ECE 901
Lecture 18: Introduction to Vapnik-Chervonenkis (VC) Theory

R. Nowak

5/17/2009

In our past lectures we have derived a number of generalization bounds, but all these had one thing in common: they required the class of candidate models to be either finite or countable. In other words the class of candidate models needed to be enumerable. We required this condition since for all the bounds derived we used a union bound, and the only way that bound is non-trivial is if the set of events involved is enumerable.

In many cases of practical importance the collection of candidate models is uncountably infinite (generally it is easier to formulate the optimization problem of finding a good model in this setting too). Let's see a motivational example, that will help us understand what are the possible ways of dealing with this issue.

Example 1. *Let the feature space be $\mathcal{X} = \mathbf{R}$ and the label space be $\mathcal{Y} = \{0, 1\}$. Consider the following class of models*

$$\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y} \ : f(x) = \mathbf{1}\{x > t\},\ t \in \mathbf{R}\} \bigcup \{f : \mathcal{X} \to \mathcal{Y} \ : f(x) = \mathbf{1}\{x < t\},\ t \in \mathbf{R}\} \ .$$

Clearly this class is uncountable, therefore we cannot apply our current generalization bounds.

1 Two Ways to Proceed

Discretize or Quantize the Collection $\mathcal{F}$: This is the technique we have used in our previous lectures for the regression problems (you also have looked at this technique for classification in a homework problem). Let's revisit our example above.

Example 2. *Let's consider a discretized version of $\mathcal{F}$. Let's assume that $\mathcal{X} = [0, 1]$ and also that $P_X(A) \leq c|A|$ for all measurable sets $A \subseteq \mathcal{X}$. Define*

$$\mathcal{F}_Q = \{f : \mathcal{X} \to \mathcal{Y} \ : f(x) = \mathbf{1}\{x > t\},\ t \in \mathcal{T}_Q\} \bigcup \{f : \mathcal{X} \to \mathcal{Y} \ : f(x) = \mathbf{1}\{x < t\},\ t \in \mathcal{T}_Q\} \ ,$$

where $\mathcal{T} = \left\{0, \frac{1}{Q}, \frac{2}{Q}, \ldots, \frac{Q}{Q}\right\}$. It's not hard to show that for any $f \in \mathcal{F}$ there exists an $f' \in \mathcal{F}_Q$ such that

$$|R(f) - R(f')| \leq \frac{c}{Q} \ ,$$

therefore if we choose Q large enough then $\mathcal{F}_Q$ and $\mathcal{F}$ are nearly equivalent.

This trick can be used in many situations, but often leads into practical difficulties when trying to implement the estimators, besides making their analysis a bit more messy.

Identical Empirical Errors: Since we are choosing a model based on a set of training data it makes sense to look at the complexity of the model class with respect to ANY set of training data with n samples. It turns out that, in many settings, this provides a notion of "effective size" of the model class, and allows us to again get generalization bounds. This is the key necessary for the developments below.

2 Vapnik-Chervonenkis Theory

For the remainder of the course we will consider only the binary classification setting. However, the ideas presented here can be generalized to other settings as well. Let $\mathcal{X}$ denote the feature space (e.g., $\mathcal{X} = \mathbf{R}^d$), and $\mathcal{Y} = \{0, 1\}$ denote the label space. Let $f : \mathcal{X} \to \mathcal{Y}$ be a classifier (also called predictor or model). Let $(X, Y) \sim P_{XY}$ where the joint probability distribution 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,

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

where $(X_i, Y_i) \overset{i.i.d.}{\sim} P_{XY}$.

Because there are only two possible labels for each feature vector each model in $\mathcal{F}$ can will give rise to a labeling sequence

$$(f(X_1), \ldots, f(X_n)) \in \{0, 1\}^n \ .$$

For a given training set D_n there are at most 2^n distinct such sequences, but often much less. Let $\mathcal{S}(\mathcal{F}, n)$ be the maximum number of labeling sequences the class $\mathcal{F}$ induces over n training points in $\mathcal{X}$. Formally let $x_1, \ldots, x_n \in \mathcal{X}$ and define

$$N_{\mathcal{F}}(x_1, \ldots, x_n) = \{(f(x_1), \ldots, f(x_n)) \in \{0, 1\}^n, \ f \in \mathcal{F}\} \ .$$

Definition 1. *The **shatter coefficient** of class $\mathcal{F}$ is defined as*

$$\mathcal{S}(\mathcal{F}, n) = \max_{x_1, \ldots, x_n \in \mathcal{X}} |N_{\mathcal{F}}(x_1, \ldots, x_n)| \ ,$$

where $|\cdot|$ denotes the number of elements in the set.

Clearly $\mathcal{S}(\mathcal{F}, n) \leq 2^n$, but often it is much smaller. Let's see some examples.

Example 3. *Let's revisit example 1. Suppose we have n feature vectors $(x_1, \ldots, x_n) \in \mathbf{R}$. Assume there are no identical points (otherwise the number of possible labelings is even smaller). Any classifier in $\mathcal{F}$ labels all points to the left of a number $t \in [0, 1]$ as "1" or "0", and points to the right as "0" or "1", respectively. For $t \in [0, x_1)$, all points are either labelled "0" or "1". For $t \in (x_1, x_2)$, x_1 is labelled "0" or "1" and $x_2 \ldots x_n$ are labeled "1" or "0" and so on. We see that there are exactly $2n$ different possible labelings, therefore $\mathcal{S}(\mathcal{F}, n) = 2n$. This is far less than the bound $\mathcal{S}(\mathcal{F}, n) \leq 2^n$.*

The number of different labelings that a class $\mathcal{F}$ can produce on a set of n training data is a measure of the "effective size" of $\mathcal{F}$. It is possible to define a meaningful dimension concept based the behavior of $\log \mathcal{S}(\mathcal{F}, n)$.

Definition 2. *The Vapnik-Chervonenkis (VC) dimension is defined as the largest integer k such that $\mathcal{S}(\mathcal{F}, k) = 2^k$. The VC dimension of a class $\mathcal{F}$ is denoted by $VC(\mathcal{F})$.*

Note that the VC dimension is not a function of the number of training data. We have the following result, presented here without a proof.

Lemma 1. *Sauer's Lemma:*

$$S(\mathcal{F}, n) \leq (n + 1)^{VC(\mathcal{F})} \ .$$

So, for example 1 we see that $VC(\mathcal{F}) = 2$. Let's see another example

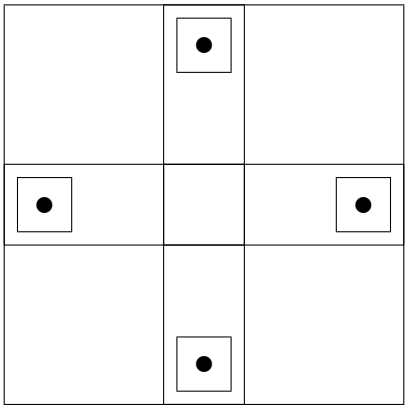

Figure 1: Labeling four feature vectors. If the points are not co-linear then we can obtain all possible 2^n labelings.

Example 4. *Let* $\mathcal{X} = \mathbf{R}^2$ *and define*

$$\mathcal{F} = \{f(x) = \mathbf{1}\{x \in A\} : \ A = [a,b] \times [c,d], a, b, c, d \in \mathbf{R}\} \ .$$

In words, the classifiers in $\mathcal{F}$ *label all the points inside a rectangle* $[a,b] \times [c,d]$ *one, and all the other points zero. Let's see what happens when* $n = 4$*. Figure 1 illustrates this when the points are not co-linear. We see that we can obtain all the possible labelings, and so* $\mathcal{S}(\mathcal{F}, 4) = 16$*. Clearly for* $n \leq 4$ *we have also* $\mathcal{S}(\mathcal{F}, n) = 2^n$*. Now if we have* $n = 5$ *things change a bit. Figure 2 illustrates this. If we have five points there is always one that stays "in the middle" of all the others, and this one cannot have a label different then all the others, therefore* $\mathcal{S}(\mathcal{F}, n) < 2^n$*. We immediately conclude that the VC dimension of this class is* $VC(\mathcal{F}) = 4$*.*

We will see more examples in the next lectures. For now let's see what kinds of results we expect to get using this approach.

3 The Shatter Coefficient and the Effective Size of a Model Class

As commented before the shatter coefficient measures the "effective" size of the class $\mathcal{F}$ when looked through the lens of n training points. Recall the generalization bound we have shown when $\mathcal{F}$ is finite

$$\begin{aligned} P(\forall f \in \mathcal{F} \quad |\hat{R}_n(f) - R(f)| > \epsilon) &= P\left(\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| > \epsilon\right) \\ &\leq 2|\mathcal{F}|e^{-2n\epsilon^2} \ . \end{aligned}$$

Our hope is that we can get something quite similar, but where $|\mathcal{F}|$ is replaced by $\mathcal{S}(\mathcal{F}, n)$. This is indeed the case, and in the next lecture we will prove the following very important result.

Theorem 1. ***The VC inequality:***

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

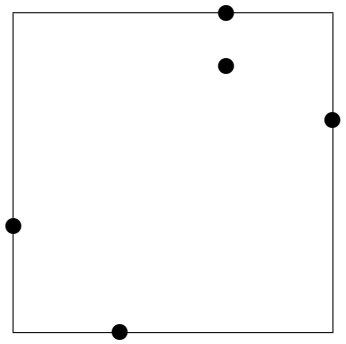

Figure 2: Labeling five feature vectors. Without loss of generality there is a point inside the convex hull of all the points, that cannot have a label different than all the others, therefore $\mathcal{S}(\mathcal{F}, n) < 2^n$ for $n \geq 5$.

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 in the theorem.

The second inequality, together with Sauer's lemma, gives rise to the following important corollary that characterizes the performance of the empirical risk minimization rule.

Corollary 1. *Let*

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

be the empirical risk minimizer (if more than one possibility is available just choose one of the possibilities). Then

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

The result of the corollary is quite similar to what we have seen before, but now it applies also to uncountable classes of models.

Proof: Note that

$$\begin{aligned}
R(\hat{f}_n) - \inf_{f\in\mathcal{F}} R(f) &= R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) + \hat{R}_n(\hat{f}_n) - \inf_{f\in\mathcal{F}} R(f) \\
&= \Big(R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)\Big) + \sup_{f\in\mathcal{F}} \Big(\hat{R}_n(\hat{f}_n) - R(f)\Big) \\
&\leq \Big(R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)\Big) + \sup_{f\in\mathcal{F}} \Big(\hat{R}_n(f) - R(f)\Big) \\
&\leq \Big|R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)\Big| + \sup_{f\in\mathcal{F}} \Big|\hat{R}_n(f) - R(f)\Big| \\
&\leq \sup_{f\in\mathcal{F}} \Big|R(f) - \hat{R}_n(f)\Big| + \sup_{f\in\mathcal{F}} \Big|\hat{R}_n(f) - R(f)\Big| \\
&= 2\sup_{f\in\mathcal{F}} \Big|\hat{R}_n(f) - R(f)\Big|
\end{aligned}$$

where the first inequality follows from the definition of the empirical risk minimizer, which implies that $\hat{R}_n(\hat{f}_n) \leq \hat{R}_n(f)$ for all $f \in \mathcal{F}$. Taking the expectation of both sides of the inequality and using the Theorem yields the first part of the result. The second part of the result follows immediately from Sauer's lemma. ■