ECE901 Spring 2004 Statistical Regularization and Learning Theory **Lecture: 20**

Applications of VC Bound

Lecturer: Rob Nowak *Scribe: Guodong Guo*

1 Linear Classifiers

Suppose $\mathcal{F}$= {linear classifiers in $\mathbf{R}^d$}, then we have

$$V_{\mathcal{F}} = d+1, \quad \hat{f}_n = \arg\min_{f\in\mathcal{F}} \hat{R}_n(f)$$

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

.

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 $\mathbf{R}^d$. Nonlinear classifiers can be obtained by a straightforward generalization.

Let $\phi_1, \cdots, \phi_{d'}$, $d' \geq d$ be a collection of functions mapping $\mathbf{R}^d \to \mathbf{R}$. These functions, applied to a feature $X \in \mathbf{R}^d$, produce a generalized set of features, $\phi = (\phi_1(X), \phi_2(X), \cdots, \phi_{d'}(X))'$. For example, if $X = (x_1, x_2)'$, then we could consider $d' = S$ and $\phi = (x_1, x_2, x_1x_2, x_1^2, x_2^2)' \in \mathbf{R}^5$. We can then construct a linear classifier in the higher dimensional generalized feature space $\mathbf{R}^{d'}$.

The VC bounds immediately extend to this case, and we have for $\mathcal{F}$' = { generalized linear classifiers based on maps $\phi : \mathbf{R}^d \to \mathbf{R}^{d'}$ },

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

.

3 Half-Space Classifiers

Theorem 1 (Steele '75, Dudley '78) *Let $\mathcal{G}$be a finite-dimensional vector space of real-valued functions on $\mathbf{R}^d$. The class of sets $\mathcal{A} = \{\{x : g(x) \geq 0\} : g \in \mathcal{G}\}$ has VC dimension $\geq$ dim($\mathcal{G}$).*

Proof: It is sufficient to show that no set of $n = dim(\mathcal{G}) + 1$ points can be shattered by $\mathcal{A}$. Take any n points and for each $g \in \mathcal{G}$, define the vector $V_g = (g(x_1), \cdots, g(x_n))$.

The set $\{V_g : g \in \mathcal{G}\}$ is a linear subspace of $\mathbf{R}^n$ of dimension $\leq$ dim $(\mathcal{G}) = n-1$. Therefore, there exists a non-zero vector $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbf{R}^n$ such that $\sum_{i=1}^n \alpha_i g(x_i) = 0$. We can assume that at least one of these α_i^S is negative (if all are positive, just negate the sum). We can then re-arrange this expression as $\sum_{i:\alpha_i\geq 0} \alpha_i g(x_i) = \sum_{i:\alpha_i<0} -\alpha_i g(x_i)$.

Now suppose that there exists a $g \in \mathcal{G}$ such that the set $\{x : g(x) \geq 0\}$ selects precisely the x_i^S on the left-hand side above. Then all terms on the left are non-negative and all the terms on the right are non-positive. Since α is non-zero, this is a contradiction. Therefore, $x_1, \cdots, x_n$ cannot be shattered by sets in $\{x : g(x) \geq 0\}$, $g \in \mathcal{G}$. ■

Example 1 *Consider half-spaces in* $\mathbf{R}^d$ *of the form* $\mathcal{A} = \{x \in \mathbf{R}^d : x_i \geq b, i \in \{1, \cdots, d\}, b \in R\}$. *Each half-space can be described by*

$$g(x) = [0, \cdots, 0, 1, 0, \cdots, 0] \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} - b$$

$$\implies dim(\mathcal{G}) = d + 1, \quad V_{\mathcal{A}} \leq d + 1$$

.

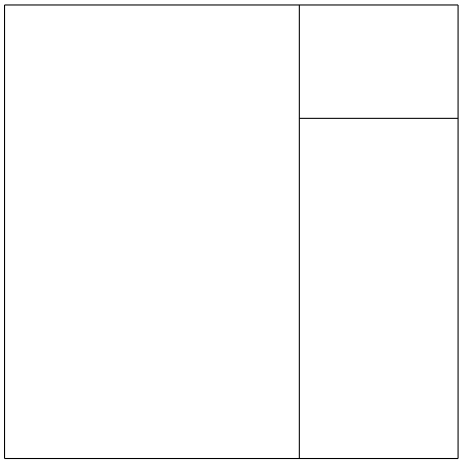
4 Tree Classifiers

Let

$$\mathcal{T}_k = \left\{ recursive \;\; rectangular \;\; partitions \;\; of \;\; \mathbf{R}^d \;\; with \;\; k+1 \;\; cells \right\}$$

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.

Example 2 $T \in \mathcal{T}_3$, $d = 2$.



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

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

.

Example 3 $d = 1$.
$k = 1$ *split shatters two points.*
$k = 2$ *splits shatters three points* < 4.

5 VC Bound for Tree Classifiers

$$\mathcal{F}_k = \{tree\ \ classifiers\ \ with\ \ k+1\ \ leafs\ \ on\ \ \mathbf{R}^d\}$$

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

.

How can we decide what dimension to choose for a generalized linear classifier?
How many leafs should be used for a classification tree?
Answer: Complexity Regularization using VC bounds!

6 Structural Risk Minimization (SRM)

SRM is simply complexity regularization using VC type bounds in place of Chernoff's bound or other concentration inequalities.

The basic idea is to consider a sequence of sets of classifiers $\mathcal{F}_1, \mathcal{F}_2, \ldots$, of increasing VC dimensions $V_{\mathcal{F}_1} \leq V_{\mathcal{F}_2} \leq \ldots$. Then for each $k = 1, 2, \ldots$ 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

$$\hat{k} = \arg\min_{k\geq 1} \left\{ \hat{R}_n(\hat{t}_n^{(k)}) + \sqrt{\frac{32 V_{\mathcal{F}_k}(\log n + 1)}{n}} \right\}$$

and $\hat{f}_n \equiv \hat{f}_n^{(\hat{k})}$ is the final choice.

The basic rational is that we know

$$R_n(\hat{f}_n^{(k)}) - \inf_{f\in\mathcal{F}_k} R(f) \leq C'\sqrt{\frac{V_{\mathcal{F}_k}\log n}{n}}$$

where C' is a constant.

The end result is that

$$E[R(\hat{f}_n)] \leq \min_{k\geq 1} \left\{ \min_{f\in\mathcal{F}_k} R(f) + 16\sqrt{\frac{V_{\mathcal{F}_k}\log n + 4}{2n}} \right\}$$

analogous to our pervious complexity regularization results, except that codelengths are replaced by VC dimensions.

In order to prove the result we use the VC probability concentration bound and assume that $\triangle = \sum_{k\geq 1} V_{\mathcal{F}_k} < \infty$. This enables a union bounding argument and leads to a risk bound of the form given above. For details see Lugosi and Zeger '96.

7 Key Point of VC Theory

Complexity of classes depends on richness (shattering capability) relative to a set of n arbitrary points. This allows us to effectively "quantize" collections of functions in a slightly data-dependent manner.

8 Application to Trees

Let

$$\mathcal{F}_k = \left\{k \;\; leaf \;\; decision \;\; trees \;\; in \;\; \mathbf{R}^d\right\}, \;\; V_{\mathcal{F}_k} \leq (d+1)(k+1)$$

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

$$\hat{k} = \arg\min_{k \geq 1} \left(\min_{f \in \mathcal{F}_k} R(f) + \sqrt{\frac{32(d+1)(k-1)(\log n + 1)}{n}} \right)$$

Then

$$\hat{f}_n = \hat{f}_n^{(\hat{k})}$$

satisfies

$$E[R(\hat{f}_n)] \leq \min_{k \geq 1} \left(\min_{f \in \mathcal{F}_k} R(f) + 16\sqrt{\frac{(d+1)(k-1)\log n + 4}{2n}} \right)$$

compare with

$$E[R(\hat{f}_n)] \leq \min_{k \geq 1} \left(\min_{f \in dyadic \; k \; leaf \; trees} R(f) + \sqrt{\frac{(3k-1)\log 2 + \frac{1}{2}\log n}{2n}} \right)$$

from Lecture 11.