

In this lecture will use the measure of VC dimension, which is a combinatorial measure of concept class complexity, to bound the sample size complexity.

## 12.1 A Polynomial Bound on the Sample Size $m$ for PAC Learning

In this section we will follow the material presented in Kearns & Vazirani pp. 57–61 and prove the following:

**Theorem 1 (Double Sampling)** *Let  $C$  be any concept class of VC dimension  $d$ . Let  $L$  be any algorithm that when given a set  $S$  of  $m$  labeled examples  $\{\mathbf{x}_i, c(\mathbf{x}_i)\}_i$ , sampled i.i.d according to some fixed but unknown distribution  $D$  over the instance space  $X$ , of some concept  $c \in C$ , produces as output a concept  $h \in C$  that is consistent with  $S$ . Then  $L$  is a learning algorithm in the formal sense provided that the sample size obeys:*

$$m \geq c_0 \left( \frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{1}{\epsilon} \right)$$

for some constant  $c_0 > 0$ .

The idea behind the proof is to build an "approximate" concept space which includes concepts arranged such that the distance between the approximate concepts  $h$  and the target concept  $c$  is at least  $\epsilon$  — where distance is defined as the weight of the region in  $X$  which is in conflict with the target concept. To formalize this story we will need few more definitions. Unless specified otherwise,  $c \in C$  denotes the target concept and  $h \in C$  denotes *some* concept.

### Definition 1

$$c\Delta h = h\Delta c = \{\mathbf{x} : c(\mathbf{x}) \neq h(\mathbf{x})\}$$

$c\Delta h$  is the region in instance space where both concepts do not agree — the error region. The probability that  $\mathbf{x} \in c\Delta h$  is equal to (by definition)  $err(h)$ .

### Definition 2

$$\begin{aligned} \Delta(c) &= \{h\Delta c : h \in C\} \\ \Delta_\epsilon(c) &= \{h\Delta c : h \in C \text{ and } err(h) \geq \epsilon\} \end{aligned}$$

$\Delta(c)$  is a set of error regions, one per concept  $h \in C$  over all concepts. The error regions are with respect to the target concept. The set  $\Delta_\epsilon(c) \subset \Delta(c)$  is the set of all error regions whose weight exceeds  $\epsilon$ . Recall that weight is defined as the probability that a point sampled according to  $D$  will hit the region.

It will be important for later to evaluate the VC dimension of  $\Delta(c)$ . Unlike  $C$ , we are not looking for the VC dimension of a class of function but the VC dimension of a set of regions in space. Recall the definition of  $\Pi_C(S)$  from the previous lecture: there were two equivalent definitions one based on a set of vectors each representing a labeling of the instances of  $S$  induced by some concept. The second, yet equivalent, definition is based on a set of subsets of  $S$  each induced by some concept (where the concept divides the sample points of  $S$  into positive and negative labeled points). So far it was convenient to work with the first definition, but for evaluating the VC dimension of  $\Delta(c)$  it will be useful to consider the second definition:

$$\Pi_{\Delta(c)}(S) = \{r \cap S : r \in \Delta(c)\},$$

that is, the collection of subsets of  $S$  induced by intersections with regions of  $\Delta(c)$ . An intersection between  $S$  and a region  $r$  is defined as the subset of points from  $S$  that fall into  $r$ . We can easily show that the VC dimensions of  $C$  and  $\Delta(c)$  are equal:

**Lemma 2**

$$VCdim(C) = VCdim(\Delta(c)).$$

**Proof:** we have that the elements of  $\Pi_C(S)$  and  $\Pi_{\Delta(c)}(S)$  are subsets of  $S$ , thus we need to show that for every  $S$  the cardinality of both sets is equal  $|\Pi_C(S)| = |\Pi_{\Delta(c)}(S)|$ . To do that it is sufficient to show that for every element  $s \in \Pi_C(S)$  there is a unique corresponding element in  $\Pi_{\Delta(c)}(S)$ . Let  $c \cap S$  be the subset of  $S$  induced by the target concept  $c$ . The set  $s$  (a subset of  $S$ ) is realized by some concept  $h$  (those points in  $S$  which were labeled positive by  $h$ ). Therefore, the set  $s \cap (c \cap S)$  is the subset of  $S$  containing the points that hit the region  $h\Delta c$  which is an element of  $\Pi_{\Delta(c)}(S)$ . Since this is a one-to-one mapping we have that  $|\Pi_C(S)| = |\Pi_{\Delta(c)}(S)|$ .  $\square$

**Definition 3 ( $\epsilon$ -net)** For every  $\epsilon > 0$ , a sample set  $S$  is an  $\epsilon$ -net for  $\Delta(c)$  if every region in  $\Delta_\epsilon(c)$  is hit by at least one point of  $S$ :

$$\forall r \in \Delta_\epsilon(c), \quad S \cap r \neq \emptyset.$$

In other words, if  $S$  hits all the error regions in  $\Delta(c)$  whose weight exceeds  $\epsilon$ , then  $S$  is an  $\epsilon$ -net. Consider as an example the concept class of intervals on the line  $[0, 1]$ . A concept is defined by an interval  $[\alpha_1, \alpha_2]$  such that all points inside the interval are positive and all those outside are negative. Given  $c \in C$  is the target concept and  $h \in C$  is some concept, then the error region  $h\Delta c$  is the union of two intervals:  $I_1$  consists of all points  $x \in h$  which are not in  $c$ , and  $I_2$  the interval of all points  $x \in c$  but which are not in  $h$ . Assume that the distribution  $D$  is uniform (just for the sake of this example) then,  $prob(x \in I) = |I|$  which is the length of the interval  $I$ . As a result,  $err(h) > \epsilon$  if either  $|I_1| > \epsilon/2$  or  $|I_2| > \epsilon/2$ . The sample set

$$S = \{x = \frac{k\epsilon}{2} : k = 0, 1, \dots, 2/\epsilon\}$$

contains sample points from 0 to 1 with increments of  $\epsilon/2$ . Therefore, every interval larger than  $\epsilon$  must be hit by at least one point from  $S$  and by definition  $S$  is an  $\epsilon$ -net.

It is important to note that if  $S$  forms an  $\epsilon$ -net then we are guaranteed that  $err(h) \leq \epsilon$ . Let  $h \in C$  be the consistent hypothesis with  $S$  (returned by the learning algorithm  $L$ ). Because  $h$  is consistent,  $h\Delta c \in \Delta(c)$  has not been hit by  $S$  (recall that  $h\Delta c$  is the error region with respect to the target concept  $c$ , thus if  $h$  is consistent then it agrees with  $c$  over  $S$  and therefore  $S$  does not hit  $h\Delta c$ ). Since  $S$  forms an  $\epsilon$ -net for  $\Delta(c)$  we must have  $h\Delta c \notin \Delta_\epsilon(c)$  (recall that by definition

$S$  hits all error regions with weight larger than  $\epsilon$ ). As a result, the error region  $h\Delta c$  must have a weight smaller than  $\epsilon$  which means that  $err(h) \leq \epsilon$ .

The conclusion is that if we can *bound* the probability that a random sample  $S$  *does not* form an  $\epsilon$ -net for  $\Delta(c)$ , then we have bounded the probability that a concept  $h$  consistent with  $S$  has  $err(h) > \epsilon$ . This is the goal of the proof of the double-sampling theorem which we are about to prove below:

**Proof (following Kearns & Vazirani pp. 59–61):** Let  $S_1$  be a random sample of size  $m$  (sampled i.i.d. according to the unknown distribution  $D$ ) and let  $A$  be the event that  $S_1$  *does not* form an  $\epsilon$ -net for  $\Delta(c)$ . From the preceding discussion our goal is to upper bound the probability for  $A$  to occur, i.e.,  $prob(A) \leq \delta$ .

If  $A$  occurs, i.e.,  $S_1$  is not an  $\epsilon$ -net, then by definition there must be some region  $r \in \Delta_\epsilon(c)$  which is not hit by  $S_1$ , that is  $S_1 \cap r = \emptyset$ . Note that  $r = h\Delta(c)$  for some concept  $h$  which is consistent with  $S_1$ . At this point the space of possibilities is infinite, because the probability that we fail to hit  $h\Delta(c)$  in  $m$  random examples is at most  $(1 - \epsilon)^m$ . Thus the probability that we fail to hit *some*  $h\Delta c \in \Delta_\epsilon(c)$  is bounded from above by  $|\Delta(c)|(1 - \epsilon)^m$  — which does not help us due to the fact that  $|\Delta(c)|$  is infinite. The idea of the proof is to turn this into a finite space by using another sample, as follows.

Let  $S_2$  be another random sample of size  $m$ . We will select  $m$  (for both  $S_1$  and  $S_2$ ) to guarantee a high probability that  $S_2$  will hit  $r$  many times. In fact we wish that  $S_2$  will hit  $r$  at least  $\frac{\epsilon m}{2}$  with probability of at least 0.5:

$$prob(|S_2 \cap r| > \frac{\epsilon m}{2}) = 1 - prob(|S_2 \cap r| \leq \frac{\epsilon m}{2}).$$

We will use the Chernoff bound (lower tail) to obtain a bound on the right-hand side term. Recall that if we have  $m$  Bernoulli trials (coin tosses)  $Z_1, \dots, Z_m$  with expectation  $E(Z_i) = p$  and we consider the random variable  $Z = Z_1 + \dots + Z_m$  with expectation  $E(Z) = \mu$  (note that  $\mu = pm$ ) then for all  $0 < \psi < 1$  we have:

$$prob(Z < (1 - \psi)\mu) \leq e^{-\frac{\mu\psi^2}{2}}.$$

Considering the sampling of  $m$  examples that form  $S_2$  as Bernoulli trials, we have that  $\mu \geq \epsilon m$  (since the probability that an example will hit  $r$  is at least  $\epsilon$ ) and  $\psi = 0.5$ . We obtain therefore:

$$prob(|S_2 \cap r| \leq (1 - \frac{1}{2})\epsilon m) \leq e^{-\frac{\epsilon m}{8}} = \frac{1}{2}$$

which happens when  $m = \frac{8}{\epsilon} \ln 2 = O(\frac{1}{\epsilon})$ . To summarize what we have obtained so far, we have calculated the probability that  $S_2$  will hit  $r$  many times *given* that  $r$  was fixed using the previous sampling, i.e., given that  $S_1$  does not form an  $\epsilon$ -net. To formalize this, let  $B$  denote the combined event that  $S_1$  does not form an  $\epsilon$ -event and  $S_2$  hits  $r$  at least  $\epsilon m/2$  times. Then, we have shown that for  $m = O(1/\epsilon)$  we have:

$$prob(B/A) \geq \frac{1}{2}.$$

From this we can calculate  $prob(B)$ :

$$prob(B) = prob(B/A)prob(A) \geq \frac{1}{2}prob(A),$$

which means that our original goal of bounding  $prob(A)$  is equivalent to finding a bound  $prob(B) \leq \delta/2$  because  $prob(A) \leq 2 \cdot prob(B) \leq \delta$ . The crucial point with the new goal is that to analyze

the probability of the event  $B$ , we need only to consider a finite number of possibilities, namely to consider the regions of

$$\Pi_{\Delta_\epsilon(c)}(S_1 \cup S_2) = \{r \cap \{S_1 \cup S_2\} : r \in \Delta_\epsilon(c)\}.$$

This is because the occurrence of the event  $B$  is equivalent to saying that there is some  $r \in \Pi_{\Delta_\epsilon(c)}(S_1 \cup S_2)$  such that  $|r| \geq \epsilon m/2$  (i.e., the region  $r$  is hit at least  $\epsilon m/2$  times) and  $S_1 \cap r = \emptyset$ . This is because  $\Pi_{\Delta_\epsilon(c)}(S_1 \cup S_2)$  contains all the subsets of  $S_1 \cup S_2$  realized as intersections over all regions in  $\Delta_\epsilon(c)$ . Thus even though we have an infinite number of regions we still have a finite number of subsets. We wish therefore to analyze the following probability:

$$\text{prob}\left(r \in \Pi_{\Delta_\epsilon(c)}(S_1 \cup S_2) : |r| \geq \epsilon m/2 \text{ and } S_1 \cap r = \emptyset\right).$$

Let  $S = S_1 \cup S_2$  a random sample of  $2m$  (note that since the sampling is i.i.d. it is equivalent to sampling  $S_1$  and  $S_2$  separately) and  $r$  satisfying  $|r| \geq \epsilon m/2$  being *fixed*. Consider some random partitioning of  $S$  into  $S_1$  and  $S_2$  and consider then the problem of estimating the probability that  $S_1 \cap r = \emptyset$ . This problem is equivalent to the following combinatorial question: we have  $2m$  balls, each colored Red or Blue, with exactly  $l \geq \epsilon m/2$  Red balls. We divide the  $2m$  balls into groups of equal size  $S_1$  and  $S_2$  and we are interested in bounding the probability that all of the  $l$  balls fall in  $S_2$  (that is, the probability that  $S_1 \cap r = \emptyset$ ). This in turn is equivalent to first dividing the  $2m$  uncolored balls into  $S_1$  and  $S_2$  groups and then randomly choose  $l$  of the balls to be colored Red and analyze the probability that all of the Red balls fall into  $S_2$ . This probability is exactly

$$\frac{\binom{m}{l}}{\binom{2m}{l}} = \prod_{i=0}^{l-1} \frac{m-i}{2m-i} \leq \prod_{i=0}^{l-1} \frac{1}{2} = \frac{1}{2^l} = 2^{-\epsilon m/2}.$$

This probability was evaluated for a *fixed*  $S$  and  $r$ . Thus, the probability that this occurs for *some*  $r \in \Pi_{\Delta_\epsilon(c)}(S)$  satisfying  $|r| \geq \epsilon m/2$  (which is  $\text{prob}(B)$ ) can be calculated by summing over all possible fixed  $r$  and applying the union bound  $\text{prob}(\sum_i Z_i) \leq \sum_i \text{prob}(Z_i)$ :

$$\begin{aligned} \text{prob}(B) &\leq |\Pi_{\Delta_\epsilon(c)}(S)| 2^{-\epsilon m/2} \leq |\Pi_{\Delta(c)}(S)| 2^{-\epsilon m/2} \\ &= |\Pi_C(S)| 2^{-\epsilon m/2} \leq \left(\frac{2\epsilon m}{d}\right)^d 2^{-\epsilon m/2} \leq \frac{\delta}{2}, \end{aligned}$$

from which it follows that:

$$m = O\left(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{d}{\epsilon} \log \frac{1}{\epsilon}\right).$$

□

Few comments are worthwhile at this point:

1. It is possible to show that the upper bound on the sample complexity  $m$  is tight by showing that the lower bound on  $m$  is  $\Omega(d/\epsilon)$  (see Kearns & Vazirani pp. 62).
2. The treatment above holds also for the unrealizable case (target concept  $c \notin C$ ) with slight modifications to the bound. In this context, the learning algorithm  $L$  must simply minimize the sample (empirical) error  $e\hat{r}(h)$  defined:

$$e\hat{r}(h) = \frac{1}{m} |\{i : h(\mathbf{x}_i) \neq y_i\}| \quad \mathbf{x}_i \in S.$$

The generalization of the double-sampling theorem (Derroye'82) states that the empirical errors converge uniformly to the true errors:

$$\text{prob} \left( \max_{h \in C} |\hat{err}(h) - err(h)| \geq \epsilon \right) \leq 4e^{(4\epsilon + 4\epsilon^2)} \left( \frac{\epsilon m^2}{d} \right)^d 2^{-m\epsilon^2/2} \leq \delta,$$

from which it follows that

$$m = O \left( \frac{1}{\epsilon^2} \log \frac{1}{\delta} + \frac{d}{\epsilon^2} \log \frac{1}{\epsilon} \right).$$

Taken together, we have arrived to a fairly remarkable result. Despite the fact that the distribution  $D$  from which the training sample  $S$  is drawn from is *unknown* (but is known to be fixed), the learner simply needs to minimize the empirical error. If the sample size  $m$  is large enough the learner is guaranteed to have minimized the true errors for some accuracy and confidence parameters which define the sample size complexity. Equivalently,

$$|Opt(C) - \hat{err}(h)| \xrightarrow{m \rightarrow \infty} 0.$$

Not only is the convergence independent of  $D$  but also the rate of convergence is independent (namely, it does not matter where the optimal  $h^*$  is located). The latter is very important because without it one could arbitrarily slow down the convergence rate by maliciously choosing  $D$ . The beauty of the results above is that  $D$  does not have an effect at all — one simply needs to choose the sample size to be large enough for the accuracy, confidence and VC dimension of the concept class to be learned over.

## 12.2 Optimality of SVM Revisited

In lectures 6 and 7 we discussed the large margin principle for finding an optimal separating hyperplane. It is natural to ask how does the PAC theory presented so far explains why a maximal margin hyperplane is optimal with regard to the formal sense of learning (i.e. to generalization from empirical errors to true errors)? We saw in the previous section that the sample complexity  $m(\epsilon, \delta, d)$  depends also on the VC dimension of the concept class — which is  $n + 1$  for hyperplanes in  $R^n$ . Thus, another natural question that may certainly arise is what is the gain in employing the "kernel trick"? For a fixed  $m$ , mapping the input instance space  $X$  of dimension  $n$  to some higher (exponentially higher) feature space might simply mean that we are compromising the accuracy and confidence of the learner (since the VC dimension is equal to the instance space dimension plus 1).

Given a fixed sample size  $m$ , the best the learner can do is to minimize the empirical error *and at the same time to try to minimize the VC dimension  $d$  of the concept class*. The smaller  $d$  is, for a fixed  $m$ , the higher the accuracy and confidence of the learning algorithm. Likewise, the smaller  $d$  is, for a fixed accuracy and confidence values, the smaller sample size is required.

There are two possible ways to decrease  $d$ . First is to decrease the dimension  $n$  of the instance space  $X$ . This amounts to "feature selection", namely find a subset of coordinates that are the most "relevant" to the learning task or perform a dimensionality reduction via PCA, for example. A second approach is to maximize the margin. Let the margin associated with the separating hyperplane  $h$  (i.e. consistent with the sample  $S$ ) be  $\gamma$ . Let the input vectors  $\mathbf{x} \in X$  have a bounded norm,  $|\mathbf{x}| \leq R$ . It can be shown that the VC dimension of the concept class  $C_\gamma$  of hyperplanes with margin  $\gamma$  is:

$$C_\gamma = \min \left\{ \frac{R^2}{\gamma^2}, n \right\} + 1.$$

Thus, if the margin is very small then the VC dimension remains  $n + 1$ . As the margin gets larger, there comes a point where  $R^2/\gamma^2 < n$  and as a result the VC dimension decreases. Moreover, mapping the instance space  $X$  to some higher dimension feature space will not change the VC dimension as long as the margin remains the same. It is expected that the margin will not scale down or will not scale down as rapidly as the scaling up of dimension from image space to feature space.

To conclude, maximizing the margin (while minimizing the empirical error) is advantageous as it decreases the VC dimension of the concept class and causes the accuracy and confidence values of the learner to be largely immune to dimension scaling up while employing the kernel trick.