Lecture 3: November 13, 2016

Lecturer: Yishay Mansour *Scribe: Yishay Mansour*

3.1 The PAC Model - Review

In the PAC Model we assume there exists a distribution $\mathcal{D}$ on the examples from $\mathcal{X}$. The learner observes a sample S composed from instances drawn i.i.d. according to $\mathcal{D}$. We assume that $\mathcal{D}$ is:

1. Fixed throughout the learning process.
2. Unknown to the learner.
3. The instances are chosen independently and identically.

The target concept is specified as a computable function c_t, and our instances are of the form $\langle x, c_t(x)\rangle$. Our goal is to find a function $h \in \mathcal{H}$ which approximates c_t with respect to $\mathcal{D}$, minimizing the error:

$$error(h) = \Pr_{\mathcal{D}}\left[c_t(x) \neq h(x)\right].$$

We would like to ensure that $error(h)$ is below a certain threshold ε, which is given as a parameter to the PAC algorithm. This parameter is a measure of the accuracy of learned hypothesis.

As a measure of our confidence in the outcome of the learning process, we add another parameter δ. We require that the following hold:

$$\Pr_{\mathcal{D}}\left[error(h) \leq \varepsilon\right] \geq 1 - \delta.$$

Namely, we would like to ensure that with high probability the output hypothesis is accurate.

The PAC algorithm has two inputs: the accuracy parameter ε and the confidence parameter δ. It also has access to instances using an oracle $EX(D, c_t)$, which generates a random example, using the distribution $\mathcal{D}$ and labeled by c_t.

Assume the realizable case, i.e., $\mathcal{C} \subseteq \mathcal{H}$. We say that an algorithm A *PAC learns* a family of concepts $\mathcal{C}$ using a hypothesis class $\mathcal{H}$ if for **any** $c_t \in \mathcal{C}$ and **any** distribution $\mathcal{D}$ on the instances in $\mathcal{X}$, A outputs a function $h \in \mathcal{H}$, such that the probability that $error(h) \leq \varepsilon$ is at least $1 - \delta$.

A PAC algorithm is *efficient* if its running time is polynomial in $\frac{1}{\epsilon}$, $\ln \frac{1}{\delta}$, the instance size and the size of the target concept c_t.

3.2 THE VC-DIMENSION

3.2.1 Motivation

We are interested in the generalization ability. How many random examples does a learning algorithm need to draw before it has sufficient information to learn an unknown target concept chosen from the hypothesis class $\mathcal{H}$? For the case of a finite concept class $\mathcal{H}$, for the realizable case, $\mathcal{C} = \mathcal{H}$,we proved that $m(\varepsilon, \delta)$ examples are sufficient for PAC learning, where,

$$m(\varepsilon, \delta) \geq \frac{1}{\epsilon} \ln \frac{|\mathcal{C}|}{\delta}.$$

We would like to be able to handle infinite concept classes, perhaps even not enumerable. We already saw an example of a threshold (last lecture) and axis-aligned rectangles (in recitation). For both concept classes we showed that the number of examples sufficient for PAC learning is $O(\frac{1}{\epsilon} \ln \frac{1}{\delta})$.

We would like to define an abstraction that will capture the complexity of a concept class $\mathcal{C}$ and show a sample bound based on it. We will introduce the definition of VC-dimension named after Vapnik and Chervonenkis, and show the connection between the VC-dimension and PAC learning. The VC-dimension, will conceptually replace $\ln |\mathcal{C}|$ for infinite concept classes.

3.2.2 Definitions

We start with a few definitions. Assume $\mathcal{C}$ is a concept class defined over the instance space $\mathcal{X}$. Let $c \in \mathcal{C}$ be identified with a set $c \subseteq \mathcal{X}$ such that $c = \{x \in \mathcal{X} \mid c(x) = 1\}$.

Definition For a concept class $\mathcal{C}$ over $\mathcal{X}$ and for any $S \subseteq X$, we define the projection of $\mathcal{C}$ on S as follows:

$$\Pi_{\mathcal{C}}(S) = \{c \cap S | c \in \mathcal{C}\}$$

Equivalently, if $S = \{x_1, \ldots, x_m\}$ then we can think of $\Pi_{\mathcal{C}}(S)$ as the set of vectors $\Pi_{\mathcal{C}}(S) \subseteq \{0,1\}^m$ defined by $\Pi_{\mathcal{C}}(S) = \{\langle c(x_1), \ldots, c(x_m)\rangle : c \in \mathcal{C}\}$.

This is the projection of the concept class $\mathcal{C}$ on a finite subset $S \subset \mathcal{X}$, namely $\Pi_{\mathcal{C}}(S)$ is all the possible functions that $\mathcal{C}$ induces on S. We are interested in how many different functions $\mathcal{C}$ induces on S. In effect we are reducing the concept class $\mathcal{C}$ to the concept class $\mathcal{C}_{|S}$, where $S = \{x_1, ..., x_m\}$. The concept class $\mathcal{C}_{|S}$ is finite with at most 2^m different concepts, thus $|\Pi_{\mathcal{C}}(S)| \leq 2^m$.

Next we define the richness of the projection. We will be interested only whether all possible functions on S are realizable, which we will call shattering.

Definition A concept class $\mathcal{C}$ *shatters* S if $2^{|S|} = |\Pi_{\mathcal{C}}(S)|$. In other words a class $\mathcal{C}$ shatters a set of inputs S if any possible Boolean function on S can be represented by some $c \in \mathcal{C}$.

Now we are ready to define the notion of VC-dimension.

Definition *VCdim (Vapnik-Chervonenkis dimension)* of $\mathcal{C}$ is the maximum size of a set S shattered by $\mathcal{C}$:

$$VCdim(\mathcal{C}) = max\{d : \exists S : |S| = d \ \ and \ \ |\Pi_{\mathcal{C}}(S)| = 2^d\}.$$

If a maximum value does not exist, i.e., for every d there is a set S_d of size d which $\mathcal{C}$ shatters, then $VCdim(\mathcal{C}) = \infty$.

3.2.3 Some examples of geometric concepts

Let us consider a few examples of simple concept classes and calculate their VC dimension. In order to show that the VC dimension of a class is at least d, we need to exhibit some shattered set S of size d. In order to show that the VC dimension is at most d, we need to show that no set S of size $d+1$ can be shattered.

$\mathcal{C}_1$: Threshold

Let $\mathcal{X} = [0,1]$ and concepts are c_α for $\alpha \in [0,1]$, where:

$$c_\alpha(x) = \begin{cases} 0 & x < \alpha \\ 1 & x \geq \alpha \end{cases}$$

Note that although the number of concepts is uncountable the concept class is learnable. We show that $VCdim(\mathcal{C}_1) = 1$.

First, we will show that $VCdim(\mathcal{C}_1) \geq 1$. Let $x = \frac{1}{2}$. We need to show two concepts such that $|\Pi_{\mathcal{C}}(\{\frac{1}{2}\})| = 2$. For example, for $c_{2/3}(1/2) = 0$ and $c_{1/3}(1/2) = 1$. Thus the VC-dimension of $\mathcal{C}_1$ is at least 1.

Second, we will show that $VCdim(\mathcal{C}_1) < 2$, by showing that for any set of size two, there exists an assignment which is not in the concept class. If $S = \{z_1, z_2\}$ where $z_1 < z_2$, the assignment that lets z_1 be 1 and z_2 be 0, is impossible. Thus, $VCdim(\mathcal{C}_1) < 2$, and we derive that $VCdim(\mathcal{C}_1) = 1$.

$\mathcal{C}_2$: A finite union of intervals

The domain is $\mathcal{X} = [0,1]$. Each concept $c_I \in \mathcal{C}_2$ is parameterize by a set of intervals $I = ([a_1, b_1], \ldots [a_k, b_k])$ concept. Let $INT(I) = \cup_{[a_i,b_i] \in I}[a_i, b_i]$ be the union of the intervals. The value of $c_I(x) = 1$ iff $x \in INT(I)$.

For any finite set of points S and any assignment that makes a subset $S_+ \subseteq S$ positive, we can cover the positive points by choosing the intervals small enough. The number of intervals is finite (at most $|S|$). Thus, $VCdim(\mathcal{C}_2) = \infty$.

$\mathcal{C}_3$: Convex polygons in the plane

The domain is $\mathcal{X} = [-1, 1]^2$. We consider concepts which are parameterized by a set of points $CP = (z_1, \ldots z_k)$ on the plane. The concept $c_{CP}(x) = 1$ iff x is in the convex hull of CP. Namely, points inside the convex polygon CP are positive and outside are negative. Again, we have no bound on the number of edges, but it has to be finite. We want to show $VCdim(\mathcal{C}_3) = \infty$, i.e., for every d there is a set S of size d that can be shattered by convex polygons.

Let S be a set of d points on the circle perimeter. For any labeling of the points in S, let S_+ be the positive point. Consider c_{S_+}. Note that c_{S_+} is positive on S_+, the vertices of the polygon, and negative on any other point on the circle, specifically $S - S'$. Thus, for any d points on the unit circle, all the 2^d classifications are possible. Therefore, $VCdim(\mathcal{C}_3) = \infty$.

C_4 - Hyperplane

Let the domain be $\mathbb{R}^d$. a hyperplane is parameterized by a set of weight $w \in \mathbb{R}^d$ and a threshold θ. It defines a concept $c_w(x)$, where

$$c_{w,\theta}(x) = sign(\sum_{i=1}^{n} w_i x_i + \theta)$$

We will prove the following bound.

Theorem 3.1

$$VCdim(\mathcal{C}_4) = d + 1$$

First we will show that there are at least $d + 1$ points that can be shattered by $\mathcal{C}_4$.

Claim 3.2

$$VCdim(\mathcal{C}_4) \geq d + 1$$

Proof: Consider the set $S = \{\vec{0}, \vec{e}_1, \ldots, \vec{e}_n\}$ of size $d + 1$, where $\vec{e}_i$ is a unit vector which has 1 in coordinate i and 0 in all the others. We will show that $\mathcal{C}_4$ shatters S. Consider an assignment of labels to S such that $S_+ \subseteq S$ is the set of positive labels. We set the weights $w_i = +1$ if $\vec{e}_i \in S_+$ and otherwise $w_i = -1$. We set the threshold $\theta = +1/2$ if $\vec{0} \in S_+$ and $\theta = -1/2$ otherwise.

For any point $\vec{e}_j$ we have that $c_{w,\theta}(\vec{e}_j) = sign(w_j + \theta)$. Since by design $w_j + \theta > 0$ iff $c_{w,\theta}(\vec{e}_j) = 1$, all those points are labeled correctly. In addition, $c_{w,\theta}(\vec{0}) = sign(\theta)$, and again, $\theta > 0$ iff $c_{w,\theta}(\vec{0}) = 1$. This implies that $c_{w,\theta}$ that we defines the correct labeling according to S_+. □

We showed that there exists a set of size $d+1$ which $\mathcal{C}_4$ shatters, hence $VCdim(\mathcal{C}_4) \geq d+1$. We will now show that $VCdim(\mathcal{C}_4) = d+1$. We start with some general definitions and Radon theorem will be shown.

Definition A subset A is *convex* if $\forall x_1, x_2 \in A$ the line connecting x_1 to x_2 is in A. Formally:

$$\forall \lambda \in [0,1]. \lambda x_1 + (1-\lambda) x_2 \in A$$

The *convex hull* of S is the smallest convex set which contains all the points of S. We denote it as $conv(S)$.

We are now ready to state Radon Theorem, which will be used in the proof of the VC-dimension.

Theorem 3.3 *(Radon Theorem) Let S be a set of $d+2$ points in $\mathbb{R}^d$. There is a non empty subset S' of S such that*

$$conv(S') \cap conv(S \setminus S') \neq \phi$$

Namely, for any set S of $d+2$ points, there is a partition to S' and $S - S'$ such that the convex hull of the two sets intersect.

Proof: Let:

$$S = \{x_0, ..., x_{d+1}\}$$

where $x_i \in \mathbb{R}^d$. Since S contains $d+2$ vectors, we can solve for the following $d+1$ equations and find $(\alpha_0, ..., \alpha_{d+1}) \neq \vec{0}$, such that,

$$\sum_{i=0}^{d+1} \alpha_i x_i = 0,$$

and

$$\sum_{i=0}^{d+1} \alpha_i = 0.$$

Thus, we have a set of $d+1$ linear equations over $d+2$ variables $\{\alpha_i\}_{i=0}^{d+1}$. There exist a non-zero vector $\langle \alpha_0, ..., \alpha_{d+1} \rangle$ satisfying the above equations, because every $d+1$ points (vectors) are linear dependent.

Assume that $\alpha_0, ..., \alpha_p$ are positive, and $\alpha_{p+1}, ..., \alpha_{d+1}$ are negative (zeros can go in to either group). We define:

- $\alpha = \sum_{i=0}^{p} \alpha_i > 0$
- $\beta_i = \frac{\alpha_i}{\alpha} > 0 \qquad 0 \leq i \leq p$

- $\gamma_i = \frac{-\alpha_i}{\alpha} > 0 \qquad p+1 \leq i \leq d+1$

We have that,

$$\sum_{i=0}^{d+1} \alpha_i x_i = 0 \;\Rightarrow\; \sum_{i=0}^{p} \beta_i x_i = \sum_{i=p+1}^{d+1} \gamma_i x_i$$

Notice that $\sum_{i=0}^{p} \beta_i = \sum_{i=p+1}^{d+1} \gamma_i = 1$.
By definition of convexity,

$$\sum_{i=0}^{p} \beta_i x_i \in conv(x_0, ..., x_p),$$

and

$$\sum_{i=p+1}^{d+1} \gamma_i x_i \in conv(x_{p+1}, ..., x_{d+1}).$$

Hence, there is a point that belongs to the intersection of

$$conv(x_0, ..., x_p) \cap conv(x_{p+1}, ..., x_{d+1}) \neq \emptyset \; .$$

□

Based on Radon Theorem we can prove our VC-dimension bound.

Claim 3.4

$$VCdim(\mathcal{C}_8) < d + 2$$

Proof: Proof by contradiction. Assume $S = \{x_1, ..., x_{n+2}\}$ can be shattered. By Radon Theorem there is a non-empty subset S' of S such that $conv(S) \cap conv(S - S') \neq \phi$.

Assume we have hyperplane $c_{w,\theta}$ which classifies S' as 1 and $S - S'$ as 0. Let P be the set of points classified as 1 by $c_{w,\theta}$ and N the remaining points (classified as 0). Both sets P and N are convex.

This implies that $conv(S') \subset P$ and $conv(S - S') \subset N$. However, by Radon Theorem there is a point $x \in conv(S') \cap conv(S - S')$. This implies that $x \in P \cap N$, but we know that $P \cap N = \emptyset$, which is a contradiction to the assumption that such a set S and a hyperplane $c_{w,\theta}$ exists. □

Combining Claim 3.2 and Claim 3.4, we derive Theorem 3.1 and conclude that $VCdim(\mathcal{C}_4) = d + 1$.

3.2.4 Sample Size Lower Bounds

We would like to show that if a concept class has a finite VC-dimension d, then there is a function $m(\epsilon, \delta, d)$ dependent on ϵ, δ and d, which characterizes the required sample. We

start by showing an impossibility result, showing that if the sample is too small, then any PAC learning algorithm would fail.

Theorem 3.5 *If a concept class $\mathcal{C}$ has a VC-dimension $d+1$, for any $\delta < 1/2$, we have,*

$$m(\epsilon, \delta = 1/2, d+1) \geq \frac{d}{16\epsilon} = \Omega\left(\frac{d}{\epsilon}\right) .$$

Proof: For contradiction assume this is possible. Let $T = \{z_0, z_1, ..., z_d\}$ such that $\mathcal{C}$ shatters T (such a set exists, since $VCdim(\mathcal{C}) = d+1$). Next, we construct a distribution $\mathcal{D}$ in the following manner:

$$\mathcal{D}(x) = \begin{cases} 1 - 8\epsilon & x = z_0 \\ \frac{8\epsilon}{d} & x = z_i, 1 \leq i \leq d \\ 0 & otherwise \end{cases}$$

Finally, select c_t randomly as follows,

$$c_t(x) = \begin{cases} 1 & x = z_0; \\ 0 \text{ or } 1 \text{ (with probability } \frac{1}{2}) & x = z_i; \end{cases}$$

Note that there exists a $c_t \in \mathcal{C}$ that agrees on T, since $\mathcal{C}$ shatters T. We claim that if we sample less than $\frac{d}{2}$ points out of $\{z_1, \ldots, z_d\}$ then the error is at least 2ε. Let $RARE$ be the set of points $\{z_1, ..., z_d\}$, and $UNSEEN \subseteq RARE$ be the points in $RARE$ which have not been samples. Then:

$$\Pr[ERROR] \geq \frac{1}{2} \cdot |UNSEEN| \cdot \frac{8\epsilon}{d} = \frac{4\varepsilon}{d} \cdot |UNSEEN|$$

The expected number of samples in $RARE$ is $m \cdot 8\varepsilon < \frac{d}{16\varepsilon} \cdot 8\varepsilon = \frac{d}{2}$. With probability of at least $\frac{1}{2}$ we will sample at most $\frac{d}{2}$ points (recall that the expected value for the binomial distribution equals its median). In such a case, $|UNSEEN| \geq \frac{d}{2}$. This implies that with probability at least $\frac{1}{2}$ we have error at least 2ε, a contradiction. □

3.3 Sample Size Upper Bound

We will now turn to the important application of the VC-dimension - deriving an upper bound on the sample size. First we will show a wrong proof for an upper bound: If S is sampled than the number of hypotheses is $\mathcal{C}_{|S} = \Pi_{\mathcal{C}}(S)$. Since $\mathcal{C}_{|S}$ is finite we can set

$$m \geq \frac{1}{\epsilon} log \frac{|\Pi_{\mathcal{C}}(S)|}{\delta}$$

This proof is obviously wrong (why?). We will now "fix" the proof:

Definition 3.6 *Given a target concept c_t, the set of ϵ-bad concepts includes all the concepts that have an error larger than ϵ. Formally:*

$$B_\epsilon(c_t) = \{h \in H | error(h, c_t) > \epsilon\}.$$

We will show that if we have a large enough sample, then none of these concepts will be consistent with the samples.

Definition 3.7 *A set of points S, is an ϵ-hitting set for c_t with respect to a distribution $\mathcal{D}$, if for each concept $h \in B_\epsilon(c_t)$ there exists a point $x \in S$ such that $h(x) \neq c_t(x)$.*

The important property of ϵ-hitting sets is that if the sample S forms an ϵ-hitting set for c_t, and the hypothesis $h \in \mathcal{C}$ is consistent with S, then this hypothesis h must have error at most ϵ. Thus, if we can bound the probability that the random sample S fails to form an ϵ-hitting set for c_t, then we have bounded the probability that a hypothesis consistent with S has error greater than ϵ.

For this discussion we will use a sample set S that is made up of two sample sets, S_1 and S_2, each is of size m, and each sampled independently according to $\mathcal{D}$. We define A as the event that S_1 is not an ϵ-hitting set. We want to bound the probability of the event A, as it bounds the probability of failure.

Assume event A holds. Then, there are concepts in $B_\epsilon(c_t)$ which are consistent with S_1. Let h be an ϵ-bad hypothesis consistent with S_1. Now, we look at the additional sample S_2 of m points. The expectation of the error of h is at least ϵ, thus, since the median of a Binomial distribution is its expectation, with a probability at least $\frac{1}{2}$, h will have at least ϵm errors on S_2.

Define B as the event that there exists a function $h \in B_\epsilon(c_t)$ such that h is consistent with S_1 and has ϵm errors on S_2. From our discussion before, $Pr[B|A] \geq \frac{1}{2}$, and since event B implies event A, we have $Pr[B] = Pr[B|A]Pr[A]$. Combining the two we have,

$$2 \cdot Pr[B] \geq Pr[A].$$

This implies that a bound on the probability of B will imply a bound on the probability of A. The main advantage is that the event B is defined on the finite set of points $S_1 \cup S_2$.

Define F as the projection of $\mathcal{C}$ to $S_1 \cup S_2$. Formally:

$$F = \Pi_{\mathcal{C}}(S_1 \cup S_2).$$

Later we will bound the size of F.

We will define the set of errors of h as follows:

$$ER(h) = \{x : x \in S_1 \cup S_2 \ \ and \ \ c_t(x) \neq h(x)\}.$$

We assumed that $ER(h)$ has at least ϵm elements because in S_2 there are at least ϵm elements from $ER(h)$. That is, $|ER(h)| \geq \epsilon m$. Now, for $h \in B_\epsilon(c_t)$, the events can be formulated as follows:

Event A**:** $ER(h) \cap S_1 = \emptyset$, and

Event B**:** $ER(h) \cap S_1 = \emptyset \wedge |ER(h) \cap S_2| = |ER(h)| = \epsilon m$.

We analyze the probability that h is consistent with S_1 and has at least ϵm errors on S_2. Since we chose both S_1 and S_2 from the distribution D, we can build the distribution on S_1 and S_2 as follows: We sample $2m$ points $S_1 \cup S_2$ and divide the sample randomly, between S_1 and S_2. This is exactly the same distribution, because any ordering of the $2m$ elements, partitioned into two sets randomly, is the same as sampling S_1 and then S_2 (due to the i.i.d property).

Our problem is now reduced to the following simple combinatorial experiment: we have $2m$ balls (the set $S = S_1 \cup S_2$), each colored black or white, with exactly $l \geq \epsilon m$ black balls (these are the points of S that h errors on them). We partition these balls randomly into two sets of equal sizes S_1 and S_2, and we are interested in bounding the probability that all the black balls fall in S_2.

We now calculate the number of possible partitions. The number of ways we can choose l elements from $2m$ elements is:

$$\binom{2m}{l}.$$

Among them, the number of partitions in which all the black balls fall into S_2 is:

$$\binom{m}{l}.$$

Thus, the probability that all of the black balls are in S_2 is exactly

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

We can now bound the probability of event and B, by doing a union bound over the assignments $f \in F$:

$$\Pr[B] \leq |F| \cdot 2^{-\epsilon m},$$

which implies that,

$$Pr[A] \leq 2Pr[B] \leq 2|F| \cdot 2^{-\epsilon m}.$$

Thus, in order for our confidence level (δ) to satisfy our goal, we will require:

$$2|F|2^{-\epsilon m} \leq \delta,$$

and we get that the sample size should be:

$$m = O(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{1}{\epsilon}\log|F|).$$

The only issue we still have to resolve is a bound on the size of F. As we recall, F is a projection of $\mathcal{C}$ on a set with $2m$ elements. We will show that:

$$|F| = |\Pi_{\mathcal{C}}(S_1 \cup S_2)| \leq (2m)^d.$$

Lemma 3.8 (Sauer-Shelah Lemma) *Let $VCdim(\mathcal{C}) = d$ and $|S| = m$, then*

$$\Pi_{\mathcal{C}}(S) \leq \sum_{i=0}^{d} \left(\begin{array}{c} m \\ i \end{array} \right).$$

Proof: We will define a recursion which we will later prove that it bounds the number of concept in the projection. We define the function J as follows:

$$J(m,d) = J(m-1,d) + J(m-1,d-1),$$

with the initial conditions: $J(m,0) = 1$, and $J(0,d) = 1$. Solving this recursion (not detailed here) gives:

$$J(m,d) = \sum_{i=0}^{d} \left(\begin{array}{c} m \\ i \end{array} \right).$$

We will use this function to bound $\Pi_{\mathcal{C}}(S)$. The proof is by induction on both d and m. For the base of the induction, the claim is easily established when $d = 0$ and m is arbitrary, and when $m = 0$ and d is arbitrary.

We assume for induction that for all m', d' such that $d' + m' < d + m$, we have $\Pi_{\mathcal{C}}(S) \leq J(m,d)$. We now show that this inductive hypothesis holds for d and m. Let $S = \{x_1, \ldots, x_m\}$ be a set of m different points and let $\mathcal{C}_{|S}$ be the projection of the concept class $\mathcal{C}$ on S. Namely,

$$\Pi_{\mathcal{C}}(S) = \mathcal{C}_{|S} = \{c \cap S | c \in \mathcal{C}\}.$$

We will show that for every S we have $|\Pi_{\mathcal{C}}(S)| \leq J(m,d)$.

We define a new set T which is the set S after extracting the last point, i.e.,

$$T = \{x_1, \ldots, x_{m-1}\} = S - \{x_m\} \quad , \quad |T| = m - 1$$

Define C_* as all the assignments over T which can be completed either by $x_m = 0$ or by $x_m = 1$. Then $|C_*|$ counts the number of pairs of sets in $\Pi_{\mathcal{C}}(S)$ that are collapsed to a single representative in $\mathcal{C}_T = \Pi_{\mathcal{C}}(S - \{x_m\})$. We thus have:

$$|\mathcal{C}_{|S}| = |\mathcal{C}_T| + |C_*|.$$

Trivially, every concept in $\mathcal{C}_{|S}$ appears in $\mathcal{C}_{|T}$, and if it appears twice it is also counted in C_*.

We now bound $\mathcal{C}_{|T}$ and C_* separately. The bound for $\mathcal{C}_{|T}$, from the induction hypothesis, is $|\mathcal{C}_{|T}| \leq J(m-1,d)$. We claim that the bound for C_* is,

$$|C_*| \leq J(m-1,d-1).$$

Note that if C_* shatters a set $\{x_1,...,x_i\}$ then $\mathcal{C}$ shatters the set $\{x_1,...,x_i,x_m\}$, since each function can be completed in two different ways. By definition of C_*, for every assignment of $x_1,...,x_i$ there exist a pair of concepts: $c_0, c_1 \in \mathcal{C}$ that are consistent with $c_1(x_m)=1$ and $c_0(x_m)=0$. Hence, if C_* shatters a set of size i, then C shatters a set of size $i+1$. Since we assume $VCdim(C)=d$, then $VCdim(C_*) \leq d-1$. Hence, from the induction hypothesis: $|C_*| \leq J(m-1,d-1)$, and

$$|\mathcal{C}_{|S}| = |\mathcal{C}_{|T}| + |C_*| \leq J(m-1,d) + J(m-1,d-1) = J(m,d),$$

which concludes the proof of the Claim. □

By Lemma 3.8 we have that

$$|F| \leq \sum_{i=0}^{d} \left(\begin{array}{c} 2m \\ i \end{array} \right).$$

This function has the following behaviors:

$$|F| \leq \sum_{i=0}^{d} \left(\begin{array}{c} 2m \\ i \end{array} \right) = \left\{ \begin{array}{ll} 2^{2m} & d \geq 2m \\ 2(2m)^d & d < 2m. \end{array} \right.$$

That is, the function grows exponentially in $2m$ until $2m$ reaches d and then it grows polynomially in m and exponentially in d. From that we can conclude that the number of functions in the projection can either grow as 2^m or fall to $2m^d$. No intermediate behavior exists. We can now return abound bound the required sample size m:

$$\begin{aligned} & m = O(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{1}{\epsilon}\log(2m)^d) \\ \Rightarrow \quad & m = O(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d}{\epsilon}\log m) \\ \Rightarrow \quad & m = O(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d}{\epsilon}\log\frac{d}{\epsilon}). \end{aligned}$$

That is, this is the number of samples required to guarantee that in probability $1-\delta$ our error will be smaller than ϵ. As we can see above, we can actually bound the size of

the sample with a function of the $VCdim$ alone. It is also worth noting that the difference between the lower bound and the upper bound we found are relatively small.

We can derive a similar upper bound for the non-realizable case for which we will have

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

3.4 Rademacher Complexity

3.4.1 Motivation

We would like to derive tighter bounds for the sample size. Using the VC dimension our bounds depend on $|\Pi_C(m)|$. First, how can we estimate $|\Pi_C(m)|$? We can choose a random assignment for S and see if there is a $h \in \mathcal{H}$ that corresponds to it. This way, we estimate $|\Pi_C(m)|/2^m$. The problem with this method is that the probability that we "hit" a legal assignment is exponentially small.

An alternative approach is taking a random assignment and estimating our overfit on it. That is, we see how well an $h \in \mathcal{H}$ estimates it. For an arbitrary fixed h, the expect error is 50%. The best $h \in \mathcal{H}$ should give us a lower error, say 40%, which would intuitively mean that it overfits by 10%.

3.4.2 Problems with VC-dim bounds

The lack of tightness of the VC-dim bounds stems from two weaknesses in our analysis can be trace back to two different issues.

1. The two sample trick. We use union bound for the samples S_1 and S_2. (This seems to be a minor issue.)

2. Ignoring the actual distribution, and concentrating on the worst S for our analysis. (This seems a major issue.)

We would like to overcome these problems. So far we showed that

$$error_D(h) \leq error_S(h) + O\left(\sqrt{\frac{\ln|\Pi_C(2m)|}{m}}\right).$$

We would like to be give a tighter bound (i.e., replace the rightmost part of the inequality by something smaller). For this we are willing to sacrifice that the result will hold for all distribution, but rather have a distribution dependent result.

3.4.3 Rademacher Averages

We now define *Rademacher averages*. First we define it for a given sample S and then for a specific distribution $\mathcal{D}$.

Definition 3.9 *Given a sample $S = \{x_1, x_2, ..., x_m\}$ and a set of functions $\mathcal{H}$, the* empirical Rademacher complexity is

$$R_S(\mathcal{H}) = E_\sigma[\max_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)],$$

where σ is the random labeling. That is, $\sigma_i \in \{\pm 1\}$, $Pr[\sigma_i = +1] = 1/2$.

Note that $h(x_i) \in \{\pm 1\}$ as well, and therefore if $\sigma_i h(x_i) = 1$ there is no error ($\sigma_i = h(x_i)$), and if $\sigma_i h(x_i) = -1$ there is an error. When we choose an h that maximizes the sum, we are maximizing the precision (minimizing the number of errors). In other words, we take a random labeling and see how well the "best" h fits it. What we are in fact approximating is, given a random labeling, how good is our hypothesis class.

Definition 3.10 *We define the* Rademacher complexity over a distribution $\mathcal{D}$ to be:

$$R_{\mathcal{D}}(\mathcal{H}) = E_{S\sim\mathcal{D}}[R_S(\mathcal{H})].$$

Note that $R_{\mathcal{D}}(\mathcal{H})$ is the expected value of $R_S(\mathcal{H})$ over a random sample. The complexity depends upon the distribution $\mathcal{D}$ and not on the worst sample. The intuition behind the Radamcher average is this: we have a random labeling σ. What is the best correlation we can get? (Notice that we always get a number between 0 and 1, so $R_{\mathcal{D}}(\mathcal{H}) = 1$ means perfect correlation). What is the relationship to VC-dim? If $\mathcal{C}$ shatters S, what is $R_S(\mathcal{C})$? If $\mathcal{C}$ shatters S, each labeling has a $c \in \mathcal{C}$ that fits it. Therefore, $R_S(\mathcal{C}) = 1$ (for any σ, there is a $c \in \mathcal{C}$ such that $\sigma_i = h(x_i)$, therefore the expected value is 1).

Theorem 3.11 *With probability $1 - \delta$, $\forall h \in \mathcal{H}$*

$$\begin{aligned} error_D(h) &\leq error_S(h) + R_D(\mathcal{H}) + \sqrt{\frac{\ln(2/\delta)}{2m}} \\ &\leq error_S(h) + R_S(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}, \end{aligned}$$

where $|S| = 2m$.

We start with a simple theorem bout the expected overfit as a function of the Rademacher complexity, and latter we give a proof of the main theorem.

Theorem 3.12

$$E_{S\sim\mathcal{D}}\left[\max_{h\in\mathcal{H}} E_{\mathcal{D}}[f] - \frac{1}{m}\sum_{i=1}^{m} f(x_i)\right] \leq 2R_{\mathcal{D}}(\mathcal{H})$$

Proof: Pick a second sample $S' = \{x'_1, \ldots, x'_m\}$.

$$\begin{aligned}
E_{S\sim\mathcal{D}}\left[\max_{h\in\mathcal{H}} E_{\mathcal{D}}[f] - \frac{1}{m}\sum_{i=1}^{m} f(x_i)\right] &= E_{S\sim\mathcal{D}}\left[\max_{h\in\mathcal{H}} E_{S'\sim\mathcal{D}}[\frac{1}{m}\sum_{i=1}^{m} f(x'_i)] - \frac{1}{m}\sum_{i=1}^{m} f(x_i)\right] \\
&\leq E_{S,S'\sim\mathcal{D}}\left[\max_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} f(x'_i) - \frac{1}{m}\sum_{i=1}^{m} f(x_i)\right] \\
&= E_{S,S'\sim\mathcal{D},\sigma_i}\left[\max_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i(f(x'_i) - f(x_i))\right] \\
&\leq E_{S'\sim\mathcal{D},\sigma_i}\left[\max_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(x'_i)\right] + E_{S\sim\mathcal{D},\sigma_i}\left[\max_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(x_i)\right] \\
&= 2R_{\mathcal{D}}(\mathcal{H})
\end{aligned}$$

□

Before we prove the theorem, we define McDiarmid's inequality, which is a generalization of the Chernoff and Hoeffding bounds.

Theorem 3.13 (McDiarmid's inequality) *Let $X_1, ..., X_m$ be independent random variables, and let $\Phi(X_1, ..., X_m)$ be a real function such that for any realization we have*

$$\forall i, |\Phi(x_1, ..., x_i, ..., x_m) - \Phi(x_1, ..., x'_i, ..., x_m)| \leq c_i,$$

then $\forall \epsilon > 0$

$$\Pr[\Phi(X) > E[\Phi(X)] + \epsilon] \leq exp(-2\epsilon^2 / \sum_{i=1}^{m} c_i^2).$$

In other words, we would like to say that if none of the random variables has too much influence on Φ (each random variable X_i's influence is less than or equal to c_i), then the probability that a random value of $\Phi(X)$ will be far from its expected value is small.

Corollary 3.14 *If $\forall i$, $0 \leq x_i \leq 1$ and $\Phi(X) = \frac{1}{m}\sum_{i=1}^{m} x_i$, then each x_i's influence is $\leq 1/m$ and we get*

$$exp(-2\epsilon^2 / \sum_{i=1}^{m} c_i^2) = exp(-2\epsilon^2 / \frac{m}{m^2}) = e^{-2\epsilon^2 m},$$

which is the Chernoff bound.

Corollary 3.15 *For* $X_i \in [a_i, b_i]$, $\Phi(X) = \frac{1}{m}\sum_{i=1}^{m} x_i$, *we get* $c_i = \frac{b_i - a_i}{m}$ *and*

$$Pr[\Phi(X) > E[\Phi(X)] + \epsilon] \leq exp(-2\epsilon^2 m^2 / \sum_{i=1}^{m}[b_i - a_i]^2),$$

which is Hoeffding's inequality.

3.4.4 Proof of the Rademacher Complexity (Theorem 3.11

Proof: We first bound the difference between the expectation and the realization in the error. Let,

$$MAXGAP(S) = \max_{h \in \mathcal{H}}[error_D(h) - error_S(h)].$$

We would like to ensure that with high probability, $1 - \delta$, we have that the difference is small, i.e., $MAXGAP(S) \leq \epsilon$.

Since the error function is an averaging function, and $|S| = m$, the dependency of $MAXGAP$ on each $x_i \in S$ is at most $1/m$. Therefore we can apply McDiarmid's inequality to $MAXGAP$. Taking $\delta = 2e^{-2\epsilon^2 m}$, we get that with probability $1 - \delta/2$,

$$MAXGAP(S) \leq E_S[MAXGAP(S)] + \sqrt{\frac{ln(2/\delta)}{m}}.$$

If we now show that $R_D(\mathcal{H}) \geq E_S(MAXGAP(S))$, this will complete the proof of Theorem 3.11.

To show that $R_D(\mathcal{H}) \geq E_S(MAXGAP(S))$, we add a second sample S' of size m, where S' is i.i.d. and independent of S. Then,

$$error_D(h) = E_{S'}[error_{S'}(h)].$$

Because S is independent of S', we can look at S's error as an average of S'. Let $S = \{x_1, ..., x_m\}$, and $S' = \{x'_1, ..., x'_m\}$.

$$\begin{aligned} E_S[MAXGAP(S)] &= E_S[\max_{h \in \mathcal{H}} E_{S'}[error_{S'}(h) - error_S(h)]] \\ &\leq E_{S,S'}[\max_{h \in \mathcal{H}}[error_{S'}(h) - error_S(h)]] \\ &= E_{S,S'}[\max_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} error_{x'_1}(h) - error_{x_i}(h)], \end{aligned}$$

where $error_x(h)$ is whether h makes a mistake on x. If h makes a mistake, $error_x(h) = 1$, otherwise, $error_x(h) = 0$.

As before, we first sample $S \cup S'$ and only then split the sample into S and S'. We do this by arbitrarily splitting the sample into m pairs, x_i and x'_i and and then for each pair, deciding which one is in S and which one is in S'. Sampling both sets and then arbitrarily splitting them up yields the same distribution as independently sampling two sets. The value of σ decides which set gets which samples. That is, for each i, if $\sigma_i = -1$, then $x_i \in S$ and $x'_i \in S'$ and if $\sigma_i = 1$, then $x_i \in S'$ and $x'_i \in S$. Continuing with our analysis:

$$E_S[MAXGAP(S)] \leq E_{S',\sigma}[\max_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i error_{x'_i}(h)] - E_{S,\sigma}[\min_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i error_{x_i}(h)]$$

$$= \frac{2}{m} E_{S,\sigma}[\max_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i error_{x_i}(h)]. \tag{3.1}$$

We are almost done with our proof. We notice now that in the Rademacher complexity we have $\sum_{i=1}^{m} \sigma_i h(x_i)$. This can be written as $\rangle\sigma, h\langle$. In our last inequality we have

$$< \sigma, error(h) > = \sum_{i=1}^{m} \sigma_i error_{x_i}(h) = \sum_{i=1}^{m} \sigma_i I(h(x_i) \neq c_t(xi)).$$

If in the Rademacher complexity we write $\langle \sigma \cdot c_t, h\rangle$ instead of $\langle \sigma, h\rangle$, we get the same distribution, because σ creates all the vectors, and if we multiply each coordinate by $c_t(x_i)$, it does not affect the distribution.

$$\langle \sigma \cdot c_t, h\rangle = \sum (\sigma_i \cdot c_t(x_i)) \cdot h(x_i) = \sum \sigma_i \cdot (c_t(x_i) \cdot h(x_i)) = \langle \sigma, c_t \cdot h\rangle.$$

And

$$c_t(x_i) \cdot h(x_i) = 1 - 2error_{x_i}(h).$$

That is, if there is an error, we get -1 and otherwise we get 1. Putting it all together,

$$R_D(\mathcal{H}) = E_\sigma[\max_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i (1 - 2error_{x_i}(h)/m)]$$

$$= 2E_{S,\sigma}[\max_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i error_{x_i}(h)/m]. \tag{3.2}$$

From equations 3.1 and 3.2, we get:

$$E_S[MAXGAP(S)] \leq R_D(\mathcal{H}),$$

which completes the proof.

To sum up,

$$error_D(h) \leq error_S(h) + R_D(\mathcal{H}) + \sqrt{\frac{ln(2/\delta)}{m}},$$

which gives us

$$R_D(\mathcal{H}) \leq R_S(\mathcal{H}) + R_D(\mathcal{H}) + \sqrt{\frac{4ln(2/\delta)}{m}},$$

by using McDiarmid's inequality for $c_i = \frac{2}{m}$ and error ϵ, noting that $exp(-\frac{1}{2}\epsilon^2 m) = exp(\frac{-2\epsilon^2}{m(4/m^2)})$. For $\epsilon = \sqrt{\frac{4ln(2/\delta)}{m}}$ we get $\delta = e^{-2\epsilon^2 m}$. □

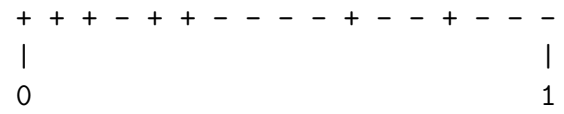
3.5 Model Selection - Introduction

So far, each learning model determined the number of examples needed in order to learn a concept class. However, in many real cases, only a limited number of examples is available, and the learning algorithm is supposed to come up with the best hypothesis it can from the available data.

In the algorithms discussed previously, we solved accuracy problems of our hypothesis by requiring a sufficiently large number of examples, which reduces the probability of the hypothesis' error. We now deal with the case in which this cannot be done.

3.5.1 Discussion

To demonstrate the problem, consider the concept class of a finite union of intervals on the line $[0, 1]$. Recall that this class has an infinite VC dimension. Let us assume that we are given the following examples in the interval $[0, 1]$:



The target concept c_t is a set of intervals within $[0, 1]$.

Obviously, if we allow a sufficiently large number of intervals, we could easily come up with a hypothesis that is completely consistent with the data (e.g., surround every positive point with its own tiny positive interval). However, we want to predict the correct classifications also for examples other than the original training set. Namely, get a good generalization.

Adding more intervals to our hypothesis reduces the hypothesis' error on the training set, but may increase its error on new unseen examples. For example, a positive interval surrounding a positive point may consist in the target concept of a 2/3 negative sub-interval and a 1/3 positive sub-interval, so adding this interval to the hypothesis can increase its

"real" error. This way we may get hypotheses which are overfitted to the data, and may not generalize well to new examples.

Returning to our example, we can make a table of the amount of errors generated by a hypothesis related to the number of intervals in the hypothesis :

Number of Intervals:	0	1	2	3	4	5	6	7	...
Number of Errors:	7	3	2	1	0	0	0	0	...

We can see that the more complex the hypothesis is, i.e., the more intervals, the lower its error on the training set. Beyond a certain complexity, all hypotheses have zero error. So far, we considered only those hypotheses which yield 0 errors on the training set, but now we are limited to the given examples and these examples may not be representative of the domain. Therefore, we want to consider simpler hypotheses, which may have some errors on the training set but generalize better to new examples.

To make the things worse, there is still the problem of noise. For a hypothesis to be completely consistent with the data, it becomes very complex. However, some of the inconsistencies in the data may be due to "noise", and the true concept may be much simpler than our consistent hypothesis. In the given example, the true concept may consist of a single interval (e.g. [0, 1/2]), and the inconsistent examples were generated due to noise. In such a case, adapting our hypothesis to the data causes the noise to get into the hypothesis.

So now we have to deal with a sample set which may be too small to accurately represent the domain, and may itself be "noisy". (Note that noise can be tiny interval of different label.) In the following we consider different models for dealing with this problem. But first we start with building a clear theoretical model.

3.6 Theoretical Model

3.6.1 The Setup

Let us consider the following theoretical model. Let H_i be the class of hypotheses, all having the same complexity-level, i (where i is some definition of complexity). Clearly, we get nested hypothesis classes

$$H_1 \subseteq H_2 \subseteq \cdots \subseteq H_i \subseteq \cdots$$

since any hypothesis of a lower complexity is included in any class of hypotheses of a higher complexity. Let $\mathcal{H} = \cup_{i=1}^{\infty} H_i$.

Let c_t be the target concept. We consider a non-realizable setting, namely, we do *not* assume that c_t is included in $\mathcal{H}$. (Alternatively, we can think of a realized setting, but having a restricted number of examples, or that $\mathcal{H}$ has infinite VC dimension.) The objective of the learning algorithm will be to produce a hypothesis which is "sufficiently close" to c_t.

3.6.2 Definitions

- $\epsilon(h)$ - the error of h with respect to the distribution $\mathcal{D}$ and c_t, namely, $\epsilon(h) = \Pr[h(x) \neq c_t(x)]$.

- ϵ_i - the lowest error found for any of the hypotheses in class H_i. Namely, $\epsilon_i = \min_{h \in H_i}\{\epsilon(h)\}$. Note that since $H_i \subseteq H_{i+1}$ it implies that $\epsilon_{i+1} \leq \epsilon_i$ (the probability of error decreases as the complexity level increases).

- ϵ^* - the optimal error rate, i.e., the value towards which ϵ_i converges as i increases. Namely, $\epsilon^* = \inf_i\{\epsilon_i\}$. It might be that ϵ^* will not be obtained by any hypothesis $h \in \mathcal{H}$, but it is the lower bound on any ϵ_i and could be approximated arbitrarily well. If for some i, $c_t \in H_i$ then $\epsilon^* = 0$.

- $\hat{\epsilon}(h)$ - the observed error, i.e., the error of hypothesis h on the given examples. Namely, $\hat{\epsilon}(h) = \frac{1}{m}\sum_{x_i \in S} I(h(x_i) \neq c_t(x_i))$, where S is the sample set of m examples.

- $\hat{\epsilon}_i$ - the lowest observed error of any of the hypotheses in H_i. Namely, $\hat{\epsilon}_i = \min_{h \in H_i}\{\hat{\epsilon}(h)\}$.

3.6.3 The Problem: Overfitting

As the complexity level i of the hypothesis increases, its error on the given data $\hat{\epsilon}_i$ is reduced. Beyond complexity level m (where m is the number of examples in the given set) all the $\hat{\epsilon}_i$ will equal 0. This will happen even when the same hypothesis' real error-level, $\epsilon(h)$, even when $\epsilon^* >> 0$. This happens because at many complex hypotheses can fit to the sample. This phenomenon is called *overfitting*.

In our case, we cannot require a sufficiently large set of examples. The given data may be too small to accurately represent the entire domain. The presence of noise makes the given data even less representative of the entire domain. Thus, the overfitted hypothesis might turn out to be quite far from the true concept.

The simplistic approach for finding a good hypothesis would be to choose a hypothesis g which has the lowest value of $\hat{\epsilon}(g)$, namely ERM which find $g = \arg\min_{h \in \cup\mathcal{H}}\{\hat{\epsilon}(h)\}$. However, using this simplistic approach for choosing g will cause us to prefer highly complex hypotheses over simple ones, and cause us to overfitted,

3.6.4 Penalty Based Model Selection

One way to overcome the overfitting problem is to impose a complexity penalty on the complexity of the chosen hypothesis; we will then try to minimize both the observed error of the chosen hypothesis and its complexity penalty.

The chosen hypothesis g^* will, therefore, be defined as

$$g^* = arg \min_{g \in \cup H_i} \left\{ \hat{\epsilon}(g) + Penalty(g) \right\} ,$$

where $Penalty(g)$ depends on the complexity of g.

We will define a measure $d(h)$ for the complexity of a hypothesis h as the lowest complexity level i such that h is found in H_i:

$$d(h) = \min_i \left\{ h \in H_i \right\} .$$

Since the penalty is calculated based on $d(h)$, which is the first class in which h is found, the penalty will be the same for all hypotheses with the same complexity.

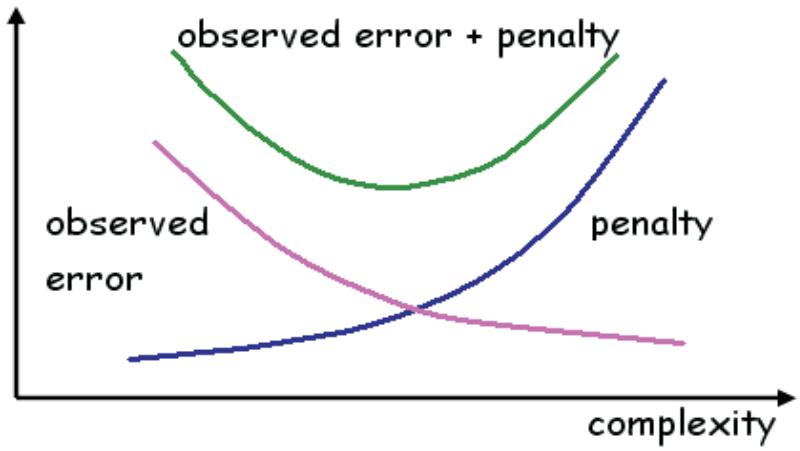


Figure 3.1: Principle of penalty based models.

Figure 3.1 shows the principle of penalty based models. As the complexity level of the hypothesis increases, its observed error is reduced but the penalty for its complexity increases. The penalty based model will try to find the minimum of the sum of the observed error and the penalty. Thus we will choose hypotheses that are not too fitted to the given examples.

3.6.5 SRM: Structural Risk Minimization

Assume that we have parameters λ_i and δ_i such that $\sum_{i=1}^{\infty} \delta_i$. We will assume that

$$\Pr[\exists h \in H_i : |\hat{\epsilon}(h) - \epsilon(h)| > \lambda_i] \leq \delta_i$$

Examples of such parameters can be $\delta_i = \delta/2^i$ and we have $\sum_{i=1}^{\infty} \delta_i = \delta$. For λ_i if H_i are finite we can have

$$\lambda_i = \sqrt{\frac{2\ln(2|H_i|/\delta_i)}{m}} = \sqrt{\frac{2\ln(2^{i+1}|H_i|/\delta)}{m}}$$

In H_i has VC-dimension i then we can have

$$\lambda_i = \sqrt{\frac{2i\ln(1/\delta_i)}{m}} = \sqrt{\frac{i^2 + i\ln(1/\delta)}{m}}$$

The *SRM* (*Structural Risk Minimization*) model is a penalty based model, which uses the following as the *Penalty* :

$$Penalty(h) = \lambda_{d(h)} \tag{3.3}$$

This penalty defines a tradeoff between the complexity of the hypothesis and the size of the sample. The hypothesis g^* chosen by the *SRM* model will therefore be:

$$g_{srm} = \arg\min_{g \in H} \left\{ \hat{\epsilon}(g) + Penalty(g) \right\} \tag{3.4}$$

Analysis

Let h^* be the best possible hypothesis there in $\mathcal{H}$, i.e., the hypothesis with the lowest actual error-level (error measured over the entire domain):

$$h^* = \arg\min_{h \in \mathcal{H}} \{\epsilon(h)\} \ . \tag{3.5}$$

Let g_{srm} be the hypothesis chosen by *SRM*, i.e., (3.4).

Theorem 3.16 (*SRM* Theorem) *With probability of at least* $1-\delta$ *the actual error of* g_{srm} *is at most the actual error of* h^* *plus twice the* SRM *complexity-penalty of* h^**. Formally :*

$$\epsilon(g_{srm}) \leq \epsilon(h^*) + 2 \cdot Penalty(h^*) \tag{3.6}$$

Recall that by definition (of h^*) the actual error of h^* is at most the actual error of g_{srm}. So, according to the *SRM* theorem, the actual error of g_{srm} is bounded on both sides by:

$$\epsilon(h^*) \ \leq \ \epsilon(g_{srm}) \ \leq \ \epsilon(h^*) + 2 \cdot Penalty(h^*) \tag{3.7}$$

It can be clearly seen from this inequality that the larger the number of examples (the larger m), the smaller the value of the complexity-penalty becomes, and the difference between the two hypotheses diminishes.

Proof of *SRM* Theorem

The proof consists of two stages. The first stage bounds the error of the hypothesis in any given class H_i. The second bounds the error across the classes H_i.

First stage : Bounding the error in H_i

Let g_i be the hypothesis with the lowest observed error in H_i:

$$g_i = \arg\min_{h \in H_i} \{\hat{\epsilon}(h)\}$$

We want to estimate the probability of difference between the actual error and the observed error of g_i:

$$\Pr\Big[|\epsilon(g_i) - \hat{\epsilon}(g_i)| > \lambda_i\Big]$$

We cannot use Chernoff to bound this probability, because g_i is determined according to the given sample set. However, we can bound this probability by the probability that *any* hypothesis in H_i will have the difference larger than λ_i:

$$\Pr\Big[|\epsilon(g_i) - \hat{\epsilon}(g_i)| > \lambda_i\Big] \leq \Pr\Big[\exists h \in H_i \mid |\epsilon(h) - \hat{\epsilon}(h)| > \lambda_i\Big].$$

By the definition of λ_i and δ_i we have,

$$\Pr\Big[\exists h \in H_i \mid |\epsilon(h) - \hat{\epsilon}(h)| > \lambda_i\Big] \leq \delta_i$$

Second stage : Bounding the error across H_i

We showed that with probability of at least $1 - \delta$, for any hypothesis $h \in \mathcal{H}$,

$$|\epsilon(h) - \hat{\epsilon}(h)| \leq \lambda_{d(h)} \ .$$

This is also true for h^* and g_{srm}, which leads to the following:

$$\hat{\epsilon}(h^*) \leq \epsilon(h^*) + \lambda_{d(h^*)} \tag{3.8}$$

and

$$\epsilon(g_{srm}) - \lambda_{d(g_{srm})} \leq \hat{\epsilon}(g_{srm}) \tag{3.9}$$

Let's define $P_{d(h^*)}$, $P_{d(g_{srm})}$ as the *SRM* complexity-penalties for h^* and g_{srm}, respectively. Therefore, from the definition of the *SRM* model we get :

$$\hat{\epsilon}(g_{srm}) + P_{d(g_{srm})} \leq \hat{\epsilon}(h^*) + P_{d(h^*)} \tag{3.10}$$

From the three inequalities (3.8), (3.9) and (3.10) we get:

$$\epsilon(g_{srm}) - \lambda_{d(g_{srm})} + P_{d(g_{srm})} \leq \epsilon(h^*) + \lambda_{d(h^*)} + P_{d(h^*)} \tag{3.11}$$

and therefore,

$$\epsilon(g_{srm}) \leq \epsilon(h^*) + \lambda_{d(h^*)} + P_{d(h^*)} + \lambda_{d(g_{srm})} - P_{d(g_{srm})} \ . \tag{3.12}$$

Now, from the definition of the penalty we get that $P_{d(g_{srm})} = \lambda_{d(g_{srm})}$. This implies that,

$$\epsilon(g_{srm}) \leq \epsilon(h^*) + \lambda_{d(h^*)} + P_{d(h^*)} + (\lambda_{d(g_{srm})} - P_{d(g_{srm})}) = \epsilon(h^*) + 2 \cdot P_{d(h^*)} \tag{3.13}$$

which proves the *SRM* theorem. □

3.7 Hypothesis Validation

The *SRM* model tackled the overfitting problem by imposing a complexity penalty on the "price" of a hypothesis, which will steer us to prefer simpler hypotheses rather than complex ones. The model shows that the chosen hypotheses will not be too fitted to the given examples; this will enable the hypotheses to correctly classify also new examples which were not used in the learning process.

The *Hypothesis Validation* method does not change the price of the hypothesis, but leaves it to be the observed error, $\hat{\epsilon}(h)$. To overcome the overfitting problem, the *Hypothesis Validation* splits the given set of examples, S, into two sets, S_1 and S_2. The set S_1, is used as the training sample set in the learning process; this yields for each H_i some hypothesis which is estimated to be the best according to the training set. The examples of the other set, S_2, are then used as a test set, to test the error of the chosen hypotheses on the "new" examples. The chosen hypothesis will be the one with the lowest observed error on the "test" set, S_2. Therefore, *Hypothesis Validation* deals with the overfitting problem by estimating how bad a hypothesis is when learning new examples (how tightly fit it is to the training sample set).

We denote by γ the fraction ($0 < \gamma < 1$) of the original set, S, which is reserved as the test set, S_2. Therefore, if the original set S contains m examples, then the test set S_2 contains γm examples, and the training set S_1 contains $(1-\gamma)m$ examples. Usually γ will be small, because after choosing the best hypotheses according to the training set S_1, the number of candidate hypotheses is reduced and thus we need less examples to choose the best one of the selected hypotheses according to the test set S_2.

We will partition the algorithm into two stages:

1. **Learning from** S_1:
   From each hypotheses class, H_i, we choose the best hypothesis g_i according to the training sample set S_1, i.e., the hypothesis which has the lowest observed error on S_i. Let
   $$g_i = arg \min_{h \in H_i} \{\hat{\epsilon}_1(h)\},$$

where $\hat{\epsilon}_1(h)$ is the observed error of hypothesis h on the training sample set S_1.

This will yield a set of hypotheses, G, with one hypothesis g_i for each class H_i.

Note that, for practicality's sake, we take $1 \leq i \leq m$; the best hypotheses from classes with complexity greater than or equal to m will have already become completely fitted to the data, and yield $\hat{\epsilon}_1(h) = 0$. We will therefore assume that $|G| = m$. (Alternatively, we will include g_i in G only if $\hat{\epsilon}_1(g_i) < \hat{\epsilon}_1(g_{i-1})$, and this can happen at most m times.)

2. **Testing on** S_2:
From G we now choose the hypothesis which has the lowest error on the test sample set S_2. Let
$$g_{srm} = arg\min_{g_i \in G}\{\hat{\epsilon}_2(g_i)\},$$
where $\hat{\epsilon}_2(h)$ is the observed error of hypothesis h on the test sample set S_2.

Analysis

We consider an arbitrary algorithm A which uses $(1-\gamma)m$ example to generate G and then selects some $g_i \in G$. Alternatively, we can compare to $\min_{g_i \in G} \epsilon(g_i)$.

Theorem 3.17 *Let $\epsilon_{HV}(m)$ be the error of* Hypothesis Validation *(HV) on m samples, and $\epsilon_A(m)$ be the error of some algorithm A on m samples. With probability $1-\delta$,*

$$\epsilon_{HV}(m) \leq \epsilon_A\big((1-\gamma)m\big) + 2 \cdot \sqrt{\frac{\ln(2m/\delta)}{\gamma m}}$$

Proof: First, we would like to bound the difference between the observed error and the actual error (both on the test set S_2) of any hypothesis g_i in G (the set of best hypothesis from each H_i, as chosen by *Hypothesis Validation*).

The probability that a hypothesis g_i in G will have the difference between its observed error on S_2 and its actual error larger than λ, is bounded as follows:

$$Prob\Big[|\epsilon(g_i) - \hat{\epsilon}_2(g_i)| \geq \lambda\Big] \leq 2 \cdot e^{-\lambda^2 \gamma m} \tag{3.14}$$

where $\hat{\epsilon}_2(g_i)$ is the observed error on the test set S_2 for hypothesis g_i.

Therefore, we can bound the probability that *any* hypothesis in G will have the difference between its actual error and observed error on S_2 larger than λ by:

$$Prob\Big[\exists g \in G \mid |\epsilon(g) - \hat{\epsilon}_2(g)| \geq \lambda\Big] \leq |G| \cdot 2e^{-\lambda^2 \gamma m} \tag{3.15}$$

Since $|G| = m$ we get:

$$Prob\Big[\exists g \in G \mid |\epsilon(g) - \hat{\epsilon}_2(g)| \geq \lambda\Big] \leq m \cdot 2e^{-\lambda^2 \gamma m} \tag{3.16}$$

If we set this upper bound to δ, we get:

$$\delta = m \cdot 2e^{-\lambda^2 \gamma m}$$

Solving for λ leads to:

$$\lambda = \sqrt{\frac{\ln(2m/\delta)}{\gamma m}} \tag{3.17}$$

Therefore, with probability $1 - \delta$ we have for any g_i in G :

$$|\epsilon(g_i) - \hat{\epsilon}_2(g_i)| \leq \lambda. \tag{3.18}$$

From this we get:

$$\epsilon(g_{HV}) - \lambda \leq \hat{\epsilon}_2(g_{HV}) \tag{3.19}$$

and

$$\hat{\epsilon}_2(g_A) \leq \epsilon(g_A) + \lambda. \tag{3.20}$$

Since *Hypothesis Validation* preferred g_j to g_k, we know:

$$\hat{\epsilon}_2(g_{HV}) \leq \hat{\epsilon}_2(g_A). \tag{3.21}$$

Now, from the three inequalities (3.19), (3.20) and (3.21) we can get:

$$\epsilon(g_j) - \lambda \leq \hat{\epsilon}_2(g_j) \leq \hat{\epsilon}_2(g_k) \leq \epsilon(g_k) + \lambda$$

which implies that

$$\epsilon(g_j) \leq \epsilon(g_k) + 2\lambda.$$

□