CS281B/Stat241B (Spring 2006) Statistical Learning Theory	**Lecture: 23**

Covering Numbers and Metric Entropy

Lecturer: Peter Bartlett *Scribe: Alex Simma*

Definitions

For a metric space (S, p) and a subset $T \subseteq S$ we say that $\hat{T} \subseteq S$ is an ϵ-*cover* of T if $\forall t \in T$, $\exists \hat{t} \in \hat{T}$ such that $p(t, \hat{t}) \leq \epsilon$ There obviously exist many different covers of T, but we are concerned with the one containing the fewest elements. We call the size of such a cover the coverning number.

Formally, the ϵ-*covering number* of T is $N(\epsilon, T, \rho) = \min\{|\hat{T}| : \hat{T} \text{ is an } \epsilon\text{-cover}\}$

A *proper cover* is one where $\hat{T} \subseteq T$, and a *proper covering number* is defined in terms of the size of the minimum proper cover.

It can be shown that covering numbers and proper covering numbers are related by the following inequality.

$$N(\epsilon, T) \quad \leq \quad N_{\text{proper}}(\epsilon, T) \leq N(\frac{\epsilon}{2}, T)$$

The *metric entropy*, defined as $H(\epsilon, T) = \log N(\epsilon, T)$, is a natural representation of how many bits you need to send in order to identify an element of a set up to precision ϵ.

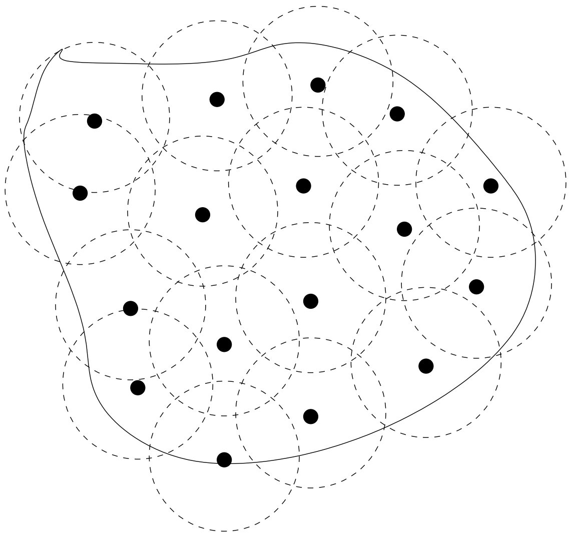

Figure 1: The cover of a set. The covering number is the size of the minimum ϵ cover and the entropy is the logarithm of the covering number

Example For the space $([0,1]^X, L_2(P_n))$, the distance is defined as

$$\rho(f,g) \quad = \quad \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-g(x_i))^2}$$

equivalent to $[0,1]^n$ with $\rho(a,b)=\sqrt{\frac{1}{n}\sum_i(a_i-b_i)^2}$

Suppose $T\subseteq[0,1]^n$ lies in a linear subspace of dimension d. We can put a grid on the subspace of points that $c\epsilon$ apart so we need $(\frac{1}{\epsilon})^d$ of them. Thus, the covering number $N(\epsilon,T,\rho)\approx\frac{1}{\epsilon^d}$ and the metric entropy $\log N(\epsilon,T,\rho)\approx d\log\frac{1}{\epsilon}$

Note that the metric dimension is $\frac{\log N(\epsilon,T,\rho)}{\log\frac{1}{\epsilon}}$ asymptotically, so the metric dimension is analagous to what the dimension does for a linear subspace.

Relationship with Rademacher averages

Theorem 1.1. In our setting: $F\subseteq[-1,1]^X$, $x_1,\ldots,x_n\in X$,

$$F_{|x_1^n}=\{(f(x_1),\ldots,f(x_n)):f\in F\}\in[-1,1]^n\ ,$$

and the $L_2(P_n)$ metric,

$$\hat{R}_n(F)\leq\inf_{\alpha}\sqrt{\frac{2\log N(\alpha,F,L_2(P_n))}{n}}+\alpha$$

Proof. Fix α. Fix a minimial α cover $\hat{F}$ of F

$$\begin{aligned}
\hat{R}_n(F) &= \mathbb{E}\sup_{f\in F}(\frac{1}{n}\sum\epsilon_i f(x_i)) && (1)\\
&= \mathbb{E}\sup_{\hat{f}\in\hat{F}}\sup_{f\in F\cap B_\alpha(\hat{f})}(\frac{1}{n}\sum\epsilon_i\hat{f}(x_i)+\frac{1}{n}\sum\epsilon_i(f(x_i)-\hat{f}(x_i)) && (2)\\
\text{(Cauchy Schwartz)} &\leq \mathbb{E}\sup_{\hat{f}\in\hat{F}}(\frac{1}{n}\sum\epsilon_i\hat{f}(x_i))+\alpha=\hat{R}_n(\hat{F})+\alpha && (3)\\
&\leq \sqrt{\frac{2\max_{\hat{f}\in\hat{F}}||\hat{f}||^2_{L_2(P_n)}\log|\hat{F}|}{n}}+\alpha && (4)\\
&\leq \sqrt{\frac{2\log N(\alpha,F,L_2(P_n))}{n}}+\alpha && (5)
\end{aligned}$$

(3) follows by Cauchy Schwartz, (4) by Massart's lemma and (5) holds since $||\hat{f}||^2_{L_2(P_n)}\leq 1$ $\square$

Example: Linear F implies that $\log N(\alpha,F)\approx d\log\frac{1}{\epsilon}$. Choose $\alpha=\frac{1}{\sqrt{n}}$ to get

$$\hat{R}(F) \leq \sqrt{\frac{cd \log n}{n}} + \sqrt{\frac{1}{n}}\ .$$