Tuesday 15th January, 2008, 09:06 EST

Statistical and Computational Foundations of Machine Learning

Shai Ben-David

David R. Cheriton School of Computer Science
University of Waterloo

200 University Avenue West
Waterloo N2L 3G1
Ontario, Canada

Contents

Introduction

This is a compilation of the letures notes from several runs of course taught at Technion in winter 2000, Cornell University in fall 2003, and at University of Waterloo in winters 2005, 2006, and 2007.

It is far from complete and probably contains many mistakes.

Chapter 1

Preliminary Notation and Basic Observations

1.1 Learners Input

- X is the *domain set*.
- Y is the *label set*. For our discussion, we will restrict Y to be $\{0, 1\}$.
- The *training data* or the *sample*, $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$, is a finite sequence of pairs in $X \times Y$.

1.2 Learners Output—"hypothesis"

A function $h : X \to Y$, that is $h : X \to \{0, 1\}$, is called a *hypothesis* or *prediction function*. (Note that h is deterministic, for simplicity.) Our goal is to find some h with small error. Our next step is to define precisely what we mean by "small error"".

1.2.1 Measure of hypothesis error

Let P be a probability distribution over $X \times Y$. Namely, P assigns probability to pairs (x, y), where $x \in X$ and $y \in Y$ is a label for the point x. Having such a P, one can measure how likely is h to make an error when labeled points are randomly drawn according to P:

$$\mathrm{Err}^P(h) = \Pr_{(x,y)\sim P}(h(x) \neq y) \ .$$

When P is the data-generating distribution, $\mathrm{Err}^P(h)$ is called the *test error* or *true error* of h. We would like to find a predictor, h, for which that error will be minimized. However, the learner does not know the data-generating distribution P. What the learner does have access to, is the training data, S.

The quantity

$$\mathrm{Err}^S(h) = \frac{|\{i \ : \ 1 \le i \le m, \ h(x_i) \ne y_i\}|}{m}$$

is called the *empirical error* or the *training error* of h on S.

Given S, a learner can compute $\mathrm{Err}^S(h)$ for any function $h : X \to \{0,1\}$. Note that $\mathrm{Err}^S(h) = \mathrm{Err}^{P(\text{uniform over } S)}(h)$.

1.3 Basic Assumption

To have any chance of success we must make some assumptions about the relationship between S and P. We shall assume that the training data S is independent and identically distributed (i.i.d.) by P—the probability distribution that determines the true error. In other words, we assume that the data given to the learner for training, S, is generated by the same procedure that generates the data on which the learner's conclusion, h, will be evaluated.

1.4 Optimal Predictor

Given any probability distribution P over $X \times \{0,1\}$, the best label predicting function from will be $h^* : X \to \{0,1\}$,

$$h^*(x) = \begin{cases} 1 & \text{if } \Pr(y = 1 \mid x) \ge 1/2 \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

Unfortunately, since we do not know P, we cannot utilize this optimal predictor h^*.

1.5 Law of Large Numbers

Theorem 1 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be real random variables that are identically and independently distributed (i.i.d.). Assume these random variables have finite mean μ and have a finite variance. Then, for every $\epsilon > 0$,*

$$\lim_{m \to \infty} \Pr\left(\left| \frac{1}{m} \sum_{i=1}^{m} X_i - \mu \right| > \epsilon \right) = 0 \ .$$

Now, given a probability distribution, P, over $X \times \{0,1\}$, we can define, for every predictor $h : X \to \{0,1\}$, a random variable, $X^h = X^h(x,y)$ by drawing a pair (x,y) according to P and setting

$$X^h(x,y) = \begin{cases} 1 & \text{if } h(x) \ne y \ , \\ 0 & \text{if } h(x) = y \ . \end{cases}$$

Note that, the mean of this random variable is just $\mathbb{E}(X^h) = \mathrm{Err}^P(h)$, and for a random sample S of size m, drawn i.i.d. by P, we get that

$$\mathrm{Err}^S(h) = \frac{1}{m} \sum_{(x,y)\in S} X^h(x,y) \ .$$

Applying the above Law of Large numbers to these variables, we get

Theorem 2. *For any predictor function $h : X \to \{0,1\}$, for every probability distribution P, over $X \times \{0,1\}$, if S is an i.i.d. random sample of size m drawn according to P,*

$$\lim_{m\to\infty} = \Pr_{S\sim P^m}\left(\left|\mathrm{Err}^S(h) - \mathrm{Err}^P(h)\right| > \epsilon\right) = 0 \ .$$

In other words, if one fixes a hypothesis h, then for every data generating distribution, the empirical error of h will converge to its true error, as sample sizes grow to infinity. This is a good start, however, it has two significant drawbacks that make it quite useless for our purposes. First, this is only an asymptotic result. It provides no information about the gap between the empirically estimated error and its true value for any given, finite, sample size. The second issue with this result is that it holds only if h is chosen independently of S. This is not the case we are interested in—we would like to analyze a scenario in which h is chosen as a result of viewing the training sample, S. In the following, we address both these issues.

We shall impose restrictions on the possible h's the learner (or learning algorithm) can choose from and develop results of the form:

$\forall \epsilon, \delta > 0$ there exists an m such that for all P and h (subject to some restrictions)

$$\Pr_{S\sim P^m}\left(\left|\mathrm{Err}^S(h) - \mathrm{Err}^P(h)\right| > \epsilon\right) < \delta \ .$$

We will call ϵ the measure of *accuracy* and δ the measure of *confidence*.

Chapter 2

First Statistical Learning Bounds—Finite Choice for h

Let H be a finite set of prediction functions. Given a sample S, we will bound the probability that there exists an $h \in H$ that looks perfect on the training data but has true error above ϵ. We will start by taking any fixed h and bounding the probability (over the samples) that $\text{Err}^S(h) = 0$ assuming that $\text{Err}^P(h) > \epsilon$.

$$\Pr_{S \sim P^m}(\text{Err}^S(h) = 0) < (1-\epsilon)^m \leq e^{-m\epsilon}$$

Note, that this probability is taken over the i.i.d. random samples of size m.

2.1 The fortunate case—we find a perfect looking predictor

Next, we apply the union bound to establish to bound the probability that for *some* $h \in H$, $\text{Err}^S(h) = 0$ in spite of having $\text{Err}^P(h) > \epsilon$.

$$\Pr_{S \sim P^m}\left[\exists h \in H \text{ s.t. } (\text{Err}^S(h) = 0 \text{ and } \text{Err}^P(h) > \epsilon)\right] < |H| e^{-m\epsilon} .$$

Therefore, if δ is such that $\delta \geq |H| e^{-m\epsilon}$, or, equivalently,

$$\ln(\delta) \geq \ln|H| - m\epsilon$$

or if

$$m \geq \frac{\ln|H| + \ln(1/\delta)}{\epsilon}$$

we get

$$\Pr_{S \sim P^m}\left[\exists h \in H \text{ s.t. } (\text{Err}^S(h) = 0 \text{ and } \text{Err}^P(h) > \epsilon)\right] < \delta .$$

2.2 The general bound for finite H

Next we address the case where H is still finite but one does not assume that $\mathrm{Err}^S(h) = 0$. We will therefore have to apply a slightly stronger argument.

Theorem 3 (Chernoff-Hoeffding Bound)**.** *Let $X_1, X_2, \ldots, X_m$ be independently identically distributed binary valued random variables, with finite expectation be μ. Then, for any $\epsilon > 0$*

$$\Pr\left(\mu - \frac{1}{m}\sum_{i=1}^{m} X_i \geq \epsilon\right) \leq e^{-2m\epsilon^2} ,$$
$$\Pr\left(\frac{1}{m}\sum_{i=1}^{m} X_i - \mu \geq \epsilon\right) \leq e^{-2m\epsilon^2} ,$$
$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2} .$$

For a sample $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$ and a hypothesis h, let

$$X_i = \begin{cases} 1 & \text{if } h(x_i) \neq y_i , \\ 0 & \text{if } h(x_i) = y_i . \end{cases}$$

Note that under this formulation $\mu = \mathrm{Err}^P(h)$ and $\frac{1}{m}\sum_{i=1}^{m} X_i = \mathrm{Err}^S(h)$. Applying Chernoff-Hoeffding bound, for any *fixed* h, we have

$$\Pr_{S\sim P^m}\left[\mathrm{Err}^P(h) \geq \mathrm{Err}^S(h) + \epsilon\right] \leq e^{-2m\epsilon^2} .$$

It follows that the probability that for *some* $h \in H$ the true error is larger by ϵ than its empirical error is bounded by $|H|e^{-2m\epsilon^2}$,

$$\Pr_{S\sim P^m}\left[\exists h \in H \text{ s.t. } \mathrm{Err}^P(h) \geq \mathrm{Err}^S(h) + \epsilon\right] \leq |H|e^{-2m\epsilon^2} .$$

Thus, repeating the same argumentation we had for the previous bound (for the case $\mathrm{Err}^S(h) = 0$), we get, for any finite H, for all $m \in \mathbb{N}$ and for all $\delta > 0$, for every probability distribution P, with probability exceeding $(1 - \delta)$,

$$\forall h \in H \qquad \mathrm{Err}^P(h) \leq \mathrm{Err}^S(h) + \sqrt{\frac{\ln|H| + \ln(1/\delta)}{2m}} .$$

2.3 Empirical Risk Minimization

The above discussion motivates the following learning paradigm, which is called *Empirical Risk Minimization* (ERM):

1. Fix a finite set H of predictors (that is, functions from $X \to \{0,1\}$).

2. Upon viewing the training sample, S, find an h^{ERM} in H that minimizes $\mathrm{Err}^S(h)$ over all h's in H.

3. Use that h^{ERM} to predict the labels of test points.

For such a learning paradigm we can now derive a relative error bound. Namely, we can bound by how much may our chosen predictor h^{ERM} be worse than the best possible predictor h^* in H.

Theorem 4 (ERM for Fininite H)**.** *For every finite set of predictor functions H, and any probability distribution P over $X \times \{0,1\}$, let h^* be an element of H that minimizes the true error $\mathrm{Err}^P(h)$. For any $\delta > 0$, if S is a random i.i.d. sample of size m generated from P, and h^{ERM} is a minimizer of $\mathrm{Err}^S(h)$ over the functions $h \in H$, then with probability exceeding $(1-\delta)$,*

$$\mathrm{Err}^P\left(h^{ERM}\right) \leq \mathrm{Err}^P\left(h^*\right) + 2\sqrt{\frac{\ln|H| + \ln(2/\delta)}{2m}} \,.$$

Proof. For each $h \in H$ using Chernoff-Hoeffding bound we have

$$\Pr_{S\sim P^m}\left[\left|\mathrm{Err}^P(h) - \mathrm{Err}^S(h)\right| \geq \epsilon\right] \leq 2e^{-2m\epsilon^2} \,.$$

By union bound we have

$$\Pr_{S\sim P^m}\left[\exists h \in H \text{ s.t. } \left|\mathrm{Err}^P(h) - \mathrm{Err}^S(h)\right| \geq \epsilon\right] \leq 2|H|e^{-2m\epsilon^2} \,.$$

Setting $\delta = 2|H|e^{-2m\epsilon^2}$ we can express

$$\epsilon = \sqrt{\frac{\ln|H| + \ln(2/\delta)}{2m}} \,.$$

And we have that with probability $1-\delta$ over the random choice of sample S,

$$\forall h \in H \qquad \left|\mathrm{Err}^P(h) - \mathrm{Err}^S(h)\right| \leq \epsilon \,. \tag{2.1}$$

Hence, with probability $1-\delta$ the following chain of inequalities holds

$$\mathrm{Err}^P\left(h^{\mathrm{ERM}}\right) \leq \mathrm{Err}^S\left(h^{\mathrm{ERM}}\right) + \epsilon \leq \mathrm{Err}^S\left(h^*\right) + \epsilon \leq \mathrm{Err}^P\left(h^*\right) + 2\epsilon \,,$$

where the middle inequality follows from that $\mathrm{Err}^S\left(h^{\mathrm{ERM}}\right)$ is a minimizer of $\mathrm{Err}^S(h)$, and the other two inequalities follow from (2.1) for $h = h^{\mathrm{ERM}}$ and $h = h^*$. □

Chapter 3

Minimum Description Length Principle and PAC-Bayes Learning

So far we have considered generalization bounds for finite hypothesis classes. Looking at the proof in Chapter 2 of these bounds, we see that the probability we assigned to the "bad event"—that for a particular $h \in H$ the empirical and true error differ greatly—was $\delta/|H|$. What if we, instead using $1/|H|$ for every h, pick some other "probability distribution" of these weights.[1]

We first present a classical result from coding theory. Recall that a *prefix free code* W over an alphabet Σ is (finite or infinite[2]) subset of Σ^*, such that for any pair of distinct words $w, w' \in W$ none of the two is a prefix of the other one. Traditionally, in coding theory, the words $w \in W$ are called the *codewords*.

Theorem 5 (Kraft's inequality)**.** *A prefix free code W over an finite alphabet Σ satisfies*

$$\sum_{w \in W} |\Sigma|^{-|w|} \leq 1 \; .$$

Proof. Consider the infinite $|\Sigma|$-ary tree where nodes correspond to the words of Σ^*. See Figure 3.1. A codeword $w \in W$ has $|\Sigma|^{d-|w|}$ descendants at level d, (provided that $d \geq |w|$). Since the code W is prefix free, in this tree no codeword can be a descendant of another codeword; and thus no word $u \in \Sigma^*$ can be descendant of two different codewords. Thus, since there are $|\Sigma|^d$ codewords at level d,

$$\sum_{\substack{w \in W \\ |w| \leq d}} |\Sigma|^{d-|w|} \leq |\Sigma|^d \; .$$

[1] There is nothing "really random" that this distribution will be used for right now. Perhaps a better way is to look at them as just weights.

[2] In coding theory, W is usually assumed to be finite. However for us the infinite case is the interesting one.

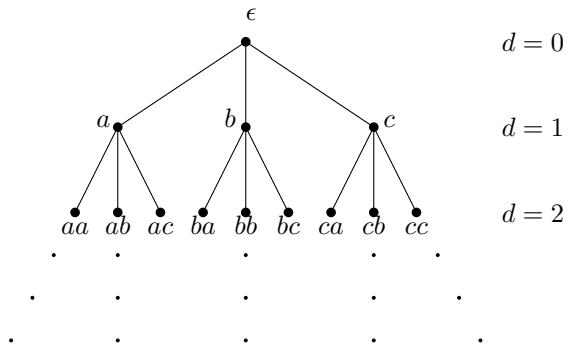

Figure 3.1: Infinite ternary tree for the alphabet $\Sigma = \{a, b, c\}$. The levels of the tree are numbered from the root starting with level $d = 0$. The nodes at level d correspond to the words of length d.

Dividing by $|\Sigma|^d$ we get

$$\sum_{\substack{w \in W \\ |w| \le d}} |\Sigma|^{-|w|} \le 1 \ ,$$

and taking limit $d \to \infty$ we have

$$\sum_{w \in W} |\Sigma|^{-w} = \lim_{d \to \infty} \sum_{\substack{w \in W \\ |w| \le d}} |\Sigma|^{-|w|} \le 1 \ . \quad \square$$

Suppose that H is (countably) infinite and we have a one-to-one mapping (a coding) $\hat{} : H \to \Sigma^*$, $h \mapsto \hat{h}$, which assigns a codeword to each hypothesis h so that the resulting code (the range of $\hat{}$) is a prefix free code. In machine learning $\hat{h}$ is called the *description* of the hypothesis h, and $|\hat{h}|$ is called the *description length* of h. For each hypothesis $h \in H$ we choose $\epsilon_h > 0$ so that the probability of the event $\mathrm{Err}^P(h) - \mathrm{Err}^S(h) > \epsilon_h$ is less than $\delta / |\Sigma|^{|\hat{h}|}$; formally,

$$\Pr_{S \in P^m}[\mathrm{Err}^P(h) - \mathrm{Err}^S(h) > \epsilon_h] < \delta \Sigma^{-|\hat{h}|} \ . \tag{3.1}$$

Applying the union bound we get

$$\begin{aligned} \Pr_{S \in P^m}[\exists h \in H : \mathrm{Err}^P(h) - \mathrm{Err}^S(h) > \epsilon_h] &\le \sum_{h \in H} \Pr_{S \in P^m}[\mathrm{Err}^P(h) - \mathrm{Err}^S(h) > \epsilon_h] \\ &< \sum_{h \in H} \delta |\Sigma|^{-|\hat{h}|} \qquad \text{(from the choice of } \epsilon_h) \\ &\le \delta \qquad \text{(by Kraft's inequality)} \ . \end{aligned}$$

It remains to compute ϵ_h for each hypothesis $h \in H$ so that (3.1) holds. From Hoeffding's inequality we know that

$$\Pr_{S \in P^m}[\mathrm{Err}^P(h) - \mathrm{Err}^S(h) > \epsilon_h] < e^{-2m\epsilon_h^2} \ .$$

Equating the right hand side with $\delta\Sigma^{-|\hat{h}|}$ we have $e^{-2m\epsilon_h^2} = \delta\Sigma^{-|\hat{h}|}$, which we solve for ϵ_h

$$\epsilon_h = \sqrt{\frac{h \ln |\Sigma| + \ln\left(\frac{1}{\delta}\right)}{2m}} \ .$$

We can we phrase our result as a theorem.

Theorem 6 (MDL Principle/Occam's razor)**.** *For any finite or countably infinite hypothesis class H, any coding $h \mapsto \hat{h}$, any $\delta > 0$, any sample size $m \in \mathbb{N}$, and any probability distribution P over $X \times \{0,1\}$ with probability greater than $1-\delta$ over the draw of an i.i.d. sample S from P*

$$\forall h \in H \qquad \mathrm{Err}^P(h) < \mathrm{Err}^S(h) + \sqrt{\frac{h \ln |\Sigma| + \ln\left(\frac{1}{\delta}\right)}{2m}} \ .$$

This theorem has a nice interpretation. Looking at the bound on the true error, $\mathrm{Err}^P(h)$, we see that the bound is better, for hypothesis h which have small empirical error $\mathrm{Err}^S(h)$ (that is the hypothesis h *explains observations well*), and for which the description length $|\hat{h}|$ is small (that is the hypothesis h has *simple description*). This interpretation is called *Occam's razor*[3] or sometimes *minimum description length principle* (MDL principle).

3.1 PAC-Bayesian Learning

Note that, in the above proof, we have assigned "probabilities" (or weights) $|\Sigma|^{-|\hat{h}|}$ to the hypotheses. However any choice of these probabilities keeps the proof valid, as long as they sum to one. (In proof of the MDL theorem this followed from Kraft's inequality.) Thus, one can have some *fixed and known* distribution D over H, assigning each $h \in H$ a probability $D(h)$—a real number in the interval $(0,1]$.[4] If we replace (3.1) and, instead, choose ϵ_h so that

$$\Pr_{S \in P^m}[\mathrm{Err}^P(h) - \mathrm{Err}^S(h) > \epsilon_h] < \delta D(h)$$

and solve $e^{-2m\epsilon_h^2} = \delta D(h)$ for ϵ_h similarly as before, we get

$$\epsilon_h = \sqrt{\frac{\ln\left(\frac{1}{D(h)}\right) + \ln\left(\frac{1}{\delta}\right)}{2m}} \ .$$

We phrase the result as a theorem.

[3] William Occam (c. 1288 - c. 1348)

[4] Do not get confused here. P is the data-generating distribution, which is unknown and can be arbitrary. On the contrary, D is chosen by us and is a distribution over hypotheses.

Theorem 7 (PAC-Bayesian Bound)**.** *For any finite or countably infinite hypothesis class H, any prior distribution D over H, $D : H \to (0, 1]$, any $\delta > 0$, any sample size $m \in \mathbb{N}$, and any probability distribution P over $X \times \{0, 1\}$ with probability greater than $1 - \delta$ over the draw of an i.i.d. sample S from P*

$$\forall h \in H \qquad \mathrm{Err}^P(h) < \mathrm{Err}^S(h) + \sqrt{\frac{\ln\left(\frac{1}{D(h)}\right) + \ln\left(\frac{1}{\delta}\right)}{2m}} \,.$$

Mathematically, there is almost no difference between Theorems 6 and 7. The difference is rather in their interpretation. The interpretation of Theorem 7 is as follows. We pick the so called *prior distribution* D before seeing the training sample S. The distribution D encodes our prior beliefs about the reality, that is, which hypotheses we think should be a priori preferred over the others. On the sample S we measure empirical error of each hypothesis $h \in H$. We will prefer the hypothesis h that minimizes the bound on the true error $\mathrm{Err}^P(h)$ given in Theorem 7. Such hypothesis will have small empirical error $\mathrm{Err}^S(h)$ (that is, *evidence supporting it*) and large prior probability $D(h)$ (high *a priori belief* in its validity). Such reasoning is closely related to Bayes's rule for conditional probabilities

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \,.$$

In our case it could be, very roughly, written as[5]

$$\Pr(h|S) = \frac{\Pr(S|h)D(h)}{\Pr(S)} \,,$$

which can be interpreted in the same fashion as Theorem 7. Namely, hypothesis h are more probable ($\Pr(h|S)$ is higher), when they better explain the data (have higher $\Pr(S|h)$), and we a priori believe that they are more likely (their prior probability $D(h)$ is higher).

[5]In order to make it formally work, we would need to define a *joint* distribution over $H \times (X \times \{0, 1\})^m$.

Chapter 4

Glivenko-Cantelli Theorem

Chapter 5

Definition of VC Dimension

The notion of VC-dimension was introduced by Vladimir Vapnik and Alexey Chervonenkis [11]; similar notions were studied by Saharon Shelah [8] and Norbert Sauer [5]. Before giving its definition, let us define one combinatorial property of sets:

Definition 8. A collection of sets $H \subseteq 2X := \{S : S \subseteq X\}$ *shatters* a set $A \subseteq X$ if

$$\{h \cap A : h \in H\} = 2^A := \{B : B \subseteq A\} .$$

In other words, H shatters A if any subset of A can be obtained by intersecting A with some set from the collection H.

Consider two simple examples:

Example 9. Let $X = \mathbb{R}$ and $H = \{(a, b) : a < b\}$. Consider two sets: $A = \{2, 3\}$ and $A' = \{2, 3, 4\}$. It is easy to see that the class H of all intervals shatters A and does not shatter A' (Fig. 3). Indeed, we can obtain sets $\emptyset$, $\{2\}$, $\{3\}$, and $\{2, 3\}$ by intersecting A with intervals. However, for set A', there is no interval that contains 2 and 4 and does not contain 3; therefore, subset $\{2, 4\}$ cannot be obtained, and A' is not shattered by intervals.

Example 10. Let $X = \mathbb{R}^2$ and

$$H = \{[a, b] \times [c, d] : a, b, c, d \in \mathbb{R},\ a \le b,\ c \le d\}$$

that is, H is the collection of axis-aligned rectangles in $\mathbb{R}^2$. You can see three situations, one with a set that can be shattered and two with a set that cannot be shattered, on Fig. 3.

Now comes the definition of VC-dimension:

Definition 11 (VC-dimension)**.** The *Vapnik-Chervonenkis dimension* (VC-dimension) of a collection H of sets is the cardinality of the largest set that is shattered by H. That is,

$$\text{VC-dim}(H) = \max\{|A| \ :\ A \text{ is shattered by } H\} .$$

If no such largest set exists, then H has infinite VC-dimension.

Let us apply this definition to the two previous examples:

Example 12. Let $X = \mathbb{R}$ and $H = \{(a, b) : a < b\}$ then VC-dim$(H) = 2$. Indeed, as we saw, there is a set of size 2 that is shattered. Now, take any set A of 3 or more points, then A contains three points $x_1 < x_2 < x_3$. The set $B = \{x_1, x_3\}$ cannot be obtained by intersecting A with an interval, because any interval that contains x_1 and x_3 must also contain x_2, but $x_2 \notin B$.

Example 13. Let $X = \mathbb{R}^2$ and let H be the collection of axis-aligned rectangles in $\mathbb{R}^2$; then VC-dim$(H) = 4$. We saw in before that VC-dim$(H) \geq 4$. Consider any set $A \subset \mathbb{R}^2$ of size 5 or more, and take a five-point subset $C \subseteq A$. In C, take a leftmost point (whose first coordinate is the smallest in C), a rightmost point (first coordinate is the largest), a lowest point (second coordinate is the smallest), and a highest point (second coordinate is the largest); let $x \in C$ be the (fifth) point that was not selected. Now, define $B = A \setminus \{x\}$. It is impossible to make B by intersecting A with an axis-aligned rectangle. Indeed, such a rectangle must contain all four selected points in C; but in this case the rectangle contains x as well, because its coordinates are within the intervals spanned by selected points (Fig. 3). So, A is not shattered by H, and therefore VC-dim$(H) = 4$.

Example 14. Let $X = \mathbb{R}$ and H be the collection of all finite subsets of $\mathbb{R}$. Then VC-dim$(H) = \infty$, because any finite set can be shattered by H.

5.1 Dudley's Theorem

We consider *function class*[1] $F \subseteq \mathbb{R}^X$ which is vector space. Recall that a vector space is a set equipped with addition and scalar multiplication. For any two functions $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$, and any scalar $\lambda \in \mathbb{R}$, the sum $f + g$ and scalar multiple λf are defined by

$$(f + g)(x) = f(x) + g(x) \ ,$$
$$(\lambda f)(x) = \lambda \cdot f(x) \ .$$

Any function f defines a subset of X (a hypothesis) where f is positive

$$\text{Pos}(f) = \{x \in X \ : \ f(x) > 0\} \ .$$

Overloading notation, any function class $F \subseteq \mathbb{R}^X$ defines hypothesis class

$$\text{Pos}(F) = \{\text{Pos}(f) : f \in F\} \ .$$

Theorem 15 (Dudley, 19??)**.** *Function class $F \subseteq X^R$, which is a real vector space,*

$$\dim(F) = \text{VC-dim}(\text{Pos}(F)) \ ,$$

where $\dim(F)$ *denotes the linear dimension of F.*

[1]The notation $\mathbb{R}^X$ denotes the set of all function from X to $\mathbb{R}$.

Before we prove Dudley's theorem, we look at some examples.

Example 16 (Halfspaces). The function class of affine functions over $\mathbb{R}^d$,

$$F = \{f : \mathbb{R}^d \to \mathbb{R} \ : \ f(x_1, x_2, \ldots, x_d) = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_d x_d, \ a_0, a_1, a_2, \ldots, a_d \in \mathbb{R}\}$$

is a vector space of dimension $d + 1$. Dudley's theorem implies that the class of half-spaces $H = \text{Pos}(F)$ has VC-dim$(H) = d + 1$. (The hypothesis class contains, together with halfspaces, the empty hypothesis $\emptyset$ and whole $\mathbb{R}^d$.)

Example 17 (Homogeneous Halfspaces). The function class of homogeneous linear functions over $\mathbb{R}^d$,

$$F = \{f : \mathbb{R}^d \to \mathbb{R} \ : \ f(x_1, x_2, \ldots, x_d) = a_1 x_1 + a_2 x_2 + \cdots + a_d x_d, \ a_1, a_2, \ldots, a_d \in \mathbb{R}\}$$

is a vector space of dimension d. Dudley's theorem implies that the class of homogeneous half-spaces $H = \text{Pos}(F)$ has VC-dim$(H) = d$. (The hypothesis class contains, together with halfspaces passing through the origin, the empty hypothesis.)

Proof of Dudley's Theorem. Note that for any subset $A \subseteq X$ is $F|_A \subseteq \mathbb{R}^A$ a vector space. We shall show that

$$A \textit{ is shattered by } \text{Pos}(F) \textit{ if and only if the linear dimension of } F|_A \textit{ is } |A|. \qquad (*)$$

If linear dimension of $F|_A$ is $|A|$, it follows that $F|_A = \mathbb{R}^d$. This means that on A the functions in F attain any combination of values, and hence in particular any combination of positive and negative signs, and thus Pos(F) shatters A.

We prove the opposite implication of (*) by contradiction. We assume that $\dim(F|_A) < |A|$ and A is shattered by Pos(F). It means that $F|_A$ is a *proper* vector subspace of the vector space $\mathbb{R}^A$. Then, there exists a non-zero function $h \in \mathbb{R}^A$ which is orthogonal to vector subspace $F|_A$ with respect to the standard dot product

$$(f, g) = \sum_{x \in A} f(x)g(x) \ .$$

In other words, for every $f \in F|_A$ we have $\sum_{x \in A} f(x)h(x) = 0$. Without loss of generality we may assume that $h(x_0)$ is positive for at least one $x_0 \in A$, otherwise $(-h)$ would have this property and still be orthogonal to all functions in $F|_A$. For every $f \in F|_A$, Pos$(h) \neq$ Pos(f), since otherwise $\sum_{x \in A} f(x)h(x)$ would consist of non-negative summands and at least one positive summand $f(x_0)h(x_0)$, and could not be equal to zero. Therefore, the subset Pos$(h) \subseteq A$ can not be cut by Pos$(F|_A)$ from A and hence it can not be be cut by Pos(F), which is a contradiction.

Using (*) we finish the proof. Since there is a shattered set of size VC-dim(Pos(F)), it follows that linear dimension of $F|_A$ is VC-dim(Pos(F)) and hence the linear dimension of F is at least VC-dim(Pos(F)). On the other hand, there exists a set A of size $\dim(F)$ such that $F|_A$ has linear dimension A, and hence A is a shattered set of size $\dim(F)$. □

Remark 18. For any fixed function $g : X \to \mathbb{R}$ and any function class $F \subseteq \mathbb{R}^X$, which is a vector space

$$\text{VC-dim}(\text{Pos}(g + F)) = \text{VC-dim}(\text{Pos}(F)) \ ,$$

where $g + F = \{h + f \ : \ f \in F\}$ is a shift of the vector space F.[2]

Proof. The remark is a slight strengthening of Dudley's theorem. It suffices to notice that $(g + F)|_A = \mathbb{R}^d$ □

Example 19 (Discs in $\mathbb{R}^2$). The hypothesis class $\text{Pos}(F')$ is the set of open discs in $\mathbb{R}^2$, where

$$F' = \{(x-a)^2 + (y-b)^2 - r^2 \ : \ a, b, r \in \mathbb{R}\} = \{\underbrace{x^2 + y^2}_{h} + \underbrace{\alpha x + \beta y + \gamma}_{f \in F} \ : \alpha, \beta, \gamma \in \mathbb{R}\} \ .$$

Since $F = \{\alpha x + \beta y + \gamma : \alpha, \beta, \gamma \in F\}$ is vector space of dimension 3, the Vapnik-Chervonenkis dimension of open discs is 3.

[2]Shift of an vector space is usually called *affine* space. However, it this might sound confusing that in Example 16 the space of affine functions is a non-shifted vector space.

Chapter 6

Shatter Function and Sauer's Lemma

The shatter function is another combinatorial notion which is closely related to the notion of VC-dimension. This function maps natural numbers to natural numbers, and it measures the "shattering ability" of a given collection of sets.

Definition 20. Given X and the collection $H \subseteq 2^X$, the shatter function of H, $\tau_H : \mathbb{N} \to \mathbb{N}$, is defined as

$$\tau_H(m) = \max_{|A|=m} |\{h \cap A \ : \ h \in H\}| \ .$$

In other words, $\tau_H(m)$ is the largest number of subsets one can get from an m-element set by intersecting it with sets from H.

If there is a set of size m that is shattered by H, then VC-dim$(H) > m$ and $\tau_H(m) = 2^m$. If there is no such set, then VC-dim$(H) < m$ and $H(m) < 2m$. [1] It turns out that for $m >$ VC-dim(H) not only $\tau_H(m) < 2m$, but in fact $\tau_H(m) =$ poly(m). This result is formulated in the following lemma:

Lemma 21 (Sauer, 1972)**.** *If H has a finite* VC-dim$(H) = d$*, then* $\forall m \in \mathbb{N}$, $\tau_H(m) \leq m^d + 1$.

Proof. We are going to prove even a stronger statement, namely: *For any finite set A and a collection of sets H,*

$$|\{h \cap A : h \in H\}| \leq |\{B \subseteq A : H \text{ shatters } B\}| \ . \tag{6.1}$$

This implies Sauers lemma, because if VC-dim$(H) = d$ then

$$|\{B \subseteq A : H \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{|A|}{i} \leq |A|^d \ .$$

We prove (6.1) by induction on the size of A. In the base case $A = \emptyset$, if H is non-emtpy, then both parts of the inequality are equal to 1, and if $H = \emptyset$, then both parts of the inequalityare equal to 0.

[1] Of course, if there is no set of size m that is shattered, then there is no set of size greater than m that is shattered: all subsets of a shattered set are themselves shattered.

Now, let $A \neq \emptyset$ and assume that (6.1) is true for all sets smaller than A and all collections of sets. Pick $x \in A$ and let $A' = A \setminus \{x\}$. Given H, define three new collections of sets

$$\begin{aligned} H_A &= \{h \cap A \ : \ h \in H\} \ , \\ H_{A'} &= \{h \cap A' \ : \ h \in H\} \ , \\ H^x_{A'} &= \{B \subseteq A' \ : \ B \in H_A \text{ and } B \cup \{x\} \in H_A\} \ . \end{aligned}$$

For every subset $B \subseteq A'$, there are four possibilities:

1. $B \in H_A$ and $B \cup \{x\} \in H_A$; then $B \in H_{A'}$ and gives two counts for $|H_A|$.
2. $B \in H_A$, but $B \cup \{x\} \notin H_A$; then $B \in H_{A'}$ and gives one count for $|H_A|$.
3. $B \notin H_A$, but $B \cup \{x\} \in H_A$; then $B \in H_{A'}$ and gives one count for $|H_A|$.
4. $B \notin H_A$ and $B \cup \{x\} \notin H_A$; then $B \notin H_{A'}$ and gives zero counts for $|H_A|$.

In the first case, B belongs to both $H_{A'}$ and $H^x_{A'}$; in the second and third cases, B belongs only to $H_{A'}$, but not to $H^x_{A'}$. Therefore, we have

$$|H_A| = |H_{A'}| + |H^x_{A'}| \ . \tag{6.2}$$

Now we apply the induction hypothesis to A' and two collections $H_{A'}$ and $H^x_{A'}$. Note that both collections contain only subsets of A', making $h \cap A' = h$. We obtain

$$\begin{aligned} |H_{A'}| &= |\{h \cap A' \ : \ h \in H_{A'}\}| \\ &\leq |\{B \subseteq A \ : \ H_{A'} \text{ shatters } B\}| \\ &= |\{B \subseteq (A \setminus \{x\}) \ : \ H \text{ shatters } B\}| \\ &= |\{B \subseteq A \ : \ x \notin B \text{ and } H \text{ shatters } B\}| \end{aligned}$$

and

$$\begin{aligned} |H^x_{A'}| &= |\{h \cap A' \ : \ h \in H^x_{A'}\}| \\ &\leq |\{B \subseteq A' \ : \ H^x_{A'} \text{ shatters } B\}| \\ &\leq |\{B \subseteq (A \setminus \{x\}) \ : \ H \text{ shatters } B \cup \{x\}\}| \\ &= |\{B \subseteq A \ : \ x \in B \text{ and } H \text{ shatters } B\}| \ . \end{aligned}$$

Combining these with (6.2), we finally obtain

$$|H_A| = |H_{A'}| + |H^x_{A'}| \leq |\{B \subseteq A \ : \ H \text{ shatters } B\}| \ ,$$

which completes the induction. □

Example 22. Suppose $A \subset \mathbb{R}^2$, $|A| = 10\,000$, and $H = HS^2$. By Sauer's lemma, there are at most $10\,000^3 = 10^{12}$ linearly separable subsets of A, out of $2^{10\,000} \approx 10^{3\,000}$ possible subsets.

Chapter 7

ϵ-nets and ϵ-approximations

Definition 23 (ϵ-net, ϵ-approximation)**.** Let P be a probability distribution over a domain X, let $H \subseteq 2^X$ be a family of (P-measurable) subsets of X, and let $\epsilon > 0$. A finite non-empty set $A \subseteq X$ is called an *ϵ-net* of P for H, when for every subset $h \in H$, $P(h) > \epsilon$ implies that $A \cap h$ is non-empty. Moreover, if for every $h \in H$

$$\left| \frac{|A \cap h|}{|A|} - P(h) \right| \leq \epsilon \, ,$$

then A is called an *ϵ-approximation* of P for h.

Note that if A is an ϵ-approximation of P for H, then it is also an ϵ-net of P for H. Indeed, $P(h) > \epsilon$ and $||A \cap h|/|A| - P(h)| \leq \epsilon$ together imply that $|A \cap h|/|A| > 0$ and hence $A \cap h$ is non-empty.

The fundamental fact is that, if H has finite Vapnik-Chervonenkis dimension, then a small subset generated i.i.d. from P is with high probability an ϵ-net, respectively an ϵ-approximation, of P for H. The good news is that required size m of the ϵ-net (respectively the ϵ-approximation) does not depeden on P, and depends only on ϵ, VC-dim(H) and the probability (δ) we require.

Theorem 24 (ϵ-net)**.** *Let H a be class of subsets of a domain set X with* VC-dim$(H) = d$. *For any $\delta, \epsilon \in (0,1)$, and any probability distribution P over X, with probability $1 - \delta$ a random i.i.d. sample S of size*

$$m \geq \max\left(\frac{4}{\epsilon} \ln \frac{2}{\delta}, \ \frac{8d}{\epsilon} \ln \frac{8d}{\epsilon} \right)$$

generated from P is an ϵ-net of P for H.

Proof. Consider the event E_1,[1]

$$E_1 = \exists h \in H \text{ s.t. } P(h) > \epsilon \text{ and } h \cap S = \emptyset \, .$$

[1]Technically, $E_1 \subseteq X^m$ is a set, that is, $E_1 = \{S \in X^m \ : \ \exists h \in H \text{ s.t. } P(h) > \epsilon \text{ and } h \cap S = \emptyset\}$. We prefer to think of E_1 as a predicate.

Our task is to show that $\Pr_{S\sim P^m}[E_1] \leq \delta$. To achieve this we use the following trick. We make an additional random choice of a second random i.i.d. sample T of size m generated from P. (The sample T is independent from S.) We consider the event E_2,

$$E_2 = \exists h \in H \text{ s.t. } P(h) > \epsilon \text{ and } h \cap S = \emptyset \text{ and } |h \cap T| \geq \frac{\epsilon m}{2} .$$

First, we show that

$$\Pr_{\substack{S\sim P^m \\ T\sim P^m}}[E_2] \geq \frac{1}{2} \Pr_{S\sim P^m}[E_1] . \tag{7.1}$$

Since E_2 is a subevent of E_1, we show that if E_1 is true for S, then E_2 is true for S and T with probability at least $1/2$ over the choice of T. In other words, we show that the conditional probability $\Pr[E_2|E_1] \geq 1/2$. Indeed, if E_1 is true, then there is some $h \in H$ with $P(h) > \epsilon$. Clearly, if $|h \cap T| \geq \epsilon m/2$ for the particular h chosen above, then E_2 holds as well. Hence, to show that $\Pr[E_2|E_1] \geq 1/2$ we show instead that $\Pr_{T\sim P^m}[|h \cap T| \geq \epsilon m/2] \geq 1/2$, or equivalently, that $\Pr_{T\sim P^m}[|h \cap T| < \epsilon m/2] \leq 1/2$. The size $|h \cap T|$ is binomially distributed[2] with mean value $\mu = \mathbb{E}(|h \cap T|) = \epsilon m$ and variance $\sigma^2 = \mathrm{Var}(|h \cap T|) = \epsilon(1-\epsilon)m \leq \epsilon m$. From Chebyshev's inequality

$$\Pr_{T\sim P^m}[|h \cap T| < \epsilon m/2] \leq \frac{\sigma^2}{(\epsilon m/2)^2} \leq \frac{\epsilon m}{4(\epsilon m)^2} = \frac{1}{4\epsilon m} \leq \frac{1}{2} .$$

The last inequality follows from the assumption $m \geq \frac{4}{\epsilon} \log \frac{2}{\delta} > \frac{1}{4\epsilon}$.

Second, in order to show that $\Pr_{S\sim P^m}[E_1]$ is small, we show that $\Pr_{S\sim P^m, T\sim P^m}[E_2]$ is small. The random choice of S and T can be described in the following way. We first draw a double sample R of size $2m$ i.i.d. according to P, and then we choose randomly from R a subset of size m to be the set S, and the remaining elements of R will form the set T.

For each hypothesis $h \in H$ let E_h be the event

$$E_h = h \cap S = \emptyset \text{ and } |h \cap T| \geq \frac{\epsilon m}{2} .$$

Assuming R is fixed, what is the probability $\Pr_{S\subseteq R}[E_h]$ of E_h? If $|h \cap R| < \epsilon m/2$, then this probability is zero, since $|h \cap T| \leq |h \cap R|$. Otherwise, let $p = |h \cap R| \geq \epsilon m/2$ be the number of points of R inside h. Then, the event E_h is the same as requiring that we do not include in S points from $h \cap R$, probability of which is precisely

$$\frac{\binom{2m-p}{m}}{\binom{2m}{m}} = \frac{(2m-p)(2m-p)\cdots(m-p+1)}{2m(2m-1)\cdots(m+1)} = \frac{m(m-1)\cdots(m-p+1)}{2m(2m-1)\cdots(2m-p+1)} \leq 2^{-p} \leq 2^{-\epsilon m/2} ,$$

since there are in total $\binom{2m}{m}$ choices for S, and in order not to hit h, S must be chosen from the set $R - h$ of size $|R - h| = 2m - p$.

[2]The number of heads in m independent tosses of a coin with probability of head p and probability of tail $1-p$ is a binomially distributed random variable. The expected number of heads is mp and the variance of heads is $p(1-p)m$. The name *binomial distribution* comes from the fact that probability of getting k heads is $\binom{m}{k}p^k(1-p)^{m-k}$.

As before assume that R is fixed. Note that E_2 is an union of (some of the) events E_h, $h \in H$. If $h \cap R = h' \cap R$ for two hypotheses $h, h' \in H$, then the events E_h and $E_{h'}$ are equal. By Sauer's lemma there are at most $\binom{2m}{\leq d}$ different possibilities for $h \cap R$. Hence given R, the probability $\Pr_{S \subseteq R}[E_2]$ is at most $\binom{2m}{\leq d} 2^{-\epsilon m/2}$. Since this is true for any R, we conclude that

$$\Pr_{\substack{S \sim P^m \\ T \sim P^m}}[E_2] \leq \binom{2m}{\leq d} 2^{-\epsilon m/2} ,$$

which combined with (7.1) gives

$$\Pr_{S \sim P^m}[E_1] \leq 2 \binom{2m}{\leq d} 2^{-\epsilon m/2} .$$

We are left with the task to show that the assumption for m implies

$$2 \binom{2m}{\leq d} 2^{-\epsilon m/2} \leq \delta .$$

TODO □

Theorem 25 (ϵ-approximation)**.** *Let H be a class of subsets of a domain set X with* VC-dim$(H) = d$. *For any $\delta, \epsilon \in (0,1)$, and any probability distribution P over X, with probability $1 - \delta$ a random i.i.d. sample S of size*

$$m \geq \max(???,\ ???)$$

generated from P is an ϵ-approximation of P for H.

Proof. TODO □

Chapter 8

Generalization Bounds with VC-Dimension

Chapter 9

Sample Compression Schemes and Generalization Bounds

For any set A we use the notation $A^{[d]}$ to denote the class of all subsets of A of size d or less, that is,

$$A^{[d]} = \{B \subseteq A \ : \ |B| \leq d\} \ .$$

If A is finite, we denote the cardinality $A^{[d]}$ by $\binom{|A|}{\leq d}$, that is,

$$\binom{|A|}{\leq d} = \binom{|A|}{0} + \binom{|A|}{1} + \cdots + \binom{|A|}{d} \ .$$

Definition 26. A *d-size compression scheme*, over a set X, is a function

$$T : X^{[d]} \to \{0,1\}^X \ ,$$

where $\{0,1\}^X$ is the set of all functions from X to $\{0,1\}$. One may think of such function as the decoding of an encoding that represents binary functions over X as small subsets of X.

Example 27 (Intervals). For $X = \mathbb{R}$ and $d = 2$, let $T(\{x,y\})$ be the characteristic function of the closed interval spanned by $\{x,y\}$. Namely, $T(\{x,y\})(z) = 1$ iff z is between x and y.

Example 28 (Axis-aligned Rectangles). Let $X = \mathbb{R}^k$ and, for each $l \leq 2k$, let $T(\{x_1, x_2, \ldots, x_l\})$ be the minimal axis-aligned rectangle that contains these points. Namely, $T(\{x_1, x_2, \ldots, x_l\})(z) = 1$ iff, for each coordinate $j \leq k$, there are points $x_{j_1}, x_{j_2} \in \{x_1, x_2, \ldots, x_l\}$ such that the j-th coordinate of z is between the j-th coordinates of x_{j_1} and x_{j_2}.

Example 29 (Positive Half-spaces). Let, again X be the Euclidean space, $\mathbb{R}^k$ and, for any set A of k many points in $\mathbb{R}^k$, $T(A)(z) = 1$ iff z is above the linear hyper-plane spanned by these points (if the set A is not linearly independent, set $T(A)$ to be the constant 0 function).

Theorem 30. *For any d-size compression scheme T over a domain set X, every sample size $m > d$ and every $\delta > 0$,*

1. *For every probability distribution P over $X \times \{0,1\}$,*

$$\Pr_{S\sim P^m}\left[\exists A \in \mathrm{Dom}(S)^{[d]} \text{ s.t. } \left|\mathrm{Err}^P(T(A)) - \mathrm{Err}^S(T(A))\right| \geq \sqrt{\frac{\ln\binom{m}{\leq d} + \ln\left(\frac{2}{\delta}\right)}{2(m-d)}} + \frac{d}{m}\right] \leq \delta$$

where $\mathrm{Dom}(S)$ *is the set of* $x \in X$ *that appear, with a label, in the sample* S*, and* $\mathrm{Dom}(S)^{[d]} = \{A \subseteq \mathrm{Dom}(S) : |A| \leq d\}$*.*

2. *For every probability distribution P over X,*

$$\Pr_{S\sim P^m}\left[\exists A \subset S^{[d]} \text{ s.t. } |P(T(A)) - S(T(A))| \geq \sqrt{\frac{\ln\binom{m}{\leq d} + \ln\left(\frac{2}{\delta}\right)}{2(m-d)}} + \frac{d}{m}\right] \leq \delta\ ,$$

where $S^{[d]} = \{A \subseteq S \ : \ |A| \leq d\}$*,* $P(T(A))$ *is the probability of drawing a point* x *such that* $T(A)(x) = 1$*, and* $S(T(A)) = \frac{|\{x\in S \ : \ T(A)(x)=1\}|}{|S|}$*. Note that this time, since* P *is a probability distribution over* X*,* S *is a (random) subset of* X*.*

In other words, there is uniform convergence of the empirical errors of hypotheses of the form $T(A)$ to their true errors.

Proof. For any $A \in X^{[d]}$, $|A| = k \leq d$ we bound the difference of the true and the empirical error of the hypothesis $T(A)$ on an random sample S' of size $m - k \geq m - d$ using Chernoff-Hoeffding bound as

$$\Pr_{S'\sim P^{m-k}}\left[\left|\mathrm{Err}^P(T(A)) - \mathrm{Err}^{S'}(T(A))\right| \geq \epsilon\right] \leq 2e^{-2(m-k)\epsilon^2} \leq 2e^{-2(m-d)\epsilon^2}\ .$$

Since A can be arbitrary, A of size $k = |A| \leq d$ can be also generated i.i.d. randomly from P independently from S', and so,

$$\Pr_{\substack{S'\sim P^{m-k}\\ A\sim P^k}}\left[\left|\mathrm{Err}^P(T(A)) - \mathrm{Err}^{S'}(T(A))\right| \geq \epsilon\right] \leq 2e^{-2(m-d)\epsilon^2}\ .$$

The set A can be though of a part of the sample, that is, $S = (S', A)$[1], and since $\mathrm{Err}^{S'}(T(A))$ and $\mathrm{Err}^{(S',A)}(T(A))$ differ by at most k/m,

$$\Pr_{S=(S',A)\sim P^m}\left[\left|\mathrm{Err}^P(T(A)) - \mathrm{Err}^{(S',A)}(T(A))\right| \geq \epsilon + \frac{k}{m}\right] \leq 2e^{-2(m-d)\epsilon^2}\ .$$

[1]Technically, a sample S is not a set, it is rather an m-tuple belonging to X^m, that is, $S = (x_1, x_2, \ldots, x_m)$. By writing $S = (S', A)$ we mean that if $S = (x_1, x_2, \ldots, x_m)$, then $S' = (x_1, x_2, \ldots, x_{m-k})$ and $A = (x_{m-k+1}, x_{m-k+2}, \ldots, x_m)$.

By union bound over all subsets A of size at most d and using the obvious inequality $k/m \leq d/m$ we have

$$\Pr_{S\sim P^m}\left[\exists A \in \mathrm{Dom}(S)^{[d]} \text{ s.t. } \left|\mathrm{Err}^P(T(A)) - \mathrm{Err}^S(T(A))\right| \geq \epsilon + \frac{d}{m}\right] \leq 2\binom{m}{\leq d} e^{-2(m-d)\epsilon^2} .$$

Setting $\delta = 2\binom{m}{\leq d} e^{-2(m-d)\epsilon^2}$ and solving for ϵ we get

$$\epsilon = \sqrt{\frac{\ln\binom{m}{\leq d} + \ln\left(\frac{2}{\delta}\right)}{2(m-d)}} ,$$

which proves the first statement.

The second statement follows from the first one by considering P over $X \times \{0,1\}$ such that the with probability 1 the label of x is $y = 1$, that is, $P(X \times \{1\}) = 1$. Such probability distribution can be thought of as simply distribution over X. With this correspondence in mind, it can be easily seen that $P(T(A)) = \mathrm{Err}^P(T(A))$ and $S(T(A)) = \mathrm{Err}^S(T(A))$, from which the result easily follows. □

The strength of this result may be appreciated by noting that, as opposed to previous uniform convergence bounds that we have seen, here the family of potential hypotheses depends on the points that come up in the samples, and can thus be very large. To emphasize this point consider the following corollaries.

Corollary 31 (Glivenko-Cantelli, 1933). For every $\delta > 0$ and every probability distribution P over the real line. For any $m \in \mathbb{N}$, if S is an i.i.d. P-sample of size m then, with probability $\geq (1-\delta)$, for every interval $I \subseteq \mathbb{R}$,

$$\left|\frac{|S \cap I|}{m} - P(I)\right| \leq \sqrt{\frac{2\ln m + \ln(1/\delta)}{2(m-2)}} .$$

A similar result holds for estimating the probabilities of axis-aligned rectangles in $\mathbb{R}^k$ (for any k). The only difference is that when we allow I to be any such rectangle (rather than just an interval), the number 2 on the right-hand side of the inequality should be replaced by $2k$—the size of the compression scheme for k-dimensional rectangles.

Chapter 10

Online Learning

So far we have studied the statistical learning model (also called PAC model, or sometimes batch learning model), in which the main assumption was that the sample is generated from a probability distribution. We now turn to the *online learning model*, which is a different, more theoretical, learning model. However, as we will see later, there are quite a lot of connections between the online learning model and the statistical learning model.

Online learning model has the following components:

- X is the *domain set*.
- $H \subseteq 2^X$ is the *hypothesis class* (or *concept class*). An element of $h \in H$ is called a *hypothesis* (or a *concept*).
- $t \in H$ is the *target concept*, which is unknown to the learner.

Notice that the model has the same components as the PAC model without noise, except that the probability distribution P is missing. In the PAC model the probability distribution P was used to evaluate the quality of a hypothesis (the true error $\mathrm{Err}^P(h)$) and subsequently a learning algorithm. In the online learning model we evaluate the quality of an online learning algorithm by measuring how many mistakes it makes in a game against a teacher.

An online learning algorithm for a hypothesis class H, *learner*, plays a game against a malicious *teacher*. Before beginning of the game the teacher secretly chooses a target concept $t \in H$.[1] Then, the game proceeds in steps, consisting of three phases each. (See also Figure 10.1.)

1. The teacher asks a question of the form: "Does $x \in X$ belongs to t?"
   (Or equivalently, he asks: "What is the correct label of x?", where 'correct' is taken with respect the target t.)
2. The learner guesses "yes" or "no". (Or equivalently he outputs a label $y \in \{0, 1\}$ of x.)

[1] An almost equivalent defintion is that the teacher plays adaptively, that is, he does not commit himself to any particular target concept t and he answers questions so that the set of concepts consistent with his answers is non-empty. The two definitions are equivalent if the learner is deterministic.

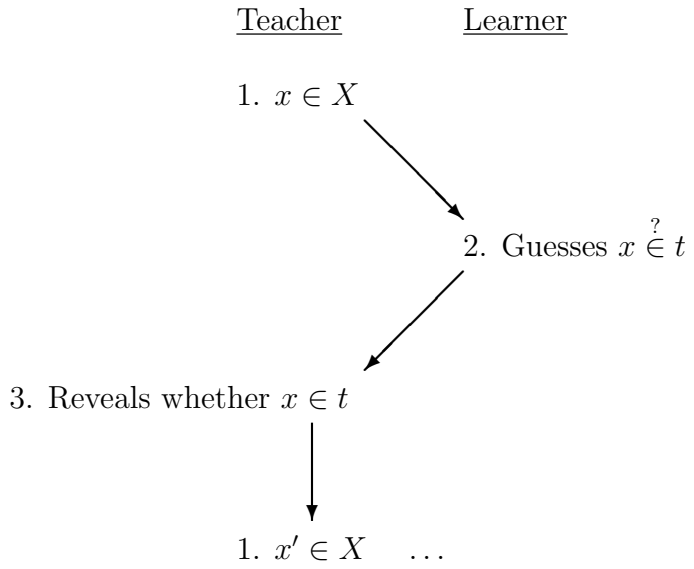

Figure 10.1: Scheme of the game between the algorithm and the teacher.

3. The teacher reveals the correct answer, whether $x \in t$ or not. (Or equivalently, he outputs $t(x)$.[2])

An incorrect reply by the learner is called a *mistake*. The number of mistakes made by the algorithm during the game measures the quality of the algorithm.

Example 32 (Thresholds). We start with a naïve learning algorithm for the concept class of "thresholds"

$$H = \{\{1\}, \{1,2\}, \ldots, \{1,2,\ldots,n\}\} = \{\{1,2,\ldots,k\} : \ 1 \leq k \leq n\} \ ,$$

over the domain $X = \{1,2,\ldots,n\}$, see Figure 10.2. The algorithm will remember the largest number ℓ labeled by the teacher with label 1, starting with label $\ell = 0$. If the algorithm is asked to label $x \leq \ell$, it outputs the label 1, and otherwise it outputs the label 0. On a mistake the algorithm updates ℓ to x.

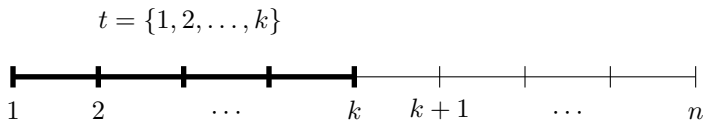

Figure 10.2: The concept class $H = \{\{1,2,\ldots,k\} : \ 1 \leq k \leq n\}$ of thresholds.

Notice that the naïve algorithm can not make more than n mistakes, because on each mistake ℓ increases by at least one, and the maximum value of ℓ is n. On the other hadn,

[2]Here we use, with little abuse, the convention that a subset $t \subseteq X$ corresponds to it characteristic function $t : X \to \{0,1\}$, where $t(x) = 1$ if $x \in t$ and $t(x) = 0$ if $t \notin t$.

the teacher can force the algorithm to make n mistakes with the target $t = \{1, 2, \ldots, n\}$ and by asking questions in the order $1, 2, \ldots, n$.

Fortunately, there is a much better algorithm resembling binary search. The learner will maintain an interval $[\ell, r]$ of target thresholds, that are consistent with the answers of the teacher. The learner starts with the largest possible interval $[1, n]$.

Suppose that at the beginning of a step the learner's interval is $[\ell, r]$ and the teacher asks for the label of x. If $x \leq \ell$, the learner outputs the label 1. Similarly, if $x > r$, the learner outputs the label 0. In none of these two cases the learner can make a mistake.

When x falls inside the interval $(\ell, r]$, the learner splits its interval into two subintervals $[\ell, x-1]$ and $[x, r]$. The learner outputs the label 1 if the first interval is shorter than the other one, and outputs the label 0 otherwise. When the teacher reveals the correct label $y \in \{0, 1\}$, the learner updates its interval: If the correct label is $y = 0$, the learner updates its interval to $[\ell, x-1]$, otherwise the learner updates his interval to $[x, r]$.

Notice that if the learner makes a mistake, the interval shrinks to half of its original length or less. When the interval has length one, i.e. $\ell = r$, the learner can not make mistakes anymore. Hence, the learner makes at most $\log_2 n = \log_2 |H|$ mistakes for arbitrary teacher and any target t.

Definition 33 (Mistake Complexity)**.** Let A be a deterministic online learning algorithm for the concept class H over the domain X. The maximum number of mistakes, denoted by $M(A)$, is called *mistake complexity* of A. Formally,

$$M(A) = \max_{\substack{t \in H \\ x_1, x_2, \cdots \in X}} (\#\text{ mistakes made on the sequence } x_1, x_2, \ldots \text{ and the target } t)$$

The mistake complexity of the concept class H is

$$M(H) = \min_A M(A) \ ,$$

that is, $M(H)$ is the mistake complexity of the best algorithm for H. For $H = \emptyset$ we define $M(H) = -1$.

10.1 Halving Algorithm

For the class H of thresholds from the example we have $M(H) \log_2 |H|$. This inequality is true in general.

Theorem 34 (Halving Algorithm)**.** *For any any finite concept class H, $M(H) \leq \log_2 |H|$.*

Proof. We construct an algorithm which makes at most $\log_2 |H|$ mistakes. This algorithm is called the *halving algorithm*. The halving algorithm will maintain a set $G \subset H$ of concept, which are consistent with the answers of the teacher. The algorithm starts with $G = H$ and the halving algorithm will arrange things so that at every mistake the set G shrinks to half or less.

If the teacher asks to label the point $x \in X$, the algorithm splits the concepts G into $G_0 = \{h \in G : h(x) = 0\}$ and $G_1 = \{h \in G : h(x) = 1\}$. It outputs the label 1 when $|G_0| \leq |G_1|$ and otherwise it outputs label 0. (In other words, the label is given according to the majority vote of the concepts in G.) When the teacher reveals the correct label $y \in \{0, 1\}$ the algorithm updates G to G_y.

When the algorithm makes a mistake, $|G|$ decreases by at least factor of two. When G consists of a single concept, the algorithm can not make mistakes anymore. Hence the upper bound on the number of mistakes is $\log_2 |H|$. □

However, the bound $M(H) \leq \log_2 |H|$ is not optimal as can be seen from the following example.

Example 35 (Singletons). Consider $X = \{1, 2, \ldots, n\}$ and the concept class of singletons, $H = \{\{x\} : 1 \leq x \leq n\}$. The mistake complexity of H is only $M(H) = 1$ as opposed to $\log_2 |H| = \log_2 n$, since there is a simple learning algorithm for H which labels any x by 0 until it makes the first and only mistake on the target t.

Note that in both Examples 32 and 35 the Vapnik-Chervnonekis dimension is 1, but the mistake complexities are different.

10.2 Optimal Algorithm

Our goal is to give complete characterization of the mistake complexity $M(H)$ of any concept class H. By doing that, we will also get an algorithm which is optimal for any class.

The *optimal algorithm* maintains, similarly as the halving algorithm, the set $G \subseteq H$ consistent with the labels of the teacher. Given a point $x \in X$ by the teacher, the algorithm computes $G_0^x = \{h \in G : h(x) = 0\}$ and $G_1^x = \{h \in G : h(x) = 1\}$, and their mistake complexities $M(G_0^x), M(G_1^x)$. Based on these two numbers the optimal algorithm, outputs a label for x. There are two options:

1. What would happen if the algorithm outputs 0? If the correct label is 0, then the maximum number of mistakes that can be done by the optimal algorithm from now on is $M(G_0^x)$. If the correct label is 1, then the algorithm makes a mistake, and then can make additional $M(G_1^x)$ mistakes. Of these two options a malicious teacher can pick larger of the two numbers, hence the maximum of number mistakes done by the algorithm would be $\max(M(G_0^x), M(G_1^x) + 1)$.

2. Likewise, if the algorithm would output label 1, the maximum of number mistakes done by the algorithm would be $\max(M(G_0^x) + 1, M(G_1^x))$.

From these two options the optimal algorithm picks the one with the minimum number of mistakes. That is, the algorithm outputs the label y such that $M(G_y^x) \geq M(G_{\overline{y}}^x)$.

10.3 Mistake Tree

A *mistake tree* for a finite concept class H over domain X is a rooted binary tree, in which every node is a non-empty subset of concepts $G \subseteq H$, and every internal node has associated a point $x \in X$. If we think of left edges as examples $(x, 0)$ and right edges as examples $(x, 1)$, then a node is a subset of all concepts consistent with the examples on the path from the root to that node. More precisely, if an internal node is a subset G and has associated a point x, then its left child (provided it exists) is the subset $G_0^x = \{h \in G \ : \ h(x) = 0\}$ and its right child (provided it exists) is the subset $G_1^x = \{h \in G \ : \ h(x) = 1\}$.

Remark 36. The number of leaves in a mistake tree for H is at most $|H|$. One can be more strict and require that there are exactly $|H|$ leaves, each corresponding to a different concept.

Note that for a concept class there exist many different non-isomorphic mistake trees. For a mistake tree, or more generally, for any rooted binary tree, we define a notion closely related to the mistake complexity.

Definition 37 (Rank of Binary Tree)**.** For a rooted binary T we define its *rank* $r(T)$ recursively as

$$r(T) = \begin{cases} -1 & \text{if } T \text{ is empty ,} \\ \max(r(T_L), r(T_R)) & \text{if } r(T_L) \neq r(T_R) \text{ ,} \\ r(T_L) + 1 & \text{if } r(T_L) = r(T_R) \text{ ,} \end{cases}$$

where T_L and T_R are respectively the left and right (possibly empty[3]) subtrees of T.

Depth (or height) of a rooted tree is the length (the number of edges) of the longest path from the root to a leaf. The depth of T will be denoted by $h(T)$.

Proposition 38. *The rank of a rooted binary tree T is equal to the depth of the largest complete binary subtree in T.*

The following proposition resembling Sauer's lemma holds.

Proposition 39. *The number of leaves in a mistake tree for concept class H is at most*

$$\sum_{i=0}^{r(T)} \binom{h(T)}{i} .$$

10.4 Littlestone's Theorem

Theorem 40 (Littlestone, 1989)**.** *For any finite concept class H, the mistake complexity $M(H)$ is equal to the maximum rank of a mistake tree for H.*

[3]If T consists of a single node, then its rank is $r(T) = 0$, since the rank of both of its subtrees is -1. If T has only one non-empty subtree, say T_L, then $r(T) = r(T_L)$.

Proof. First, we show that $M(H)$ is at least the maximum rank of a mistake tree for H. Let T be a mistake tree for H with maximum possible rank $r(T)$. Saying that $M(H) \geq r(T)$ is the same as saying that any deterministic online learning algorithm A for H can be forced to make at least $r(T)$ mistakes. An adversarial argument goes as follows. The teacher keeps track of a node G in T, the node is the set of concepts consistent with the answers of the teacher. At the beginning the G is the root of T, that is, $G = H$.

In every step, the teacher asks the learner for a point x associated with the node G. After A makes a guess of the label, the teacher gives the "correct" answer and moves to either left or right child of G accordingly. If ranks of both subtrees at G are equal, then the teacher answers so that A makes a mistake. Otherwise, it answers so that it moves to the subtree with larger rank; note that in this case the rank of the node G remains unchanged. Since, by the definition, the rank decreases by one only when both left and right subtree have equal ranks, that is, only when A makes a mistake.

Second, we show by induction on $|H|$ that $M(H)$ is at most the maximum rank $r(T)$ of a mistake tree T for H. More precisely, we show that the optimal algorithm for H makes at most $r(T)$ mistakes on any sequence of examples presented by the teacher. For $|H| = 1$ the optimal algorithm does not make any mistake on any sequence of examples, and the mistake tree for H with the maximum rank consists of a single node and has rank zero.

In the inductive case assume that $|H| > 1$, and let T be a mistake tree for H with the maximum rank. Consider any sequence of examples $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$, $x_i \in X, y_i \in \{0, 1\}$, presented by the teacher. Assume that the optimal algorithm makes k mistakes on S. Our task is to show that $m \leq r(T)$. Assume that the optimal algorithm makes the first mistake on the j-th example (x_j, y_j). Let T' be the tree for the subset of concepts $H' \subseteq H$ consistent with $S' = ((x_1, y_1), (x_2, y_2), \ldots, (x_j, y_j))$ with the maximum rank. Also, let T'' be the tree for the subset of concepts $H'' \subseteq H$ consistent with $S'' = ((x_1, y_1), (x_2, y_2), \ldots, (x_{j-1}, y_{j-1}), (x_j, \overline{y}_j))$ with maximum rank. By the definition of $M(H')$, the number of mistakes k does not exceed $M(H') + 1$, that is, $k \leq M(H') + 1$.

By the definition of the optimal algorithm $M(H') \leq M(H'')$. Note that $|H'| + |H''| \leq |H|$ since H' and H'' are disjoint and both are subsets of H. The case $H' = H$ and $H'' = \emptyset$ is impossible, since it would mean that $M(H'') = -1$ and obviously $M(H') \geq 0$ which contradicts the inequality $M(H') \leq M(H'')$. The case $H' = \emptyset$ and $H'' = H$ is impossible since there is no concepts which is consistent with the subsequence S' given by the teacher.

Hence both $|H'| < |H|$ and $|H''| < |H|$ and by the induction hypothesis $M(H') \leq r(T')$ and $M(H'') \leq r(T'')$. Consider the mistake tree T''' for $H' \cup H''$ with root $H' \cup H''$ and associated point x_j and subtrees T' and T''. We argue that $M(H') + 1 \leq r(T''')$. Indeed, if $M(H') < M(H'')$ then $r(T''')$ is at least $M(H'') \geq M(H') + 1$. The case $M(H') = M(H'')$ is more tedious, either $r(T') = r(T'')$ and thus $r(T''') = r(T') + 1 \geq M(H') + 1$, or $r(T') \neq r(T'')$ and thus $r(T''') = \max(r(T'), r(T'')) > \min(r(T'), r(T'')) \geq M(H')$.

The tree T''' can be augmented to a mistake tree $T^{(4)}$ for the whole H in such way that T''' is a subtree of $T^{(4)}$; clearly $r(T''') \leq r(T^{(4)})$. Since T is the tree with the maximum rank for H, $r(T^{(4)}) \leq r(T)$. Thus combining all the inequalities we obtain $k \leq M(H') + 1 \leq r(T''') \leq r(T^{(4)}) \leq r(T)$, and we see that the number of mistakes k is bounded by the maximum rank

$r(T)$. □

Let us define $r(H)$ as the maximum rank of a tree for H. Then the Littlestone's theorem simply says that $r(H) = M(T)$. Using $r(H)$ the optimal algorithm can be phrased more explicitly: When the algorithm's set of consistent concepts is G, the algorithm outputs the label y such that $r(G_y^x) \geq r(G_{\overline{y}}^x)$.

Chapter 11

Perceptron Algorithm

Perceptron is a simple online learning algorithm for learning the halfspaces. It is efficient in terms of time and space complexity. Also, if the positive and negative examples are "well separated", the number of mistakes the perceptron algorithm makes is small. This is precisely characterized by the Novikoff's theorem, also sometimes called perceptron convergence theorem.

Theorem 41 (Novikoff)**.** *Let $(x_1, y_1), (x_2, y_2), \ldots$ be (finite of infinite) sequence of labeled points, $x_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$, $i = 1, 2, \ldots$, such that*

(a) For all $i = 1, 2, \ldots$,

$$\|x_i\|_2 \leq R \,,$$

that is, all points x_i lie in a ball of radius R centered at the origin.

(b) There exists a unit vector $w_* \in \mathbb{R}^d$ such that for all $i = 1, 2, \ldots$

$$y_i (w^*)^T x_i \geq \gamma$$

*that is, there exists a hyperplane defined by w^* such that the positively labeled points lie on side of the hyperplane, the negative labeled points lie on the other side of the hyperplane, and the distance of any point from the hyperplane (*margin*) is at least γ.*

Then, the perceptron algorithm makes at most $(\gamma/R)^2$ mistakes on this sequence.

Proof. □

Chapter 12

Query Learning

Bibliography

[1] Martin Anthony and Peter L. Barlett. *Neural Network Learning: Theoretical Foundations*. Cambridge, 1999.

[2] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[3] Umesh V. Vazirani Michael J. Kearns. *An Introduction to Computation Learning Theory*. MIT Press, 1994.

[4] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[5] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory*, 13:145–147, 1972.

[6] John Shawe-Taylor and Nello Cristianini. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

[7] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[8] Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

[9] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[10] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[11] Vladimir N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.