ECE901 Spring 2004 Statistical Regularization and Learning Theory **Lecture: 18**

Vapnik-Chervonenkis Theory

Lecturer: Rob Nowak *Scribe: Jonathan Liang*

1 Review of Past Lecture

In our past lectures we considered collections of candidate function $\mathcal{F}$ that were either *finite* or *enumerable*. We then constructed penalties, usually codelengths, for each candidate $c(f)$, $f \in \mathcal{F}$, such that $\sum_{f\in\mathcal{F}} 2^{c(f)} \leq 1$ This allowed us to derive uniform concentration inequalities over the entire set $\mathcal{F}$ using the union bound. However, in many cases the collections $\mathcal{F}$ may be uncountably infinite. A simple example is the collection $\mathcal{F}$ of a single threshold classifier in 1-d having the form

$$f_t(x) = \mathbf{1}_{\{x\geq t\}}$$

and their complements

$$f_s(t) = \mathbf{1}_{\{x<s\}}$$

Thus, $\mathcal{F}$ contains an uncountable number of classifiers, and we cannot apply the union bound argument in such cases.

2 Two Ways to Proceed

2.1 Discretize or Quantize the Collection $\mathcal{F}$

Example: to quantize $\mathcal{F}$

$$F_q = \left\{f, f(x) = \mathbf{1}_{\{x\leq 1/q_i i\in\{0,1,\ldots,q\}\}}\right\}$$

q is positive, such that $\forall f_q \in \mathcal{F}_q$

$$\int |f - f_q| \leq c/q$$

if the density of x is bounded by $c > 0$. $q < n^{1/2}$

2.2 Identical Empirical Errors

Consider the fact that given only n training data, many of the classifiers in such a collection may produce identical empirical errors. Also, many $f \in \mathcal{F}$ will produce identical label assignments on the data. We will have at most 2^n unique labels.

f is uncountable, its interceptions are countable and bounded by 2^n. n intervals with 2 classifier per interval.

The number of distinct labeling assignments that a class $\mathcal{F}$ can produce on a set of n points is denoted

$$S(\mathcal{F}, n) \leq 2^n$$

The VC dimension is $logS(\mathcal{F}, n)$. Specifically, $VC(\mathcal{F}) = k$, where k is largest integer such that $S(\mathcal{F}, k) = 2^k$ Ex. $2n = 2^n$, $n = 2$, $VC(\mathcal{F}) = 2$.

Ex. Consider

$$\mathcal{F} = \left\{f : f(x) = \mathbf{1}_{\{x\geq t\}} or f(x) = \mathbf{1}_{\{x<t\}}, t \in [0, 1]\right\}$$

Let q be a positive integer and

$$\mathcal{F}_q = \left\{ f : f(x) = \mathbf{1}_{\{x \geq i/q\}} or f(x) = \mathbf{1}_{\{x < i/q\}}, i \in \{0, 1, \ldots, q\} \right\}$$

and,

$$|f_q| = 2(q+1)$$

Moreover, for any $f \in \mathcal{F}$ there exists an $f_1 \in \mathcal{F}_q$ such that

$$\int |f(x) - f_q(x)| dx \leq \int_{(i-1)/q}^{i/q} 1 dx = 1/q$$

Now suppose we have n training data and suppose $f^* \in \mathcal{F}$. We know that in general, the minimum empirical risk classifier will converge to the Bayes classifier at the rate of $n^{-1/2}$ or slower. Therefore, it is unnecessary to drive the approximation error down faster than $n^{-1/2}$ So, we can restrict our attention of $\mathcal{F}_{n^{-1/2}}$ and, provided that the density of x is bound above. We have

$$min_{f \in \mathcal{F}_{n^{-1/2}}} R(f) - R(f^*) \leq C_{f_q} min \int |f^*(x) - f(x)| dx \leq c/n^{1/2}$$

Vapnik-Chervonenkis theory is based not on explicitly quantizing the collection of candidate functions, but rather on recognizing that the richness of $\mathcal{F}$ is limited in a certain sense by the number of training data. Indeed, given n i.i.d. training data, there are at most 2^n different binary labelings. Therefore, any collection $\mathcal{F}$ may be divided into 2^n subsets of classifiers that are "equilvalent" with respect to the training data. In many cases a collection may not even be capable of producing 2^n different labellings.

3 Example

Consider $X = [0, 1]$

$$\mathcal{F} = \left\{ f : f(x) = \mathbf{1}_{\{x \geq t\}} or f(x) = \mathbf{1}_{\{x < t\}} t \in [0, 1] \right\}$$

Suppose we have n training data: $(x_1, \ldots, x_n) \in [0, 1]$. With x^s denotes the location of each training point in [0,1]. Associated with each x is a label $y \in \{0, 1\}$. Any classifier in $\mathcal{F}$ will label 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 label "1" or "0" and so on. We see that there are exactly $2n$ different labellings; far less than 2^n!

The number of different labellings that a class $\mathcal{F}$ can produce on a set of n training data is a measure of the "effective size" of $\mathcal{F}$. The Vapnik-Chervonenkis (VC) dimension of $\mathcal{F}$ is proportional to the log of the effective size. Let $V(\mathcal{F}, n)$ denote the VC dimension of $\mathcal{F}$, typically a constant, independent of n. The VC inequality states that for all $f \in \mathcal{F}$

$$P\left(|\widehat{R}_n(f) - R(f)| > \epsilon \right) \leq 8 e^{V(\mathcal{F}, h)} e^{-n\epsilon^2/32}$$

This type of uniform concentration inequality can be used in a similar fashion to our use of Hoeffding's inequality plus union bound.

4 Hyperplane Classifiers

We will go into the details of VC Theory next lecture, and the remainder of this lecture will introduce the key ideas with an example Consider the following setup. Let $X = [0, 1]^d$, $Y = \{0, 1\}$ Let

$$\mathcal{F} = \left\{ f : f(x) = \mathbf{1}_{\{w^T x + w_0 > 0\}} \right\}$$

with w_0 and $w \in R^{d+1}$ This is the collection of all hyperplane classifiers. $\mathcal{F}$ is infinite and uncountable.

Suppose that we have n training data

$$\{X_i, Y_i\}_{i=1}^n$$

There are at most $2\binom{n}{d}$ unique classifiers in $\mathcal{F}$ with respect to these data. To see this, consider d arbitrary data points $x_1, \ldots, x_{i_d}$, and let $w^T x + w_0 > 0$ be a hyperplane containing these points. To e specific, take the hyperplane with

$$||w_0 w|| = 1$$

. this hyperplane coincides with two possible classification rules:

$$f_1(x) = \mathbf{1}_{\{w^T x + w_0 > 0\}}$$

$$f_2(x) = \mathbf{1}_{\{w^T x + w_0 < 0\}}$$

Each d-tuple of training data produces two distinct classifiers, assuming the data are not co-linear. Thus, there are at most $2 * \binom{n}{d}$ unique classifiers in $\mathcal{F}$ with respect to the training data. (All other $f \in \mathcal{F}$ produce the same labels and empirical risk as one of the classifiers.) Let's enumerate the unique hyperplane classifiers $f_1, \ldots, f_{2*\binom{n}{d}}$, and let

$$\widehat{f}_n = \arg \min_{f \in \{f_1, \ldots, f_{2\binom{n}{d}}\}} \widehat{R}_n(f)$$

and let

$$R^* = inf_{f \in \mathcal{F}} R(f)$$

and define

$$f^* = argmin_{f \in \mathcal{F}} R(f)$$

If multiple $f \in \mathcal{F}$ achieve R^*, pick f^* to be one of them in an arbitrary fixed number.

Theorem: Assume that P_x has a density, but that the distribution of (x, y) is other arbitrary. If $n \geq d$ and $2d/n \leq \epsilon \leq 1$ then

$$P\left(R(\widehat{f}_n) - R(f) > \epsilon\right) \leq e^{2d\epsilon}\left(2\binom{n}{d} + 1\right) e^{-n\epsilon^2/2}$$

NOTE: The assumption that P_x has a density insures that no d+1 points are co-planar. This in turns, guarantees that there are exactly $2\binom{n}{d}$ unique classifier and that the $2\binom{n}{d}$ under consideration are fully representative of all possible classifiers in $\mathcal{F}$, with respect to the data.

Proof: The proof is a specialization of the basic ingredients of VC Theory to the case at hand. Here we follow the proof in DGL '96. First we note that,

$$R(\widehat{f}_n) - R(f^*) = R(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n) + \widehat{R}_n(\widehat{f}_n) - R(f^*)$$

$$\leq R(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n) + \widehat{R}_n f^*) - R(f^*) + d/n$$

and since $\widehat{R}_n(\widehat{f}_n) \leq \widehat{R}_n(f) + d/n$ for any $f \in \mathcal{F}$

$$\leq max_{i=1,\ldots,2\binom{n}{d}}(R(f_i) - \widehat{(R)}_n(f_i)) + \widehat{(R)}_n(f^*) - R(f^*) + d/n$$

Therefore, by the union bound:

$$P(R(\widehat{f}_n) - R(f^*) > \epsilon)$$

$$\leq \sum_{i=1}^{2\binom{n}{d}} P\left(R(f_i) - \widehat{R}_n(f_i) > \epsilon/2\right) + P\left(\widehat{R}_n(f^*) - R(f^*) + d/n > \epsilon/2\right)$$

We can bound the second term of the above bound using Chernoff's/Hoeffding's inequality:

$$P\left(\widehat{R}_n(f^*) - R(f^*) > \epsilon/2 - d/n\right)$$

$$\leq e^{-2n(\epsilon/2 - d/n)^2}$$

$$\leq e^{2d\epsilon} e^{-n\epsilon^2/2}$$

Next, let's bound one of the terms in the summation. For example, take

$$P\left(R(f_i) - \widehat{R}_n(f_i) > (\epsilon/2)\right)$$

Note that by symmetry all $2\binom{n}{d}$ terms will have identical bounds. Since the bounds are independent of $P_x y$.

Assume that f_1 is determined by the first d data points $x_1 4, \ldots, x_d$. By the smoothing property of expectations we can write,

$$P\left(R(f_i) - \widehat{R}_n(f) > \epsilon/2\right) = E\left[P\left(R(f_i) - \widehat{R}_n(f_i) > \epsilon/2 | x_1, \ldots, x_d\right)\right]$$

From here, we will bound the conditional probability inside the expectation. Let $(X_1^{"}, Y_1^{"}), \ldots, (X_d^{"}, Y_d^{"})$ be d additional random samples that are independent and identically distributed as the data $(X_1, Y_1), \ldots, (X_d, Y_d)$. $\left\{X_i^{"}, Y_i^{"}\right\}_{i=1}^{d}$ are often called the "ghost sample" since they are not actually observed. They are a fictious sample leads to a simple bound on the conditional probability. Define if $i \leq d$

$$(X_i^{'}, Y_i^{'}) = (X_i^{"}, Y_i^{"})$$

or if $i > d$

$$(X_i^{'}, Y_i^{'}) = (X_i, Y_i)$$

That is, $\left\{X_i^{'}, Y_i^{'}\right\}_{i=1}^{d}$ agrees with our observed data on i¿d, but the first d samples are replaced with the ghost sample. Then,

$$\begin{aligned}
&P\left(R(f_i) - \widehat{R}_n(f_1) > \epsilon/2 | x_1, \ldots, x_d\right) \\
&\leq P\left(R(f_i) - 1/n \sum_{i=d+1}^{n} 1_{f_1(x_i) \neq y_i} > \epsilon/2 | x_1, \ldots, x_d\right) \\
&\leq P\left(R(f_i) - 1/n \sum_{1}^{n} 1_{f_1(x_i) \neq y_i} + d/n > \epsilon/2 | x_1, \ldots, x_d\right) \\
&= P\left(R(f_i) - \widehat{(R)}_n^{'}(f_1) > t/2 - d/n | x_1, \ldots, x_d\right)
\end{aligned}$$

where,

$$\widehat{R}_n^{'}(f_1) = 1/n \sum_{i=1}^{n} \mathbf{1}_{\{f_1(x_i^{'}) \neq y_i^{'}\}}$$

Note that $n\widehat{(R)}_n^{'}(f_1)$ is binomially distributed with mean $R(f_1)$ and it is independent of $x_1, \ldots, x_d$ Therefore,

$$\begin{aligned}
&P\left(R(f_i) - \widehat{R}_n^{'}(f_1) > \epsilon/2 - d/n | x_1, \ldots, x_d\right) \\
&= P(R(f_i) - \widehat{R}_n^{'}(f_1) > t/2 - d/n | x_1, \ldots, x_d) \\
&\leq e^{-2n(\epsilon/2 - d/n)^2} \\
&\leq e^{2d\epsilon} e^{-n\epsilon^2/2}
\end{aligned}$$

In conclusion,

$$\begin{aligned}
&P\left(R(\widehat{f}_n) - R^* > \epsilon\right) \\
&\leq \sum_{i=1}^{2\binom{n}{d}} P\left(R(f)_i - \widehat{R}_n(t_i) > \epsilon/2\right) + P\left(\widehat{R}_n(f^*) - R(f^*) + d/n > \epsilon/2\right)
\end{aligned}$$

$$\leq 2\binom{n}{d}e^{2d\epsilon}e^{-n\epsilon^2/2} + e^{2d\epsilon}e^{-n\epsilon^2/2}$$

$$= e^{2d\epsilon}\left(2\binom{n}{d}+1\right)e^{-n\epsilon^2/2}$$

Lastly, Corollary If $n \geq d$, then

$$E\left[R(\widehat{f}_n) - min_{f\in\mathcal{F}}R(f)\right] \leq \sqrt{2(d+1)(logn+2)/n}$$