ECE901(spring 2004): Statistical learning Theory Lecture:19

The Vapnik-Chervonenkis Inequality

Lecturer: Robert Nowak Scribe: Sirin Nitinawarat

The Vapnik-Chervonenkis Inequality
The VC inequality is a powerful generalization of the bounds we obtained for the hyperplane classifier in the previous lecture. The basic idea of the proof is quite similar. Before starting the inequality, we need to introduce the concept of shatter coefficients and VC dimension.

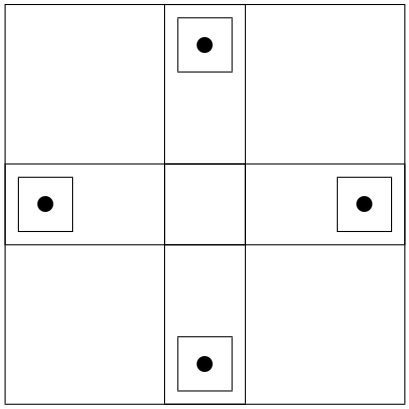
Shatter Coefficients
Let $\mathcal{A}$ be a collection of subsets of $\mathcal{R}^d$,
definition: The n^{th} shatter coefficient of $\mathcal{A}$ is defined by

$$\mathcal{S}_{\mathcal{A}}(n) = \max_{x_1,\ldots,x_n \,\epsilon\, \mathcal{R}^d} \left|\{\{x_1,\ldots,x_n\} \bigcap A, A \,\epsilon\, \mathcal{A}\}\right|$$

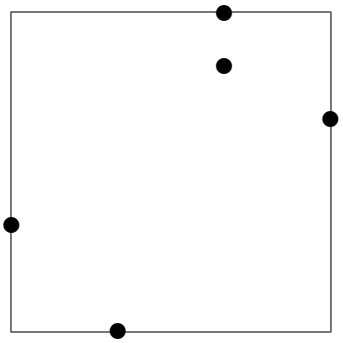
The shatter coefficients are a measure of the richness of the collection $\mathcal{A}$. $\mathcal{S}_{\mathcal{A}}(n)$ is the largest number of different subsets of a set of n points that can be generated by intersecting the set with elements of $\mathcal{A}$.

example: In 1-d, Let $\mathcal{A} = \{(-\infty, t], t \,\epsilon\, \mathcal{R}\}$
Possible subsets of $\{x_1,\ldots,x_n\}$ generated by intersecting with sets of the form $(-\infty, t]$ are $\{x_1,\ldots,x_n\}$, $\{x_1, \ldots, x_{n-1}\}$, $\ldots$, $\{x_1\}$, ϕ.
Hence $\mathcal{S}_d(n) = n + 1$.

example: In 2-d, Let $\mathcal{A} =$ { all rectangles in $\mathcal{R}^2$ }



Consider a set $\{x_1, x_2, x_3, x_4\}$ of training points. If we arrange the four points into the corner of a diamond shape. It's easy to see that we can find a rectangle in $\mathcal{R}^2$ to cover any subsets of the four points as the above picture, i.e. $\mathcal{S}_{\mathcal{A}}(4) = 2^4 = 16$.
Clearly, $\mathcal{S}_{\mathcal{A}}(n) = 2^n, n = 1, 2, 3$ as well.



However, for $n = 5, \mathcal{S}_{\mathcal{A}}(n) < 2^5$. This is because we can always select four points such that the rectangle, which just contains four of them, contains

the other point. Consequently, we cannot find a rectangle classifier which contains the four outer points and does not contain the inner point as shown above.

Note the $\mathcal{S}_{\mathcal{A}} \leq 2^n$.
If $|\{\{x_1, \ldots, x_n\} \bigcap A, A \,\epsilon\, \mathcal{A}\}| = 2^n$ then we say that $\mathcal{A}$ shatters $x_1, \ldots, x_n$.

VC Dimension
definition: The VC dimension $V_{\mathcal{A}}$ of a collection of sets $\mathcal{A}$ is defined as the largest interger n such that $S_{\mathcal{A}}(n) = 2^n$.

example: $\mathcal{A} = \{(-\infty, t]\ ; t \,\epsilon\, \mathcal{R}\}, \mathcal{S}_{\mathcal{A}} = n + 1$ hence $\mathcal{V}_{\mathcal{A}} = 1$.

example: $\mathcal{A} = \{$ all rectangles in $\mathcal{R}^2\ \}$
$\mathcal{S}_{\mathcal{A}} = 2^n, n = 1, 2, 3, 4$ and $\mathcal{S}_{\mathcal{A}} \leq 2^n, n = 4$, Hence $V_{\mathcal{A}} = 4$.

The VC dimension provides a useful bound on the growth of the shatter coefficients.

Sauer's Lemma: Let $\mathcal{A}$ be a collection of set with VC dimension $V_{\mathcal{A}} < \infty$. Then $\forall n, \mathcal{S}_{\mathcal{A}}(n) \leq \sum_{i=0}^{V_{\mathcal{A}}} \begin{pmatrix} n \\ i \end{pmatrix}$, also $\mathcal{S}_{\mathcal{A}}(n) \leq (n+1)^{V_{\mathcal{A}}}, \forall n$.

VC Dimension and Classifiers
Let $\mathcal{F}$ be a collection of classifiers of the form $f : \mathcal{R}^d \to \{0, 1\}$
Define $\mathcal{A} = \{\{x : f(x) = 1\} \times \{0\} \bigcup \{x : f(x) = 0\} \times \{1\}, f \,\epsilon\, \mathcal{F}\}$
In words, this is collection of subsets of $\mathcal{X} \times \mathcal{Y}$ for which on $f \epsilon \mathcal{F}$ maps the features x to a label opposite of y. The size of $\mathcal{A}$ expresses the richness of $\mathcal{F}$. The larger $\mathcal{A}$ is the more likely it is that there exists an $f \epsilon \mathcal{F}$ for which $R(f) = P(f(X) \neq Y)$ is close to the Bayes risk $R^* = P(f^*(X) \neq Y)$ where f^* is the Bayes classifier. The n^{th} shatter coefficient of $\mathcal{F}$ is defined as $\mathcal{S}_{\mathcal{F}}(n) = \mathcal{S}_{\mathcal{A}}(n)$ and the VC dimesion of $\mathcal{F}$ is defined as $V_{\mathcal{F}} = V_{\mathcal{A}}$.

example: linear (hyperplane) classifiers in $\mathcal{R}^d$
Consider $d = 2$. Let n be the number of trainning points, It is easy to see that when $n = 1$, Let $\mathcal{A}$ be as above. By using linear classifiers in $\mathcal{R}^2$, it is easy to see that we can assign 1 to all possible subsets $\{\{x_1\}, \phi\}$ and 0 to their complements. Hence $\mathcal{S}_{\mathcal{F}}(1) = 2$.

When $n = 2$, we can also assign 1 to all possible subsets $\{\{x_1, x_2\}, \{x_1\}, \{x_2\}, \phi\}$ and 0 to their complements, and vice versa. Hence $\mathcal{S}_{\mathcal{F}}(2) = 4 = 2^2$.

When $n = 3$, we can arrange arrange the point x_1, x_2, x_3(non-colinear) so that the set of linear classifiers shatters the three points, hence $\mathcal{S}_{\mathcal{F}}(3) = 8 = 2^3$

When $n = 4$, no matter where the points x_1, x_2, x_3, x_4 and what designated binary values y_1, y_2, y_3, y_4 are. It's clear that $\mathcal{A}$ does not shatter the four points. To see the claim, first observe that the four points will form a 4-gon (if the four points are co-linear, or if the three points are co-linear then clearly linear classifiers cannot shatter the points). The two points that belong to the same diagonal lines form 2 groups and no linear classifier can assign different values to the 2 groups. Hence $\mathcal{S}_{\mathcal{F}}(4) < 16 = 2^4$ and $V_{\mathcal{F}} = 3$.

We state here without proving it that in general the class of linear classifiers in $\mathcal{R}^d$ has $V_{\mathcal{F}} = d + 1$.

The VC Inequality
Let $X_1, , \ldots, X_n$ be i.i.d. $\mathcal{R}^d$-valued random variables. Denote the common distribution of $X_i, 1 \leq i \leq n$ by $\mu(A) = P(X_1 \, \epsilon \, A)$ for any subset $A \subset \mathcal{R}^d$. Similarly, define the empirical distribution $\mu_n(A) = \frac{1}{n} \sum_1^n 1_{\{X_i \epsilon A\}}$.

Thm(VC'71)
For any probablilty measure μ and collection of subsets $\mathcal{A}$, and for any $\epsilon > 0$.

$$P\left(\underset{A \epsilon \mathcal{A}}{sup} \; |\mu_n(A) - \mu(A)| > \epsilon \right) \leq 8 \mathcal{S}_{\mathcal{A}}(n) e^{-n\epsilon^2/32}$$

and

$$E\left[\underset{A \epsilon \mathcal{A}}{sup} \; |\mu_n(A) - \mu(A)| \right] \leq 2\sqrt{\frac{\log 2\mathcal{S}_{\mathcal{A}}(n)}{n}}$$

Before giving a proof to the theorem. We present a Corollary.
Corollary: Let $\mathcal{F}$ be a collection of classifiers of the form$f : \mathcal{R}^d \to \{0, 1\}$ with VC dimension $V_{\mathcal{F}} < \infty$, Let $R(f) = P(f(X) \neq Y)$ and $\hat{R}_n(f) =$

$\frac{1}{n}\sum_1^n 1_{\{f(X_i)\neq Y_i\}}$, where $X_i, Y_i, 1 \leq i \leq n$ are i.i.d. with joint distributionP_{XY}.

Define

$$\hat{f}_n = \underset{f\epsilon\mathcal{F}}{argmin}\ \hat{R}_n(f).$$

Then

$$E[R(\hat{f}_n)] - \underset{f\epsilon\mathcal{F}}{inf}\ R(f) \leq 4\sqrt{\frac{\mathcal{V}_\mathcal{F}\log(n+1)+\log 2}{n}}$$

Proof:
Let $\mathcal{A} = \{\{x : f(x) = 1\} \times \{0\} \bigcup \{x : f(x) = 0\} \times \{1\}, f\,\epsilon\,\mathcal{F}\}$
Note that

$$P(f(X) \neq Y) = P((X, Y)\,\epsilon\,A) := \mu(A)$$

where $A = \{x : f(x) = 1\} \times \{0\} \bigcup \{x : f(x) = 0\} \times \{1\}$
Similarly,

$$\frac{1}{n}\sum_1^n 1_{\{f(X_i)\,\neq\, Y_i\}} = \frac{1}{n}\sum_1^n 1_{\{(X_i,Y_i)\,\epsilon\,A\}} := \mu(A)$$

Therefore, according to the VC theorem.

$$\begin{aligned} E\left[\underset{f\,\epsilon\,\mathcal{F}}{sup}\ \left|\hat{R}_n(f) - R(f)\right|\right] = E\left[\underset{A\,\epsilon\,\mathcal{A}}{sup}\ |\mu_n(A) - \mu(A)|\right] &\leq 2\sqrt{\frac{\log 2\mathcal{S}_\mathcal{A}(n)}{n}} \\ &= 2\sqrt{\frac{\log 2\mathcal{S}_\mathcal{F}(n)}{n}} \end{aligned}$$

Since $V_\mathcal{F} < \infty, \mathcal{S}_\mathcal{F}(n) \leq (n+1)^{V_\mathcal{F}}$ and

$$E\left[\underset{f\epsilon\mathcal{F}}{sup}\ \left|\hat{R}_n(f) - R(f)\right|\right] \leq 2\sqrt{\frac{V_\mathcal{F}\log(n+1)+\log 2}{n}}$$

Next, note that

$$\begin{aligned} R\left(\hat{f}_n\right) - \underset{f\,\epsilon\,\mathcal{F}}{inf}\ R(f) &= \left[R\left(\hat{f}_n\right) - \hat{R}_n\left(\hat{f}_n\right)\right] + \left[\hat{R}_n\left(\hat{f}_n\right) - \underset{f\,\epsilon\,\mathcal{F}}{inf}\ R(f)\right] \\ &= \left[R\left(\hat{f}_n\right) - \hat{R}_n\left(\hat{f}_n\right)\right] + \left[\underset{f\,\epsilon\,\mathcal{F}}{sup}\ \left(\hat{R}_n(\hat{f}_n) - R(f)\right)\right] \end{aligned}$$

$$
\begin{aligned}
&\leq \left[R\left(\hat{f}_n\right) - \hat{R}_n\left(\hat{f}_n\right)\right] + \left[\sup_{f \,\epsilon\, \mathcal{F}} \left(\hat{R}_n(f) - R(f)\right)\right] \\
&\leq 2 \sup_{f \,\epsilon\, \mathcal{F}} \left|\hat{R}_n(f) - R(f)\right|
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
E\left[R\left(\hat{f}_n\right)\right] - \inf_{f \,\epsilon\, \mathcal{F}} R(f) &\leq 2E\left[\sup_{f \,\epsilon\, \mathcal{F}} \left|\hat{R}_n(f) - R(f)\right|\right] \\
&\leq 4\sqrt{\frac{V_{\mathcal{F}} \log(n+1) + \log 2}{n}} \qquad //
\end{aligned}
$$