XENICAL®
orlistat

The PAC Learning Framework (Mohri)

The Probably Approximately Correct learning framework helps define the class of learnable concepts in terms of the number of sample points needed to achieve an approximate solution, sample complexity and the time and space complexity of the learning algorithm.

X: all possible examples of instances (input space)
Y: set of all possible labels (for now it will be binary)
$c: X \to Y$ a concept, $C \to$ concept class, set of concepts to learn

Assumption: examples are i.i.d. according to some distribution D.
The learner considers a fixed set of possible concepts H, called a hypothesis set, which might not necessarily coincide with C. It receives a sample $S = (x_1, \dots, x_m)$ drawn i.i.d. from D as well as the labels $(c(x_1), \dots, c(x_m))$, which are based on a specific target concept $c \in C$ to learn. The task is to use the labeled sample S to select a hypothesis $h_S \in H$ that has a small generalization error with respect to the concept c.

↳ $R(h) = \underset{x \sim D}{\mathbb{P}}[h(x) \neq c(x)] = \underset{x \sim D}{\mathbb{E}}[1_{h(x) \neq c(x)}]$ ← indicator function

Since D and c are unknown the generalization error is not directly

accessible, so the learner has to measure the empirical error:

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} 1_{h(x_i) \neq c(x_i)}$$

For a fixed $h \in \mathcal{H}$, the expectation of the empirical error based on an i.i.d. sample S is equal to the generalization error:

$$\mathop{\mathbb{E}}_{S \sim D^m} [\hat{R}_S(h)] = R(h)$$

Now, let n be a number such that the computational cost of representing any element $x \in \mathcal{X}$ is at most $\mathcal{O}(n)$ and denote by size(c) the maximal cost of the computational representation of $c \in \mathcal{C}$. Let h_S denote the hypothesis returned by algorithm $\mathcal{A}$ after receiving a labeled sample S. Then

A concept class $\mathcal{C}$ is said to be PAC-learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function poly(·,·,·,·) such that for any $\epsilon > 0$ and $\delta > 0$, for all distributions D on $\mathcal{X}$ and for any target concept $c \in \mathcal{C}$, the following holds for any sample size $m \geq \text{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \text{size}(c))$:

$\epsilon \sim$ accuracy
$\delta \sim$ confidence

$$\mathop{\mathbb{P}}_{S \sim D^m} [\underbrace{R(h_S) \leq \epsilon}_{\text{approximately}}] \geq \underbrace{1-\delta}_{\text{probably}}$$

If $\mathcal{A}$ further runs in poly$(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \text{size}(c))$, then $\mathcal{C}$ is said to be efficiently PAC-learnable. When such an algorithm $\mathcal{A}$ exists, it is called a PAC-learning algorithm for $\mathcal{C}$.

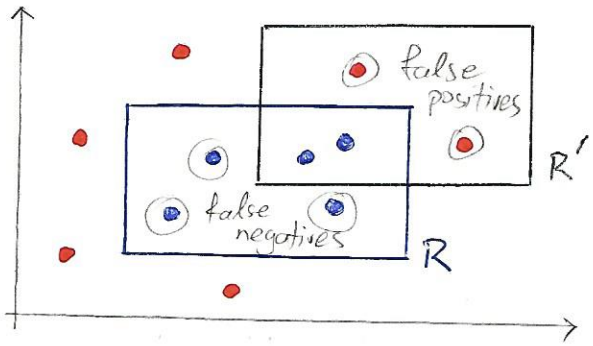
→ The PAC framework is a distribution-free model.

→ The question of learnability for a concept class $\mathcal{C}$ and not a particular concept c is addressed. The concept class is known to the algorithm, but the concept isn't.

XENICAL®
orlistat

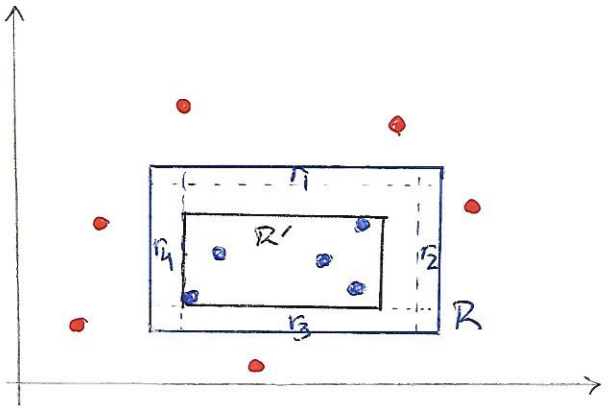
Learning axis-aligned rectangles

Let $X = \mathbb{R}^2$ and C is the set of all axis-aligned rectangles lying in $\mathbb{R}^2$ and therefore c is the set of points inside a particular axis-aligned rectangle. The learning problem consists of determining with small error a target axis-aligned rectangle, using the labeled training sample.



R is a target and R' is a hypothesis. The error regions of R' are formed by the area within the rectangle R but outside the rectangle R' and the area within R' but outside R.

To show that the concept is PAC-learnable, we describe a simple PAC-learning algorithm A. Given a labeled sample S, the algorithm consists of returning the tightest axis-aligned rectangle $R' = R_S$ containing the blue points. Then, R_S does not produce any false positives, since its points must be included in the target concept R. Thus, the error region of R_S is included in R.



Let R be a target concept. Fix $\epsilon > 0$. Let $\mathbb{P}[R]$ denote the probability mass of the region defined by R, that is the probability of a point randomly drawn according to D falling within R. Since errors made by our algorithm can only be due to points falling inside R, we can assume that $\mathbb{P}[R] > \epsilon$; otherwise, the error of R_S is less than or equal to ϵ regardless of the sample S received.

Since $\mathbb{P}[R] > \epsilon$, we can define four rectangular regions r_1, r_2, r_3, r_4 along the sides of R, each with probability at least $\epsilon/4$. These can be constructed by starting with the full rectangle R and then decreasing the size by moving one side as much as possible while keeping a distribution mass of at least $\epsilon/4$. If R is defined as $R = [\ell, r] \times [b, t]$, then, for example, $r_4 = [\ell, s_4] \times [b, t]$, where $s_4 = \inf\{s : \mathbb{P}[[\ell, s] \times [b, t]] \geq \epsilon/4\}$. Similarly for r_3, r_2, r_1.

If R_S meets all of these four regions r_i, $i = 1, \ldots, 4$, then, because it is a rectangle, it will have one side in each of these regions. Its error area, which is the part of R that it does not cover, is thus included in the union of the regions r_i and cannot have probability mass more than ϵ. If $R(R_S) > \epsilon$, then R_S must miss at least one of the regions r_i. As a result one can write:

$$\mathbb{P}_{S \sim D^m}[R(R_S) > \epsilon] \leq \mathbb{P}_{S \sim D^m}\left[\bigcup_{i=1}^{4} \{R_S \cap r_i = \emptyset\}\right] \leq \sum_{i=1}^{4} \mathbb{P}_{S \sim D^m}[\{R_S \cap r_i = \emptyset\}]$$

XENICAL®
orlistat

because $1-x \leq e^{-x}$. For any $\delta > 0$, to ensure that $\mathbb{P}_{S\sim D^m}[R(R_S) > \epsilon] \leq \delta$, we can impose:

$$4e^{-\frac{m\epsilon}{4}} \leq \delta \iff m \geq \frac{4}{\epsilon}\ln\left(\frac{4}{\delta}\right) \quad (*)$$

Thus, for any $\epsilon > 0$ and $\delta > 0$, if the sample size m is greater than $4/\epsilon \cdot \ln(4/\delta)$, then $\mathbb{P}_{S\sim D^m}[R(R_S) > \epsilon] \leq \delta$ holds. Therefore the concept class is learnable with complexity $\mathcal{O}\left(\frac{1}{\epsilon}\ln\left(\frac{1}{\delta}\right)\right)$.

An equivalent way of presenting sample complexity results like (*) is to give a **generalization bound**. A generalization bound states that with probability at least $1-\delta$, $R(R_S)$ is upper bounded by some quantity that depends on the sample size m and δ. Using the above results we get $R(R_S) \leq \frac{4}{m}\ln\left(\frac{4}{\delta}\right)$. Can we find more general results to use in more complex cases?

Note: We shall consider consistent hypotheses in the case where the cardinality $|\mathcal{H}|$ of the hypothesis set is finite (unlike in the above case). Since we consider consistent hypotheses, we will assume that the target concept c is in $\mathcal{H}$.

Let $\mathcal{H}$ be a finite set of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$. Let A be an algorithm that for any target concept $c \in \mathcal{H}$ and i.i.d. sample S returns a consistent hypothesis h_S: $\hat{R}_S(h_S) = 0$. Then, for any $\epsilon, \delta > 0$

the inequality $\mathbb{P}_{S\sim D^m}[R(h_S) \leq \epsilon] \geq 1-\delta$ holds if

$$m \geq \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right).$$

This sample complexity result admits the following equivalent statement as a generalization bound: for any $\epsilon, \delta > 0$, with probability at least $1-\delta$,

$$R(h_S) \leq \frac{1}{m}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

Inconsistent case

In the most general case, there may be no hypothesis in $\mathcal{H}$ consistent with the labeled training sample. To derive learning guarantees in this more general setting, we will use Hoeffding's inequality, which relates the generalization error and empirical error of a single hypothesis.

→ Fix $\epsilon > 0$. Then, for any hypothesis $h: X \to \{0,1\}$, the following hold:

$$\left.\begin{aligned} \mathbb{P}_{S\sim D^m}[\hat{R}_S(h) - R(h) \geq \epsilon] &\leq e^{-2m\epsilon^2} \\ \mathbb{P}_{S\sim D^m}[\hat{R}_S(h) - R(h) \leq -\epsilon] &\leq e^{-2m\epsilon^2} \end{aligned}\right\} \text{ combining these yields}$$

$$\mathbb{P}_{S\sim D^m}[|\hat{R}_S(h) - R(h)| \geq \epsilon] \leq 2e^{-2m\epsilon^2}$$

Fix a hypothesis $h: X \to \{0,1\}$. Then, for any $\delta > 0$, the following inequality holds with probability at least $1-\delta$:

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2m}}$$

As we did in the consistent case, we need to derive a uniform convergence bound which holds for all hypotheses $h \in \mathcal{H}$.

XENICAL®
orlistat

Let $\mathcal{H}$ be a finite hypothesis set. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds:

$$\forall h \in \mathcal{H}, \quad R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\ln|\mathcal{H}| + \ln\left(\frac{2}{\delta}\right)}{2m}}$$

Note: Now we have $\mathcal{O}\left(\sqrt{\frac{\ln|\mathcal{H}|}{m}}\right)$, while before we had $\mathcal{O}\left(\frac{\ln|\mathcal{H}|}{m}\right)$. This means that, for a fixed $|\mathcal{H}|$, to attain the same guarantee as in the consistent case, a quadratically larger labeled sample is needed.

Rademacher Complexity & VC-dimension (Mohri)

We now want to examine the infinite hypothesis space scenario.
Idea: reduction to finite sets and perform the previous analysis

Rademacher Complexity

Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[a,b]$ and $S = (z_1, \dots, z_m)$ a fixed sample of size m with elements in $\mathcal{Z}$.
Then, the empirical Rademacher complexity of $\mathcal{G}$ with respect to the sample is defined as

$$\hat{\mathcal{R}}_S(\mathcal{G}) = \mathop{\mathbb{E}}_{\vec{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i)\right],$$

where $\vec{\sigma} = (\sigma_1, \dots, \sigma_m)^T$, with σ_i: independent uniform random variables taking values in $\{-1, +1\}$, which are called Rademacher variables.

$L: Y \times Y \to \mathbb{R}$: loss function
$\mathcal{G}$: family of loss functions associated to $\mathcal{H}$

$$\mathcal{G} = \{g: (x,y) \mapsto L(h(x), y) : h \in \mathcal{H}\}$$

The inner product $\vec{\sigma} \cdot \vec{g}_S$, where $\vec{g}_S = (g(z_1), \ldots, g(z_m))^T$ measures the correlation of $\vec{g}_S$ with the vector of random noise $\vec{\sigma}$. The supremum is a measure of how well the function class $\mathcal{G}$ correlates with $\vec{\sigma}$ over the sample S. Thus, the empirical Rademacher complexity measures on average how well the function class $\mathcal{G}$ correlates with random noise on S.

Let D denote the distribution according to which samples are drawn. For any integer $m \geq 1$, the Rademacher complexity of $\mathcal{G}$ is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to D:

$$\mathcal{R}_m(\mathcal{G}) = \mathop{\mathbb{E}}_{S \sim D^m}\left[\hat{\mathcal{R}}_S(\mathcal{G})\right]$$

Compare this to the 2nd equation of page (2).

Let $\mathcal{G}$ be a family of functions mapping from Z to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m, each of the following holds for all $g \in \mathcal{G}$:

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(\mathcal{G}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \quad \text{and}$$

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\hat{\mathcal{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}$$

We shall then relate the empirical Rademacher complexities of a hypothesis set $\mathcal{H}$ to the family of loss functions $\mathcal{G}$ associated to $\mathcal{H}$ in the case of binary loss (zero-one loss):

Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ and let $\mathcal{G}$ be the family of loss functions associated to $\mathcal{H}$ for the zero-one loss: $\mathcal{G} = \{(x,y) \mapsto \mathbb{1}_{h(x) \neq y} : h \in \mathcal{H}\}$. For any sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of elements in $\mathcal{X} \times \{-1, +1\}$, let $S_{\mathcal{X}}$ denote the projection over $\mathcal{X}$: $S_{\mathcal{X}} = (x_1, \ldots, x_m)$.

Then, the following relation holds:

XENICAL®
orlistat

Using this equation and taking expectations one receives:

$$\mathcal{R}_m(\mathcal{G}) = \tfrac{1}{2}\mathcal{R}_m(\mathcal{H})$$

Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ and let D be the distribution over the input space X. Then, for any $\delta > 0$, with probability at least $1-\delta$ over a sample S of size m drawn according to D, each of the following holds for any $h \in \mathcal{H}$:

$$R(h) \leq \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \quad (i)$$

and $$R(h) \leq \hat{R}_S(h) + \hat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}$$

Problem: Computing $\hat{\mathcal{R}}_S(\mathcal{H})$, which is data-dependent, is hard. For this reason we will relate the Rademacher complexity to the growth function.

The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ for a hypothesis set $\mathcal{H}$ is defined by:

$$\forall m \in \mathbb{N}, \quad \Pi_{\mathcal{H}}(m) = \max_{\{x_1, \ldots, x_m\} \subseteq X} \left| \{(h(x_1), \ldots, h(x_m)) : h \in \mathcal{H}\} \right|$$

In other words, $\Pi_{\mathcal{H}}(m)$ is the maximum number of distinct ways in which m points can be classified using hypotheses in $\mathcal{H}$. Each one of these distinct classifications is called a dichotomy and, thus, the growth function counts the number of dichotomies that are realized by the

This provides another measure of the richness of the hypothesis set $\mathcal{H}$. However, unlike the Rademacher complexity, this measure does not depend on the distribution, it is purely combinatorial.

Using Massart's lemma:

$$\mathbb{E}_{\vec{\sigma}}\left[\frac{1}{m}\sup_{\vec{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\right] \leq \frac{r\sqrt{2\ln|A|}}{m},$$

one can prove the following:

Let $\mathcal{G}$ be a family of functions taking values in $\{-1, +1\}$. Then, the following holds:

$$\mathcal{R}_m(\mathcal{G}) \leq \sqrt{\frac{2\ln \Pi_{\mathcal{G}}(m)}{m}}$$

As a result, one can write (i) as:

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{2\ln \Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}} \quad (ii)$$

Since $\Pi_{\mathcal{H}}(m)$ needs to be computed for all $m \geq 1$, the computation of the growth function will not always be convenient. We can thus introduce an alternative measure of the complexity of a hypothesis set $\mathcal{H}$ that is based on a single scalar, which will turn out to be deeply related to the behaviour of the growth function.

VC-dimension

Shattering: A set S of $m \geq 1$ points is said to be shattered by a hypothesis set $\mathcal{H}$ when $\mathcal{H}$ realizes all possible dichotomies of S, i.e. when $\Pi_{\mathcal{H}}(m) = 2^m$.

The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be shattered by $\mathcal{H}$:

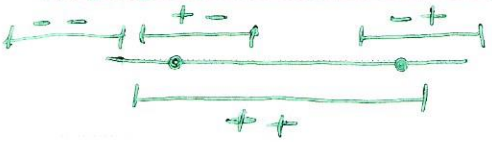
$$VCdim(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}$$

By definition, if $VCdim(\mathcal{H}) = d$, there exists a set of size d that can be shattered. However, not <u>all</u> sets of size d are shattered.

XENICAL®
orlistat

To compute the VC-dimension we show a lower bound for its value and then a matching upper bound. To give a lower bound d for $VCdim(\mathcal{H})$, it suffices to show that a set S of cardinality d can be shattered by $\mathcal{H}$. To give an upper bound, we need to prove that no set S of cardinality $d+1$ can be shattered by $\mathcal{H}$.

Intervals on the real line



The VC-dimension is at least 2, since all four dichotomies $(+,+)$, $(-,+)$, $(+,-)$, $(-,-)$ can be realized. By definition of intervals, no set of three points can be shattered, as the labelings $(+,-,+)$ and $(-,+,-)$ cannot be realized. Hence, $VCdim(\text{intervals in } \mathbb{R}) = 2$.

In general, $VCdim(\text{hyperplanes in } \mathbb{R}^d) = d+1$

Also, the VC-dimension of any vector space of dimension $r < \infty$ can be shown to be at most r.

So, how is the growth function related to the VC-dimension?
Let $\mathcal{H}$ be a hypothesis set with $VCdim(\mathcal{H}) = d$. Then, for all $m \in \mathbb{N}$, the following inequality holds:

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

This leads to an important result:

Let $\mathcal{H}$ be a hypothesis set with $VCdim(\mathcal{H}) = d$. Then for all $m \geq d$:

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d = O(m^d)$$

This means that if $d < \infty$, $\Pi_{\mathcal{H}}(m) = O(m^d)$, while if $d = +\infty$, $\Pi_{\mathcal{H}}(m) = 2^m$. Furthermore, we can extend (ii) as follows:

Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in \mathcal{H}$:

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{2d \ln\left(\frac{em}{d}\right)}{m}} + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}},$$

so the form is $R(h) \leq \hat{R}_S(h) + O\left(\sqrt{\frac{\ln(m/d)}{m/d}}\right)$, which emphasizes the importance of the ratio m/d for generalization.

See Mohri for discussion on lower bounds.