COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Berk Kapicioglu

Lecture #5
February 21, 2006

1 An Outline of the Lecture

In this lecture, we continue to study general conditions under which one could perform learning successfully. In particular, under mild restrictions on the hypothesis class and the learning algorithm, we want to find out how many training examples we need for successful learning (sample complexity). Equivalently, we want to find out, under similar mild restrictions, upper bounds on the generalization error given a fixed number of training examples.

2 Recap

Last lecture, as a first step in deriving such results, we proved a theorem called "Occam's Razor":

Theorem 1 (Occam's Razor) *Let $\mathcal{H}$ be finite. With probability at least $1-\delta$, $\forall h \in \mathcal{H}$, if h is consistent then $err_D(h) \leq \frac{\ln|\mathcal{H}|+ln\frac{1}{\delta}}{m}$.*

Note that "$\lg|\mathcal{H}|$" is the number of bits needed to specify a hypothesis from the hypothesis class $\mathcal{H}$. Since $\ln|\mathcal{H}| = \frac{\lg|\mathcal{H}|}{\lg e}$, the bound on $err_D(h)$ grows proportional to the "complexity" of the hypothesis class $\mathcal{H}$.

A natural question to ask at this point is: can we obtain similar results for hypothesis classes that are infinitely large?

3 Vapnik-Chervonenkis Theory

The answer is yes! First, let's introduce some new concepts and notation that will help us in the upcoming discussion.

S	$\stackrel{\text{def}}{=}$	$(x_1, ..., x_m)$	(sample set)
$\Pi_{\mathcal{H}}(S)$	$\stackrel{\text{def}}{=}$	$\{\langle h(x_1), ..., h(x_m)\rangle : h \in \mathcal{H}\}$	(dichotomies of $\mathcal{H}$ in S)
$\Pi_{\mathcal{H}}(m)$	$\stackrel{\text{def}}{=}$	$\max_{\lvert S\rvert=m}\lvert\Pi_{\mathcal{H}}(S)\rvert$	(growth function)

Note that the number of dichotomies of a hypothesis class $\mathcal{H}$ in S is at most 2^m. In other words:

$$\forall\mathcal{H}, \forall m, \Pi_{\mathcal{H}}(m) \leq 2^m$$

However, there are many hypothesis classes where the number of dichotomies is much smaller than 2^m. For example, when $\mathcal{H} = \{\text{positive half-lines}\}$, $\Pi_{\mathcal{H}}(m) = m+1$. In fact, we will show later in the course that:

$$\begin{array}{lll} \forall\mathcal{H}, \text{either} & & \\ & (1) & \forall m, \Pi_{\mathcal{H}}(m) = 2^m \\ OR & (2) & \forall m, \Pi_{\mathcal{H}}(m) = O(m^d). \end{array}$$

Now, let's prove the main theorem of this lecture.

Theorem 2 *With probability at least* $1-\delta$, $\forall h \in \mathcal{H}$, *if* h *is consistent then* $err_D(h) \leq O\left(\frac{\ln \Pi_{\mathcal{H}}(2m)+ln\frac{1}{\delta}}{m}\right)$.

Proving the above theorem is equivalent to proving that

$$Pr_S[\exists h \in \mathcal{H} : h \text{ is consistent and } err(h) > \epsilon] \leq \delta.$$

In order to prove the above bound, we will use something called the "Double Sample Trick". That is, we will choose an "imaginary" sample S' and use it to prove the above bound. The main idea is to use a large number of mistakes on S' as a proxy for a hypothesis having a large generalization error since this latter event is much harder to handle when dealing with an infinite hypothesis space.

Before we proceed, we introduce some additional notation:

S	$\stackrel{\text{def}}{=}$	$(x_1, ..., x_m)$	("real" sample set)
S'	$\stackrel{\text{def}}{=}$	$(x'_1, ..., x'_m)$	("imaginary" sample set)
$M(h, S)$	$\stackrel{\text{def}}{=}$	# of mistakes of h on S	
B	$\stackrel{\text{def}}{=}$	$[\exists h \in \mathcal{H} : (h \text{ is consistent with } S) \wedge (err(h) > \epsilon)]$	(bad event)
B'	$\stackrel{\text{def}}{=}$	$[\exists h \in \mathcal{H} : (h \text{ is consistent with } S) \wedge (M(h, S') \geq \frac{m\epsilon}{2})\ ]$	

Step 1 $Pr[B'|B] \geq \frac{1}{2}$.

Proof: If we write out the probability in its full form, we are interested in:

$$Pr[\{\exists h \in \mathcal{H} : (M(h, S) = 0) \wedge (M(h, S') \geq \frac{m\epsilon}{2})\} | \{\exists h' \in \mathcal{H} : (M(h', S) = 0) \wedge (err(h') > \epsilon)\}]$$

In plain English, we are interested in the conditional probability that there exists $h \in \mathcal{H}$ such that h is consistent with S and h has at least $\frac{m\epsilon}{2}$ mistakes on S' given there exists $h' \in \mathcal{H}$ such that h' is consistent with S and $err(h') > \epsilon$. Note that given a hypothesis h that satisfies event B, if it satisfies $M(h, S') \geq \frac{m\epsilon}{2}$, event B' is also satisfied. Thus, it suffices to give a lower bound for the following event:

$$Pr[B'|B] \geq Pr\left[M(h, S') \geq \frac{m\epsilon}{2} | B\right]$$

where h satisfies B. This is essentially the chance of getting at least $\frac{m\epsilon}{2}$ heads when flipping m identical coins, each with probability of heads at least ϵ. Since h satisfies B, $err(h) > \epsilon$, thus

$$E[M(h, S')|B] > m\epsilon,$$

so the probability of getting half the expected number of heads (mistakes) will be small. This argument can be made more precise using techniques that we will learn later in the course (namely Chernoff bounds) to show that

$$Pr\left[M(h, S') < \frac{m\epsilon}{2} | B\right] < \frac{1}{2}$$

which proves the claim.

Step 2 $Pr[B] \leq 2Pr[B']$.

Proof:

$$Pr[B'] \geq Pr[B' \wedge B] = Pr[B]Pr[B'|B] \geq Pr[B]\frac{1}{2}$$

where we used Step 1 for the last inequality.

An Alternative Way to Sample S and S':
In order to proceed with the VC proof, we provide an alternative way to sample S and S'. However, we need to be careful, because our arguments rely on the fact that S and S' are chosen *iid* from $\mathcal{D}$. Thus, if we want to change the sampling method while preserving the validity of our arguments, we need to argue that the alternative sampling method preserves the probability distribution over S and S' induced by the original sampling method.

Let's define the two different sampling methods:

Experiment 1 (Original Sampling):
Choose S and S' *iid* from $\mathcal{D}$.

Experiment 2 (Alternative Sampling):
Choose S and S' *iid* from $\mathcal{D}$. Then, flip m fair coins. For each coin flip i, if the coin flip comes up heads then exchange x_i and x'_i; if it comes up tails then do not exchange x_i and x'_i. Call the new samples T and T'.

The probability distribution over S and S' is the same as the probability distribution over T and T' since each sample is chosen *iid* from $\mathcal{D}$. Now we continue with the proof of the VC Theorem.

Step 3 $Pr[B'] = Pr[B'']$.

where

$$\begin{aligned} b(h) &\stackrel{\text{def}}{=} [(M(h,T) = 0) \wedge (M(h,T') \geq \tfrac{m\epsilon}{2})] \\ B'' &\stackrel{\text{def}}{=} [\exists h \in \mathcal{H} : b(h)] \end{aligned}$$

Proof: Result follows from the discussion above since S, S' are distributed exactly the same as T, T'.

Step 4 *Let $h \in \mathcal{H}$ be arbitrary.* $Pr[b(h)|S, S'] \leq 2^{m\epsilon/2}$.

Proof: In case the notation is unclear, the lemma states that there is an upper bound on the probability that a particular hypothesis h will satisfy $b(h)$ after S and S' are swapped using random coin flips to form T and T'. We can prove this lemma by analyzing the different ways the condition could be realized.
Let x_i and x'_i be corresponding elements of S and S' respectively.

Case 1: In this case, $\exists x_i \in S, x'_i \in S'$ such that h misclassifies both x_i and x'_i. Then, no matter how we flip the coins and form T and T', $M(h,T) \neq 0$. Thus,

$$Pr[b(h)|S, S'] = 0$$

Case 2: In this case, we are interested in the event where r of the m pairs of S and S' have exactly one instance, either x_i or x'_i, that h misclassifies. The rest of the m pairs, in particular $m - r$ of them, are classified correctly by h. Furthermore, we assume that

$r < \frac{m\epsilon}{2}$ in this case. Then, no matter what the coin flips are and how T and T' are formed, $M(h, T') \leq r < \frac{m\epsilon}{2}$, so $b(h)$ cannot be realized:

$$Pr[b(h)|S, S'] = 0$$

Case 3: Let r be as above, except in this final case, $r \geq \frac{m\epsilon}{2}$. Then, in order to satisfy $M(h, T) = 0$ and $M(h, T') \geq \frac{m\epsilon}{2}$, we need just the right sequence of coin flips that would assure that all of the r pairs are arranged correctly (i.e., with all the mistakes on T', and none on T). Since coin flips are independent:

$$Pr[b(h)|F] = \left(\frac{1}{2}\right)^r \leq 2^{-m\epsilon/2}.$$

Thus, in every case, the claim holds.

Step 5 $Pr_{T,T'}[B''] \leq \Pi_{\mathcal{H}}(2m)2^{-m\epsilon/2}$.

Proof: Let $\mathcal{H}'(S) \subseteq \mathcal{H}$ be a set of hypotheses containing exactly one representative hypothesis for each dichotomy in $\Pi_{\mathcal{H}}(S)$. Then $b(h)$ holds for some $h \in \mathcal{H}$ if and only if it holds for some $h \in \mathcal{H}'(S \cup S')$ since $b(h)$ is a property only of points in $S \cup S'$. Thus,

$$\begin{array}{rcll} Pr[B''] & = & E_{S,S'}[Pr[\exists h \in \mathcal{H} : b(h)|S, S']] & \\ & = & E[Pr[\exists h \in \mathcal{H}'(S \cup S') : b(h)|S, S']] & \\ & \leq & E[\sum_{h \in \mathcal{H}'(S \cup S')}[Pr[b(h)|S, S']] & \text{(by union bound)} \\ & \leq & E[|\mathcal{H}'(S \cup S')|2^{-m\epsilon/2}] & \text{(by Step 4)} \\ & \leq & \Pi_{\mathcal{H}}(2m)2^{-m\epsilon/2} & \end{array}$$

Conclusion: We combine the results of all the steps for the final touch and prove the theorem:

$$\begin{array}{rcll} Pr[B] & \leq & 2Pr[B'] & \text{(by Step 2)} \\ & = & 2Pr[B''] & \text{(by Step 3)} \\ & \leq & 2\Pi_{\mathcal{H}}(2m)2^{-m\epsilon/2} & \text{(by Step 5)} \\ & = & \delta & \text{(because we want to upper bound it)} \end{array}$$

Solving for ϵ yields the desired result:

$$err(h) \leq \frac{2}{m}(\lg(\Pi_{\mathcal{H}}(2m)) + \lg(1/\delta) + 1).$$

4 Vapnik-Chervonenkis Dimension

In this section, we introduce the notion of Vapnik-Chervonenkis (VC) dimension. In order to motivate the concept, let's start off with an example.

Example 1: Suppose that $\mathcal{H}$ ={intervals}. Furthermore, let's suppose that our sample consists of 2 (distinct) points: $S = \{x_1, x_2\}$. It's easy to see that all dichotomies over S could be realized by $\mathcal{H}$. In other words, S is shattered by $\mathcal{H}$:

$$S \text{ is said to be } \textit{shattered} \text{ by } \mathcal{H} \text{ if } |\Pi_{\mathcal{H}}(S)| = 2^{|S|}$$

Note that in the example above, if sample S consisted of 3 points instead of 2, then not all dichotomies would have been realized. For example, given 3 consecutive points $x_1 < x_2 < x_3$, no interval would be able to classify them as $+, -, +$. Thus, in the example above, no sample S of size 3 or more is shattered by $\mathcal{H}$.

Now, let's define the VC dimension of a hypothesis class $\mathcal{H}$:

$$\begin{aligned} \text{VC-dim}(\mathcal{H}) &= \text{cardinality of largest set shattered by } \mathcal{H} \\ &= \text{largest } m \text{ such that } \exists \text{ set of size } m \text{ shattered by } \mathcal{H} \end{aligned}$$

Example 2: Let $\mathcal{H} = \{\text{axis-parallel rectangles}\}$. Let S be a sample that consists of points in $\mathbb{R}^2$ that form the four corners of a diamond shape. It's easy to see that $\mathcal{H}$ realizes all dichotomies of S, hence VC-dim$(\mathcal{H}) \geq 4$. Furthermore, there doesn't exist a sample of size 5 where all the dichotomies of $\mathcal{H}$ are realized. The reason is, if we label the topmost, rightmost, leftmost and the bottommost example with a positive label, and label the middle example with a negative label, no axis-aligned rectangle would be able to classify the examples in such a way. Thus, VC-dim $(\mathcal{H}) = 4$.

Example 3: Let $\mathcal{H}$ be finite. If a sample S of size d is shattered by $\mathcal{H}$, then $|\mathcal{H}| \geq 2^d$. Thus, VC-dim$(\mathcal{H}) = d \leq \lg(|\mathcal{H}|)$.

In the upcoming lecture, we will see how the VC-dimension of a hypothesis class affects its sample complexity and its generalization error.