67577 – Intro. to Machine Learning **Fall semester, 2008/9**

Lecture 11: PAC II

Lecturer: Amnon Shashua *Scribe: Amnon Shashua*[1]

The result of the PAC model (also known as the "formal" learning model) is that if the concept class C is PAC-learnable then the learning strategy must simply consist of gathering a sufficiently large training sample S of size $m > m_o(\epsilon, \delta)$, for given accuracy $\epsilon > 0$ and confidence $0 < \delta < 1$ parameters, and finds a hypothesis $h \in C$ which is consistent with S. The learning algorithm is then guaranteed to have a bounded error $err(h) < \epsilon$ with probability $1 - \delta$. The error measurement includes data *not seen* by the training phase.

This state of affair also holds (with some slight modifications on the sample complexity bounds) when there is no consistent hypothesis (the unrealizable case). In this case the learner simply needs to minimize the empirical error $\hat{err}(h)$ on the sample training data S, and if m is sufficiently large then the learner is guaranteed to have $err(h) < Opt(C) + \epsilon$ with probability $1 - \delta$. The measure $Opt(C)$ is defined as the minimal $err(g)$ over all $g \in C$. Note that in the realizable case $Opt(C) = 0$. More details in Lecture 10.

The property of bounding the true error $err(h)$ by minimizing the sample error $\hat{err}(h)$ is very convenient. The fundamental question is *under what conditions this type of generalization property applies?* We saw in Lecture 10 that a satisfactorily answer can be provided when the cardinality of the concept space is bounded, i.e. $|C| < \infty$, which happens for Boolean concept space for example. In that lecture we have proven that:

$$m_o(\epsilon, \delta) = O(\frac{1}{\epsilon} \ln \frac{|C|}{\delta}),$$

is sufficient for guaranteeing a learning model in the formal sense, i.e., which has the generalization property described above.

In this lecture and the one that follows we have two goals in mind. First is to generalize the result of finite concept class cardinality to infinite cardinality — note that the bound above is not meaningful when $|C| = \infty$. Can we learn in the formal sense any non-trivial infinite concept class? (we already saw an example of a PAC-learnable infinite concept class which is the class of axes aligned rectangles). In order to answer this question we will need to a general measure of concept class complexity which will replace the cardinality term $|C|$ in the sample complexity bound $m_o(\epsilon, \delta)$. It is tempting to assume that the number of parameters which fully describe the concepts of C can serve as such a measure, but we will show that in fact one needs a more powerful measure called the *Vapnik-Chervonenkis* (VC) dimension. Our second goal is to pave the way and provide the theoretical foundation for the *large margin principle* algorithm (SVM) we derived in lectures 6,7.

11.1 The VC Dimension

The basic principle behind the VC dimension measure is that although C may have infinite cardinality, the restriction of the application of concepts in C to a finite sample S has a finite outcome.

[1]

This outcome is typically governed by an exponential growth with the size m of the sample S — but not always. The point at which the growth stops being exponential is when the "complexity" of the concept class C has exhausted itself, in a manner of speaking.

We will assume C is a concept class over the instance space X — both of which can be infinite. We also assume that the concept class maps instances in X to $\{0,1\}$, i.e., the input instances are mapped to "positive" or "negative" labels. A training sample S is drawn i.i.d according to some fixed but unknown distribution D and S consists of m instances $\mathbf{x}_1, ..., \mathbf{x}_m$. In our notations we will try to reserve $c \in C$ to denote the target concept and $h \in C$ to denote *some* concept. We begin with the following definition:

Definition 1

$$\Pi_C(S) = \{(h(\mathbf{x}_1), ..., h(\mathbf{x}_m) \; : \; h \in C\}$$

which is a set of vectors in $\{0,1\}^m$.

$\Pi_C(S)$ is set whose members are m-dimensional Boolean vectors induced by functions of C. These members are often called dichotomies or behaviors on S induced or realized by C. If C makes a full realization then $\Pi_C(S)$ will have 2^m members. An equivalent description is a collection of subsets of S:

$$\Pi_C(S) = \{h \cap S \; : \; h \in C\}$$

where each $h \in C$ makes a partition of S into two sets — the positive and negative points. The set $\Pi_C(S)$ contains therefore subsets of S (the positive points of S under h). A full realization will provide $\sum_{i=0}^{m} \binom{m}{i} = 2^m$. We will use both descriptions of $\Pi_C(S)$ as a collection of subsets of S and as a set of vectors interchangeably.

Definition 2 *If* $|\Pi_C(S)| = 2^m$ *then* S *is considered* **shattered** *by* C. *In other words,* S *is shattered by* C *if* C *realizes all possible dichotomies of* S.

Consider as an example a finite concept class $C = \{c_1, ..., c_4\}$ applied to three instance vectors with the results:

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
c_1	1	1	1
c_2	0	1	1
c_3	1	0	0
c_4	0	0	0

Then,

$$\begin{array}{ll} \Pi_C(\{\mathbf{x}_1\}) = \{(0),(1)\} & \text{shattered} \\ \Pi_C(\{\mathbf{x}_1, \mathbf{x}_3\}) = \{(0,0),(0,1),(1,0),(1,1)\} & \text{shattered} \\ \Pi_C(\{\mathbf{x}_2, \mathbf{x}_3\}) = \{(0,0),(1,1)\} & \text{not shattered} \end{array}$$

With these definitions we are ready to describe the measure of concept class complexity.

Definition 3 (VC dimension) *The VC dimension of* C*, noted as* $VCdim(C)$*, is the cardinality* d *of the largest set* S *shattered by* C*. If all sets* S *(arbitrarily large) can be shattered by* C*, then* $VCdim(C) = \infty$.

$$VCdim(C) = \max\{d \mid \exists |S| = d, \; and \; |\Pi_C(S)| = 2^d\}$$

The VC dimension of a class of functions C is the point d at which *all* samples S with cardinality $|S| > d$ are *no longer shattered* by C. As long as C shatters S it manifests its full "richness" in the sense that one can obtain from S all possible results (dichotomies). Once that ceases to hold, i.e., when $|S| > d$, it means that C has "exhausted" its richness (complexity). An infinite VC dimension means that C maintains full richness for all sample sizes. Therefore, the VC dimension is a combinatorial measure of a function class complexity.

Before we consider a number of examples of geometric concept classes and their VC dimension, it is important clarify the lower and upper bounds (existential and universal quantifiers) in the definition of VC dimension. The VC dimension is at least d if there *exists* some sample $|S| = d$ which is shattered by C — this does not mean that all samples of size d are shattered by C. Conversely, in order to show that the VC dimension is *at most* d, one must show that no sample of size $d + 1$ is shattered. Naturally, proving an upper bound is more difficult than proving the lower bound on the VC dimension. The following examples are shown in a "hand waiving" style and are not meant to form rigorous proofs of the stated bounds — they are shown for illustrative purposes only.

Intervals of the real line: The concept class C is governed by two parameters α_1, α_2 in the closed interval $[0, 1]$. A concept from this class will tag an input instance $0 < x < 1$ as positive if $\alpha_1 \leq x \leq \alpha_2$ and negative otherwise. The VC dimension is at least 2: select a sample of 2 points x_1, x_2 positioned in the open interval $(0, 1)$. We need to show that there are values of α_1, α_2 which realize all the possible four dichotomies $(+, +), (-, -), (+, -), (-, +)$. This is clearly possible as one can place the interval $[\alpha_1, \alpha_2]$ such the intersection with the interval $[x_1, x_2]$ is null, (thus producing $(-, -)$), or to fully include $[x_1, x_2]$ (thus producing $(+, +)$) or to partially intersect $[x_1, x_2]$ such that x_1 or x_2 are excluded (thus producing the remaining two dichotomies). To show that the VC dimension is at most 2, we need to show that *any* sample of three points x_1, x_2, x_3 on the line $(0, 1)$ cannot be shattered. It is sufficient to show that one of the dichotomies is not realizable: the labeling $(+, -, +)$ cannot be realizable by any interval $[\alpha_1, \alpha_2]$ — this is because if x_1, x_3 are labeled positive then by definition the interval $[\alpha_1, \alpha_2]$ must fully include the interval $[x_1, x_3]$ and since $x_1 < x_2 < x_3$ then x_2 must be labeled positive as well. Thus $VCdim(C) = 2$.

Axes-aligned rectangles in the plane: We have seen this concept class in Lecture 2 — a point in the plane is labeled positive if it lies in an axes-aligned rectangle. The concept class C is thus governed by 4 parameters. The VC dimension is at least 4: consider a configuration of 4 input points arranged in a cross pattern (recall that we need only to show *some* sample S that can be shattered). We can place the rectangles (concepts of the class C) such that all 16 dichotomies can be realized (for example, placing the rectangle to include the vertical pair of points and exclude the horizontal pair of points would induce the labeling $(+, -, +, -)$). It is important to note that in this case, not all configurations of 4 points can be shattered — but to prove a lower bound it is sufficient to show the existence of a single shattered set of 4 points. To show that the VC dimension is at most 4, we need to prove that any set of 5 points cannot be shattered. For any set of 5 points there must be some point that is "internal", i.e., is neither the extreme left, right, top or bottom point of the five. If we label this internal point as negative and the remaining 4 points as positive then there is no axes-aligned rectangle (concept) which cold realize this labeling (because if the external 4 points are labeled positive then they must be fully within the concept rectangle, but then the internal point must also be included in the rectangle and thus labeled positive as well).

Separating hyperplanes: Consider first linear half spaces in the plane. The lower bound on the

VC dimension is 3 since any three (non-collinear) points in R^2 can be shattered, i.e., all 8 possible labelings of the three points can be realized by placing a separating line appropriately. By having one of the points on one side of the line and the other two on the other side we can realize 3 dichotomies and by placing the line such that all three points are on the same side will realize the 4th. The remaining 4 dichotomies are realized by a sign flip of the four previous cases. To show that the upper bound is also 3, we need to show that no set of 4 points can be shattered. We consider two cases: (i) the four points form a convex region, i.e., lie on the convex hull defined by the 4 points, (ii) three of the 4 points define the convex hull and the 4th point is internal. In the first case, the labeling which is positive for one diagonal pair and negative to the other pair cannot be realized by a separating line. In the second case, a labeling which is positive for the three hull points and negative for the interior point cannot be realize. Thus, the VC dimension is 3 and in general the VC dimension for separating hyperplanes in R^n is $n + 1$.

Union of a finite number of intervals on the line: This is an example of a concept class with an infinite VC dimension. For any sample of points on the line, one can place a sufficient number of intervals to realize any labeling.

The examples so far were simple enough that one might get the wrong impression that there is a correlation between the number of parameters required to describe concepts of the class and the VC dimension. As a counter example, consider the two parameter concept class:

$$C = \{sign(\sin(\omega x + \theta) : \omega\}$$

which has an infinite VC dimension as one can show that for every set of m points on the line one can realize all possible labelings by choosing a sufficiently large value of ω (which serves as the frequency of the sync function) and appropriate phase.

We conclude this section with the following claim:

Theorem 1 *The VC dimension of a finite concept class $|C| < \infty$ is bounded from above:*

$$VCdim(C) \leq \log_2 |C|.$$

Proof: if $VCdim(C) = d$ then there exists at least 2^d functions in C because every function induces a labeling and there are at least 2^d labelings. Thus, from $|C| \geq 2^d$ follows that $d \leq \log_2 |C|$. □

11.2 The Relation between VC dimension and PAC Learning

We saw that the VC dimension is a combinatorial measure of concept class complexity and we would like to have it replace the cardinality term in the sample complexity bound. The first result of interest is to show that if the VC dimension of the concept class is infinite then the class is not PAC learnable.

Theorem 2 *Concept class C with $VCdim(C) = \infty$ is not learnable in the formal sense.*

Proof: Assume the contrary that C is PAC learnable. Let L be the learning algorithm and m be the number of training examples required to learn the concept class with accuracy $\epsilon = 0.1$ and $1 - \delta = 0.9$. That is, after seeing at least $m(\epsilon, \delta)$ training examples, the learner generates a concept h which satisfies $p(err(h) \leq 0.1) \geq 0.9$.

Since the VC dimension is infinite there exist a sample set S with $2m$ instances which is shattered by C. Since the formal model (PAC) applies to any training sample we will use the set S as follows. We will define a probability distribution on the instance space X which is uniform on S (with probability $\frac{1}{2m}$) and zero everywhere else.

Because S is shattered, then any target concept is possible so we will choose our target concept c in the following manner:

$$prob(c_t(\mathbf{x}_i) = 0) = \frac{1}{2} \quad \forall \mathbf{x}_i \in S,$$

in other words, the labels $c_t(\mathbf{x}_i)$ are determined by a coin flip. The learner L selects an i.i.d. sample of m instances $\bar{S}$ — which due to the structure of D means that the $\bar{S} \subset S$ and outputs a consistent hypothesis $h \in C$. The probability of error for each $\mathbf{x}_i \notin \bar{S}$ is:

$$prob(c_t(\mathbf{x}_i) \neq h(\mathbf{x}_i)) = \frac{1}{2}.$$

The reason for that is because S is shattered by C, i.e., we can select any target concept for any labeling of S (the $2m$ examples) therefore we could select the labels of the m points not seen by the learner arbitrarily (by flipping a coin). Regardless of h, the probability of mistake is 0.5. The expectation on the error of h is:

$$E[err(h)] = m \cdot 0 \cdot \frac{1}{2m} + m \cdot \frac{1}{2} \cdot \frac{1}{2m} = \frac{1}{4}.$$

This is because we have $2m$ points to sample (according to D as all other points have zero probability) from which the error on half of them is zero (as h is consistent on the training set $\bar{S}$) and the error on the remaining half is 0.5. Thus, the average error is 0.25. Note that $E[err(h)] = 0.25$ for any choice of ϵ, δ as it is based on the sample size m. For any sample size m we can follow the construction above and generate the learning problem such that if the learner produces a consistent hypothesis the expectation of the error will be 0.25.

The result that $E[err(h)] = 0.25$ is not possible for the accuracy and confidence values we have set: with probability of at least 0.9 we have that $err(h) \leq 0.1$ and with probability 0.1 then $err(h) = \beta$ where $0.1 < \beta \leq 1$. Taking the worst case of $\beta = 1$ we come up with the average error:

$$E[err(h)] \leq 0.9 \cdot 0.1 + 0.1 \cdot 1 = 0.19 < 0.25.$$

We have therefore arrived to a contradiction that C is PAC learnable. □

We next obtain a bound on the growth of $|\Pi_S(C)|$ when the sample size $|S| = m$ is much larger than the VC dimension $VCdim(C) = d$ of the concept class. We will need few more definitions:

Definition 4 (Growth function)

$$\Pi_C(m) = \max\{|\Pi_S(C)| \ : \ |S| = m\}$$

The measure $\Pi_C(m)$ is the maximum number of dichotomies induced by C for samples of size m. As long as $m \leq d$ then $\Pi_C(m) = 2^m$. The question is what happens to the growth pattern of $\Pi_C(m)$ when $m > d$. We will see that the growth becomes polynomial — a fact which is crucial for the learnability of C.

Definition 5 *For any natural numbers m, d we have the following definition:*

$$\begin{aligned} \Phi_d(m) &= \Phi_d(m-1) + \Phi_{d-1}(m-1) \\ \Phi_d(0) &= \Phi_0(m) = 1 \end{aligned}$$

By induction on m, d it is possible to prove the following:

Theorem 3

$$\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$$

Proof: by induction on m, d. For details see Kearns & Vazirani pp. 56.□

For $m \leq d$ we have that $\Phi_d(m) = 2^m$. For $m > d$ we can derive a polynomial upper bound as follows.

$$\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} \leq \sum_{i=0}^{m} \left(\frac{d}{m}\right)^i \binom{m}{i} = (1 + \frac{d}{m})^m \leq e^d$$

From which we obtain:

$$\left(\frac{d}{m}\right)^d \Phi_d(m) \leq e^d.$$

Dividing both sides by $\left(\frac{d}{m}\right)^d$ yields:

$$\Phi_d(m) \leq e^d \left(\frac{m}{d}\right)^d = \left(\frac{em}{d}\right)^d = O(m^d).$$

We need one more result before we are ready to present the main result of this lecture:

Theorem 4 (Sauer's lemma) *If $VCdim(C) = d$, then for any m, $\Pi_C(m) \leq \Phi_d(m)$.*

Proof: By induction on both d, m. For details see Kearns & Vazirani pp. 55–56.□

Taken together, we have now a fairly interesting characterization on how the combinatorial measure of complexity of the concept class C scales up with the sample size m. When the VC dimension of C is infinite the growth is exponential, i.e., $\Pi_C(m) = 2^m$ for all values of m. On the other hand, when the concept class has a bounded VC dimension $VCdim(C) = d < \infty$ then the growth pattern undergoes a discontinuity from an exponential to a polynomial growth:

$$\Pi_C(m) = \left\{ \begin{array}{cc} 2^m & m \leq d \\ \leq (\frac{em}{d})^d & m > d \end{array} \right\}$$

As a direct result of this observation, when $m >> d$ is much larger than d the entropy becomes much smaller than m. Recall than from an information theoretic perspective, the entropy of a random variable Z with discrete values $z_1, ..., z_n$ with probabilities p_i, $i = 1, ..., n$ is defined as:

$$H(Z) = \sum_{i=0}^{n} p_i \log_2 \frac{1}{p_i},$$

where $I(p_i) = \log_2 \frac{1}{p_i}$ is a measure of "information", i.e., is large when p_i is small (meaning that there is much information in the occurrence of an unlikely event) and vanishes when the event is certain $p_i = 1$. The entropy is therefore the expectation of information. Entropy is maximal for a uniform distribution $H(Z) = \log_2 n$. The entropy in information theory context can be viewed as the number of bits required for coding $z_1, ..., z_n$. In coding theory it can be shown that the entropy of a distribution provides the lower bound on the average length of any possible encoding of a uniquely decodable code fro which one symbol goes into one symbol. When the distribution is uniform we will need the maximal number of bits, i.e., one cannot compress the data. In the case of concept class C with VC dimension d, we see that one when $m \leq d$ all possible dichotomies are realized and thus one will need m bits (as there are 2^m dichotomies) for representing all the outcomes of the sample. However, when $m >> d$ only a small fraction of the 2^m dichotomies can be realized, therefore the distribution of outcomes is highly non-uniform and thus one would need much less bits for coding the outcomes of the sample. The technical results which follow are therefore a formal way of expressing in a rigorous manner this simple truth — *If it is possible to compress, then it is possible to learn.* The crucial point is that learnability is a direct consequence of the "phase transition" (from exponential to polynomial) in the growth of the number of dichotomies realized by the concept class.

In the next lecture we will continue to prove the "double sampling" theorem which derives the sample size complexity as a function of the VC dimension.