

ELEN6887

Lecture 20: Applications of the Vapnik-Chervonenkis Theory

R. Castro

4/29/2009

In the last lecture we have proved the VC inequality

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 did also showed that if we choose a model using empirical risk minimization

$$\hat{f}_n = \arg \min_{f \in \mathcal{F}} \hat{R}_n(f) ,$$

then

$$\begin{aligned} E \left[R(\hat{f}_n) \right] &\leq \inf_{f \in \mathcal{F}} R(f) + 4\sqrt{\frac{\log \mathcal{S}(\mathcal{F}, n) + \log 2}{n}} \\ &\leq \inf_{f \in \mathcal{F}} R(f) + 4\sqrt{\frac{VC(\mathcal{F}) \log(n+1) + \log 2}{n}} . \end{aligned}$$

We will now use these results in a variety of settings, in some cases generalizing ideas we had seen before.

1 Linear Classifiers

Let $\mathcal{X} = \mathbf{R}^d$ and \mathcal{F} be the class of linear classifiers (hyperplane classifiers). Formally

$$\mathcal{F} = \{ f(x) = \mathbf{1}\{\mathbf{w}^T \mathbf{x} + w_0 > 0\} ; \mathbf{x}, \mathbf{w} \in \mathbf{R}^d, w_0 \in \mathbf{R} \} .$$

This class is illustrated in Figure 1.

Consider $d = 2$. It is clear to see that this class can shatter at least 3 points, that is, it is possible to obtain all possible labelings for a set of three points (provided these are not colinear). Figure 2(a) illustrates this. Trivially then $\mathcal{S}(\mathcal{F}, n) = 2^n$ for all $n \leq 3$. Now for four points the picture changes. There is no possible arrangement of four points such that we can obtain all possible $2^4 = 16$ distinct labelings. Figure 2(b) illustrates this. Therefore $\mathcal{S}(\mathcal{F}, 4) < 2^4 = 16$, and so $\mathcal{S}(\mathcal{F}, n) < 2^n$ for $n > 3$. We therefore conclude that the VC dimension of \mathcal{F} for $d = 2$ is $VC(\mathcal{F}) = 3$.

Actually the result generalizes for all d , and we have that $VC(\mathcal{F}) = d + 1$. Proving the result directly is rather tricky and cumbersome, but the following result allows us to get a tight upper bound on the dimension.

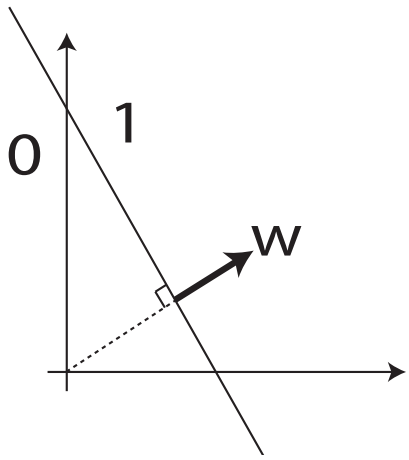


Figure 1: Hyperplane classifiers in two dimensions.

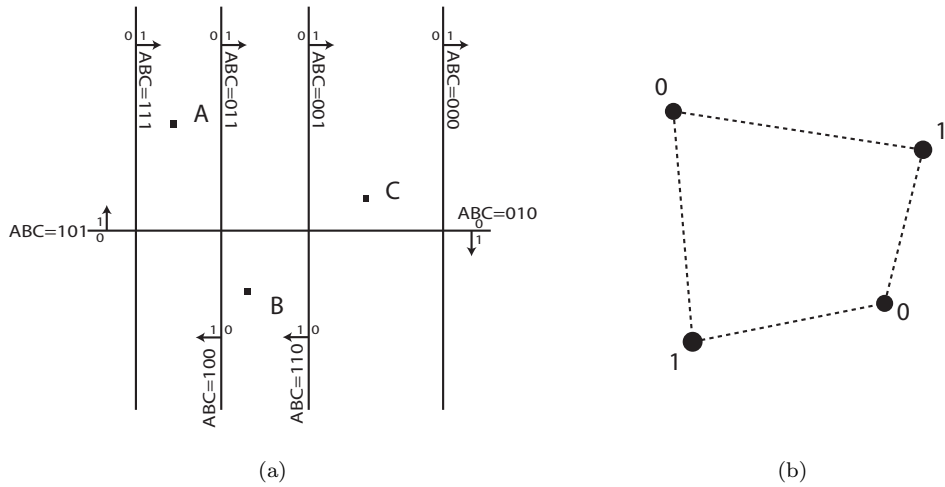


Figure 2: Shattering n points using hyperplanes when $d = 2$: (a) The class \mathcal{F} can shatter $n = 3$ points (in the figure you see all the hyperplanes and corresponding labelings of the points - there is a total of $2^3 = 8$ labelings); (b) Four points can never be shattered. The labeling in the figure cannot be obtained.

Theorem 2 (Steele '75, Dudley '78). Let \mathcal{G} be a vector space of real-valued functions on \mathbf{R}^d , with dimension $\dim(\mathcal{G})$. The class of rules $\mathcal{F} = \{f(x) = \mathbf{1}\{g(x) \geq 0\}; g \in \mathcal{G}\}$ has $VC(\mathcal{F})$.

This immediately implies that, for hyperplane classifiers $VC(\mathcal{F}) \leq d+1$. It turns out that a more careful argument shows the bound is actually tight. We can therefore apply the generalization bounds we derived before in this setting. Let $\mathcal{X} \in \mathbf{R}^d$ and \mathcal{F} be the class of hyperplane classifiers. If

$$\hat{f}_n = \arg \min_{f \in \mathcal{F}} \hat{R}_n(f) ,$$

then

$$E \left[R(\hat{f}_n) \right] \leq \inf_{f \in \mathcal{F}} R(f) + 4 \sqrt{\frac{(d+1) \log(n+1) + \log 2}{n}} .$$

2 Generalized Linear Classifiers

Normally, we have a feature vector $X \in \mathbf{R}^d$. A hyperplane in \mathbf{R}^d provides a linear classifier in \mathbf{R}^d . Although appealing linear classifiers are a little limited. It turns out nonlinear classifiers can be obtained using a straightforward generalization.

Let $\phi_1, \dots, \phi_{d'}$, $d' \geq d$ be a collection of functions mapping \mathbf{R}^d to \mathbf{R} . These functions, applied to a feature vector $X \in \mathbf{R}^d$ produce a higher dimensional feature $\phi(X) = (\phi_1(X), \phi_2(X), \dots, \phi_{d'}(X))^T \in \mathbf{R}^{d'}$. For example, if $X = (x_1, x_2)^T$ then we could consider $d' = 5$ and $\phi = (x_1, x_2, x_1 x_2, x_1^2, x_2^2)^T \in \mathbf{R}^5$. We can then construct a linear classifier in the higher dimensional generalized feature space $\mathbf{R}^{d'}$. In other words, we use an hyperplane classifier to fit the dataset

$$(\phi(X_1), Y_1), \dots, (\phi(X_n), Y_n) .$$

The VC bounds immediately extend to this case, and we have for

$$\mathcal{F}' = \{f(x) = \mathbf{1}\{w^T \phi(x) + w_0 > 0\}; w \in \mathbf{R}^{d'}, w_0 \in \mathbf{R}\} ,$$

$$E[R(\hat{f}_n)] - \inf_{f \in \mathcal{F}'} R(f) \leq 4 \sqrt{\frac{(d'+1) \log(n+1) + \log 2}{n}} .$$

Note that for the case of the example above (where $d = 2$ and $d' = 5$) the class \mathcal{F}' consists of classification rules where the decision boundary is given by conical curves (e.g., ellipses, hyperboles, etc.).

3 Decision Trees

Consider the following subset of the hyperplane classifiers:

$$\mathcal{F} = \{f(x) = \mathbf{1}\{x_i > b\} \text{ or } f(x) = \mathbf{1}\{x_i < b\} : i = 1, \dots, d, b \in \mathbf{R}\} .$$

Note these are hyperplane classifiers that are aligned with the coordinate axis.

Perhaps surprisingly the VC dimension of this class is the same as the class of all hyperplanes, that is $VC(\mathcal{F}) = d+1$, although it is clear that an element of \mathcal{F} can be described essentially by two parameters: b and which coordinate axis are perpendicular too. This illustrates that the VC dimension might be large even for simple classes.

These kind of aligned hyperplanes can be used recursively, to create a very general class of tree classifiers. Let $k \geq 1$ and define the class of tree classifiers

$$\mathcal{T}_k = \{\text{classifiers based on recursive rectangular partitions of } \mathbf{R}^d \text{ with } k+1 \text{ cells}\} .$$

Let's illustrate what are the elements of \mathcal{T}_k . Let $T \in \mathcal{T}_k$. Each cell of T results from splitting a rectangular region into two smaller rectangles parallel to one of the coordinate axes. This process is illustrated in Figure 3.

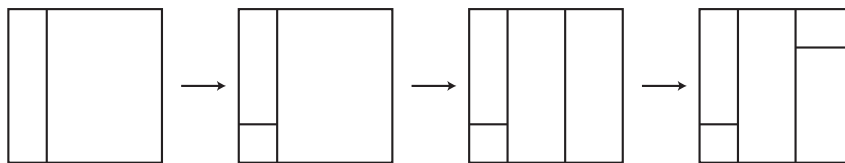


Figure 3: Growing trees.

Each additional split is analogous to a half-space set. Therefore each additional split can potentially shatter $d + 1$ points. This implies that

$$V_{\mathcal{T}_k} \leq (d + 1)k .$$

This bound is not completely tight but it is more than enough for our purposes. Notice also that if we don't restrict the number of cells in the tree then the VC dimension of the class is $VC(\mathcal{T}_\infty) = \infty$ since for a set of n points we can always construct a recursive rectangular partition that has exactly one point per cell.

Let's now apply the VC bounds to trees. Let

$$\hat{f}_n = \arg \min_{f \in \mathcal{T}_k} \hat{R}_n(f) ,$$

then

$$E \left[R(\hat{f}_n) \right] - R^* \leq \inf_{f \in \mathcal{T}_k} R(f) - R^* + 4 \sqrt{\frac{k(d + 1) \log(n + 1) + \log 2}{n}} .$$

Clearly if we take k large we can make $\inf_{f \in \mathcal{T}_k} R(f) - R^*$ small, but this will make the estimation error (bounded by the square-root term) large. How can we decide what dimension to choose for a generalized linear classifier? How many leaves should be used for a classification tree? The answer is complexity Regularization using VC bounds! The approach is in all identical to what we have done before, but now we don't need to restrict ourselves to finite classes of models anymore.

4 Structural Risk Minimization (SRM)

SRM is simply complexity regularization using VC type bounds in place of the Hoeffding inequality we used before. Let's derive a simple version of a SRM result.

Assume you have a sequence of sets of classifiers

$$\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_k, \dots ,$$

of increasing VC dimension

$$VC(\mathcal{F}_1) \leq VC(\mathcal{F}_2) \leq \dots .$$

In the typically cases we might consider the inequalities above are strict, that is $VC(\mathcal{F}_k) < VC(\mathcal{F}_{k+1})$ for all $k \geq 1$. For each $k = 1, 2, \dots$ we find the minimum empirical risk classifier

$$\hat{f}_n^{(k)} = \arg \min_{f \in \mathcal{F}_k} \hat{R}_n(f)$$

and then select the final classifier according to

$$\hat{k} = \arg \min_{k \geq 1} \left\{ \hat{R}_n(\hat{f}_n^{(k)}) + \text{penalty}(k, n) \right\} .$$

Finally we take $\hat{f}_n \equiv \hat{f}_n^{(\hat{k})}$.

The basic rationale is to choose $\text{penalty}(k, n)$ as an increasing function of k , so that we choose models that are simple, but also “explain” the training data well. The choice of the penalty function can be made quite obvious by looking at our VC bounds.

Begin by writing the result of Theorem 1 in a slightly different way. Let $\delta_k > 0$. Then

$$P \left(\sup_{f \in \mathcal{F}_k} \left| \hat{R}_n(f) - R(f) \right| > 8 \sqrt{\frac{\mathcal{S}(\mathcal{F}_k, n) + \log 8 + \log(1/\delta_k)}{2n}} \right) \leq \delta_k .$$

Sauer’s lemma implies that

$$P \left(\sup_{f \in \mathcal{F}_k} \left| \hat{R}_n(f) - R(f) \right| > 8 \sqrt{\frac{VC(\mathcal{F}_k) \log(n+1) + 3 + \log(1/\delta_k)}{2n}} \right) \leq \delta_k .$$

Now define $\mathcal{F} = \bigcup_{k=1}^{\infty} \mathcal{F}_k$ and take $\delta_k = \delta 2^{-k}$. Let’s see how big is the difference between empirical and true risk for ANY element of \mathcal{F} . In other words, we want to see how big can

$$\left| \hat{R}_n(f) - R(f) \right| ,$$

be for any $f \in \mathcal{F}$.

Let $k(f)$ be the smallest integer k such that $f \in \mathcal{F}_k$. Now note that

$$\begin{aligned} & P \left(\sup_{f \in \mathcal{F}} \left\{ \left| \hat{R}_n(f) - R(f) \right| - 8 \sqrt{\frac{VC(\mathcal{F}_{k(f)}) \log(n+1) + 3 + \log(1/\delta_{k(f)})}{2n}} \right\} > 0 \right) \\ &= P \left(\sup_k \left\{ \sup_{f \in \mathcal{F}_k} \left\{ \left| \hat{R}_n(f) - R(f) \right| - 8 \sqrt{\frac{VC(\mathcal{F}_k) \log(n+1) + 3 + \log(1/\delta_k)}{2n}} \right\} \right\} > 0 \right) \\ &= P \left(\bigcup_{k=1}^{\infty} \left\{ \sup_{f \in \mathcal{F}_k} \left\{ \left| \hat{R}_n(f) - R(f) \right| - 8 \sqrt{\frac{VC(\mathcal{F}_k) \log(n+1) + 3 + \log(1/\delta_k)}{2n}} \right\} > 0 \right\} \right) \\ &= P \left(\bigcup_{k=1}^{\infty} \left\{ \sup_{f \in \mathcal{F}_k} \left| \hat{R}_n(f) - R(f) \right| > 8 \sqrt{\frac{VC(\mathcal{F}_k) \log(n+1) + 3 + \log(1/\delta_k)}{2n}} \right\} \right) \\ &\leq \sum_{k=1}^{\infty} P \left(\sup_{f \in \mathcal{F}_k} \left| \hat{R}_n(f) - R(f) \right| > 8 \sqrt{\frac{VC(\mathcal{F}_k) \log(n+1) + 3 + \log(1/\delta_k)}{2n}} \right) \\ &\leq \sum_{k=1}^{\infty} \delta_k = \sum_{k=1}^{\infty} \delta 2^{-k} = \delta , \end{aligned}$$

where the first inequality is just a consequence of the union of events bound. In other words, with probability at least $1 - \delta$

$$\forall f \in \mathcal{F}, \quad \left| \hat{R}_n(f) - R(f) \right| < \underbrace{8 \sqrt{\frac{VC(\mathcal{F}_{k(f)}) \log(n+1) + 3 + \log(1/\delta_{k(f)})}{2n}}}_{C(f, n, \delta)} .$$

Note that

$$C(f, n, \delta) = 8 \sqrt{\frac{VC(\mathcal{F}_{k(f)}) \log(n+1) + 3 + k(f) \log 2 + \log(1/\delta)}{2n}} .$$

We are now under the setting of Lecture 10, and it seems quite sensible to take

$$\hat{f}_n \equiv \arg \min_{f \in \mathcal{F}} \hat{R}_n(f) + C(f, n, \delta) .$$

Using the results of Lecture 10 we get immediately that

$$E[R(\hat{f}_n)] - R^* \leq \inf_{f \in \mathcal{F}} \{R(f) - R^* + C(f, n, \delta)\} + \delta ,$$

so taking $\delta = 1/\sqrt{n+1}$ is a very sensible choice, yielding the bound.

$$E[R(\hat{f}_n)] - R^* \leq \inf_{f \in \mathcal{F}} \left\{ R(f) - R^* + 8\sqrt{\frac{(VC(\mathcal{F}_{k(f)}) + 1/2) \log(n+1) + 3 + k(f) \log 2}{2n}} \right\} + \frac{1}{\sqrt{n+1}} .$$

5 Application to Trees

Let $\mathcal{T}_1, \mathcal{T}_2, \dots$ be the classes of decision trees with $k+1$ leafs. Then $VC(\mathcal{T}_k) \leq k(d+1)$. Now define

$$\hat{f}_n^{(k)} = \arg \min_{f \in \mathcal{T}_k} \hat{R}_n(f)$$

and then select the final classifier according to

$$\hat{k} = \arg \min_{k \geq 1} \left\{ \hat{R}_n(\hat{f}_n^{(k)}) + 8\sqrt{\frac{(k(d+1) + 1/2) \log(n+1) + 3 + k \log 2}{2n}} \right\} .$$

Finally we take $\hat{f}_n \equiv \hat{f}_n^{(\hat{k})}$. This yields the bound

$$E[R(\hat{f}_n)] - R^* \leq \inf_k \left\{ \inf_{f \in \mathcal{F}_k} \left\{ R(f) - R^* + 8\sqrt{\frac{(k(d+1) + 1/2) \log(n+1) + 3 + k \log 2}{2n}} \right\} \right\} + \frac{1}{\sqrt{n+1}} .$$

Compare this with the bound of Lecture 11, where we considered recursive dyadic partitions with a much more stringent structure. You see that bound is has essentially the same form, although we are now considering a much richer class of classification rules, that is uncountable. These kinds of trees are used frequently for classification and regression (see for example the methods under the name of CART), and are quite useful in practice.