left[\sup_{f \in \mathcal{F}} P\left(\frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X_i) \neq Y_i\} \right| > \frac{\epsilon}{4} \middle| D_n\right)\right] \\
&\leq \mathcal{S}(\mathcal{F}, n) E\left[2e^{-\frac{n\epsilon^2}{32}} \middle| D_n\right] \\
&= 2\mathcal{S}(\mathcal{F}, n)e^{-\frac{n\epsilon^2}{32}}.
\end{aligned}$$

Using this in step (2) yields the desired result

$$\begin{aligned}
& P\left(\sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| > \epsilon\right) \\
&\leq 4P\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \mathbf{1}\{f(X_i) \neq Y_i\} \right| > \frac{\epsilon}{4}\right) \\
&\leq 8\mathcal{S}(\mathcal{F}, n)e^{-\frac{n\epsilon^2}{32}},
\end{aligned}$$

concluding the proof. ■