Generalization theory

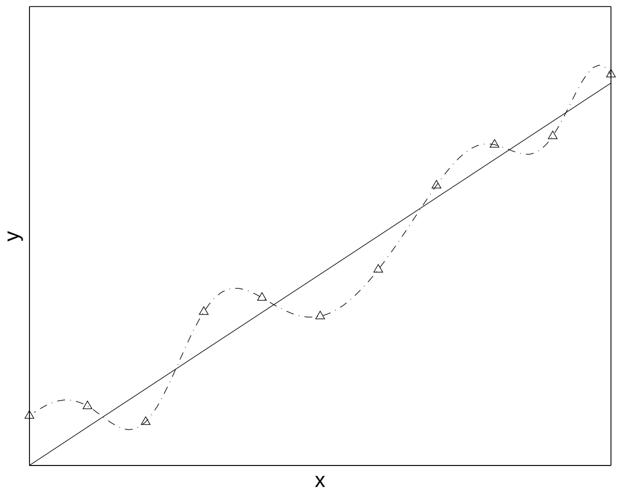
Chapter 4

March 19, 2003

T.P. Runarsson (tpr@hi.is) and S. Sigurdsson (sven@hi.is)

Introduction

Suppose you are given the empirical observations, $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell) \subseteq (\mathcal{X} \times \mathcal{Y})^\ell$. Consider the regression estimation problem where $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and the data are the following points:



- where the dash-dot line represents are fairly complex model and fits the data set perfectly, and
- the straight line does not completely “explain” the data but seems to be a “simpler” model, if we argue that the residuals are possibly measurement errors.

Is it possible to characterize the way in which the straight line is simpler and in some sense closer in explaining the true functional dependency of x and y?
$\longrightarrow$ *Statistical learning theory* provides a solid mathematical framework for studying these questions in depth.

Underfitting versus overfitting

In the machine learning literature the more complex model is said to show signs of *overfitting*, while the simpler model *underfitting*.

Often several heuristic are developed in order to avoid overfitting, for example, when designing neural networks one may:

1. limit the number of hidden nodes (equivalent to reducing the number of support vectors),

2. stop training early to avoid a perfect explanation of the training set, and

3. apply weight decay to limit the size of the weights, and thus of the function class implemented by the network (equivalent to the regularization used in ridge regression).

Minimizing risk (expected loss)

The risk or expected loss (L) is defined as:

$$\mathsf{err}_{\mathcal{D}}(h) = \mathsf{E}[\tfrac{1}{\ell}\mathsf{err}_{S_t}(h)] = \int_{\mathcal{X}\times\mathcal{Y}} L(\boldsymbol{x}, y, h(\boldsymbol{x}))d\mathcal{D}(\boldsymbol{x}, y)$$

where S_t is a test set. This problem is intractable since we do not know $\mathcal{D}(x, y)$ which is the probability distribution on $\mathcal{X} \times \mathcal{Y}$ which governs the data generation and underlying functional dependency.

The only thing we do have it our training set S sampled from the distribution $\mathcal{D}$ and we can use this to approximate the above integral by the finite sum:

$$\frac{1}{\ell}\mathsf{err}_S(h) = \frac{1}{\ell}\sum_{i=1}^{\ell} L(\boldsymbol{x}_i, y_i, h(\boldsymbol{x}_i))$$

If we allow h to be taken from a very large class of function $H = \mathcal{Y}^{\mathcal{X}}$, we can always find a h that leads to a rather small value of $\frac{1}{\ell}\mathsf{err}_S(h)$, but yet be distant from minimizer of $\mathsf{err}_{\mathcal{D}}(h)$. For example, a look-up-table for the training set would give excellent results for that set, but contain no information about any other points.

The message: *if we make no restriction on the class of functions from which we choose our estimate h, we cannot hope to learn anything.*

A learning model

- **Instance space**: is the set $\mathcal{X}$ of encodings of instances or objects in the learner's world. For example, the instance space could be points in the Euclidean plane $\mathbb{R}^n$.

- A **concept** over $\mathcal{X}$ is a subset $c \subseteq \mathcal{X}$ of the instance space. For example, a concept may be a hyperplane in $\mathbb{R}^n$. We define a concept to be a boolean mapping $c : \mathcal{X} \mapsto \{-1, 1\}$, with $y = c(\boldsymbol{x}) = 1$ indicating that $\boldsymbol{x}$ is a positive example of c and $y = c(\boldsymbol{x}) = -1$ implying a negative one.

- A **concept class** C over $\mathcal{X}$ is a collection of concepts over $\mathcal{X}$.

- The **learning algorithm** will have access to positive and negative examples of an unknown **target concept** c, chosen from a known concept class C. *Important:* learning algorithms know the target class C, the designer of a learning algorithm must design the algorithm to work for any $c \in C$.

- For the reasons above $\mathcal{X}$ is sometimes called the **input space**.

- The **target distribution** $\mathcal{D}$ is any fixed probability distribution over the instance space $\mathcal{X}$.

- Let the **oracle** EX$(c, \mathcal{D})$ be a procedure that returns a labeled example $(\boldsymbol{x}, y)$, $y = c(\boldsymbol{x})$, on each call, where $\boldsymbol{x}$ is drawn randomly and independently according to $\mathcal{D}$.

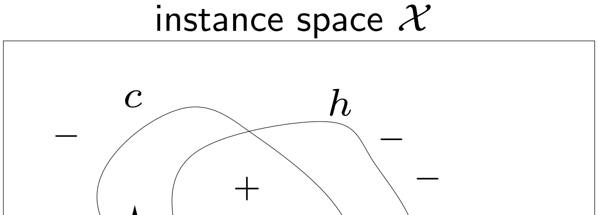
- We now define a **hypothesis space** H as concepts over $\mathcal{X}$. Elements of H are called hypotheses and are denoted by h. Usually, we would like the hypothesis space to be C itself. However, sometimes it is computationally advantageous to use another H to approximate concepts in C.

- If h is any concept over $\mathcal{X}$, then the distribution of $\mathcal{D}$ provides a natural measure of **error** between h and the target concept:

$$\text{err}_{\mathcal{D}}(h) = \Pr_{\boldsymbol{x} \in \mathcal{D}}[y \neq h(\boldsymbol{x})]$$

(Note that in the book uses the notation: $\text{err}_{\mathcal{D}}(h) = \mathcal{D}\{(\boldsymbol{x}, y) : h(\boldsymbol{x}) \neq y)$, p. 53.)

An alternative way of viewing the error:


instance space $\mathcal{X}$
c
h
$-$
$-$
$-$
$-$
$+$
$+$
$+$
where $c(\boldsymbol{x}) = y$ and $h(\boldsymbol{x})$ disagree

The error is simply the probability w.r.t. $\mathcal{D}$ of falling in the area where the target concept and hypothesis disagree.

Probably approximately correct (PAC) learning

Definition: *Let C be a concept class over $\mathcal{X}$. We say that C is* **PAC learnable** *if there exists an algorithm L_r with the following properties: for every concept $c \in C$, for every distribution $\mathcal{D}$ on $\mathcal{X}$, and for all $0 < \varepsilon < 1/2$ and $0 < \delta < 1/2$, if L_r is given access to EX$(c, \mathcal{D})$ and inputs ε and δ, then with probability at least $1 - \delta$, L_r outputs a hypothesis concept $h \in C$ satisfying err$_{\mathcal{D}}(h) \leq \varepsilon$. This probability is taken over the random examples drawn by calls to EX$(c, \mathcal{D})$, and any internal randomization of L_r. That is* [1],

$$\Pr[\mathit{err}_{\mathcal{D}}(h) \leq \varepsilon] \geq 1 - \delta \quad \textit{or} \quad \Pr[\mathit{err}_{\mathcal{D}}(h) > \varepsilon] < \delta$$

If L_r runs in time polynomial in $1/\varepsilon$ and $1/\delta$, we say the C is **efficiently** *PAC learnable. We sometimes refer to the input ε as the* **error parameter**, *and the input δ as the* **confidence parameter**.

The hypothesis $h \in C$ of the *PAC* learning algorithm is thus *approximately correct* with high probability, hence the name $\underline{P}robably\ \underline{A}pproximately\ \underline{C}orrect\ learning$.

[1] Note that the bounds must hold for whatever distribution $\mathcal{D}$ generated the examples, a property known as *distribution free* learning. Since some distribution will make learning harder than others this is a pessimistic way of evaluating learning. As a result, many of the results in the literature about PAC-learning are negative, showing that a certain concept class is not learnable.

Empirical (sample) error

The **empirical error** (sample error) of hypothesis $h(\boldsymbol{x})$ with respect to the target $y = c(\boldsymbol{x})$ and data sample

$$S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_\ell, y_\ell)\}$$

is the number k of examples h misclassifies out of ℓ,

$$\mathsf{err}_S(h) = k.$$

A hypothesis h is **consistent with the training examples** S, when $\mathsf{err}_S(h) = 0$.

Consider a fixed inference rule for selecting a hypothesis h_S from the class H on a set S of ℓ training examples chosen i.i.d. according to $\mathcal{D}$. In this setting we can view the generalization error $\mathsf{err}_\mathcal{D}(h_S)$ as a random variable depending on the random selection of the training set.

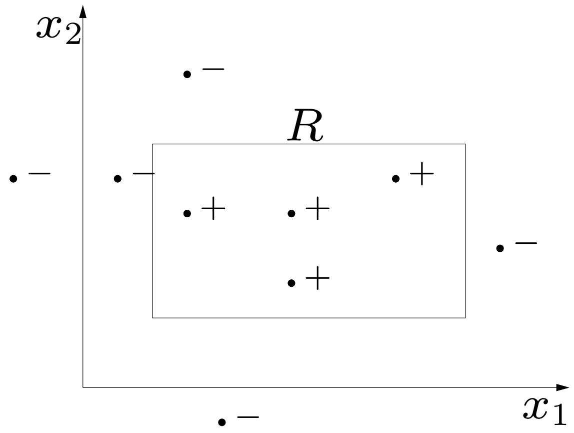
The PAC model of learning requires a generalization error bound that is unlikely to fail. It therefore bounds the tail of the distribution of the generalization error random variable $\mathsf{err}_\mathcal{D}(h_S)$. The size of the tail is determined by δ, hence

$$\mathsf{err}_\mathcal{D}(h_S) \leq \varepsilon(\ell, \delta, H)$$

the book uses the notation $\mathcal{D}^\ell\{S : \mathsf{err}_\mathcal{D}(h_S) > \varepsilon(\ell, H, \delta)\} < \delta$ here.

Example: the rectangle learning game

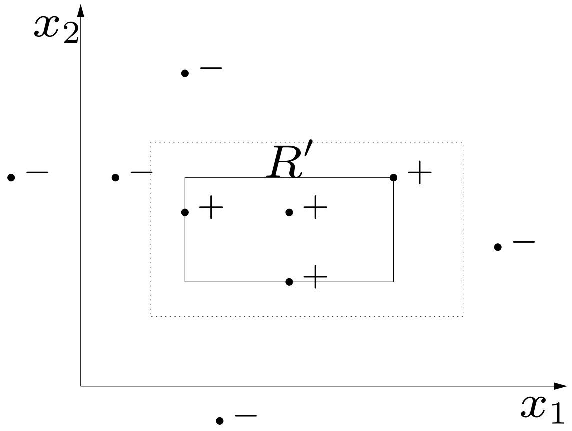
Consider the problem of learning the concept of "medium build" of a man defined as having weight (x_1) between 150 and 185 pounds and height (x_2) between $5'4''$ and $6'$. These bounds form a *target* rectangle R whose sides are parallel to the coordinate axis. A man's build is now represented by a point in the Euclidean plane $\mathbb{R}^2$. The target rectangle R in the plane $\mathbb{R}^2$ along with a sample (S) of positive and negative examples is depicted below:



The learner now encounters every man in the city with equal probability. This does not necessarily mean that the points will be uniformly distributed on $\mathbb{R}^2$, since not all heights and weights are equally likely. However, they obey some fixed distribution $\mathcal{D}$ on $\mathbb{R}^2$ which may be difficult to characterize.

The distribution $\mathcal{D}$ is therefore arbitrary, but fixed, and each example is drawn independently from this distribution. As $\mathcal{D}$ is arbitrary it is not necessary that the learner encounters every man with equal probability.

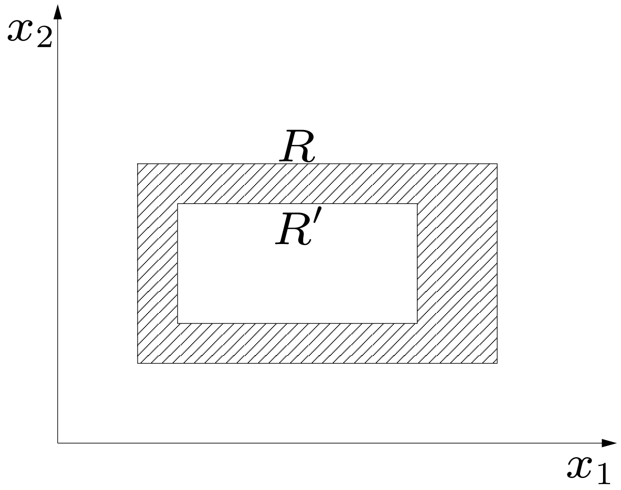
A simple algorithm to learn this concept is the following: after drawing ℓ samples and keeping track of all the positive examples, predict that the concept is the tightest-fit rectangle $h = R'$ as shown here:



If no positive examples are drawn then $h = R' = \emptyset$. Note that R' is always contained in R (i.e. $R' \subseteq R$).

We want to assert with probability at least $1 - \delta$ that the tightest-fit rectangle hypothesis R' has error at most ε, with respect to the concept R and distribution $\mathcal{D}$, for a given sample size ℓ.

The error of the hypothesis, $\text{err}_{\mathcal{D}}(h)$, is the probability of generating a sample in the hatched region between the rectangle R' and target rectangle R as shown below:



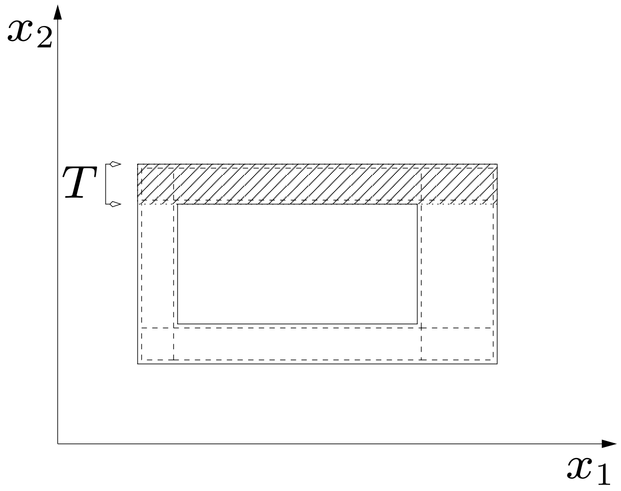
The error of the hypothesis, $\text{err}_{\mathcal{D}}(h)$, is the weight in the hatched region.

First note that if the weight inside the rectangle R is less than ε then clearly $\text{err}_{\mathcal{D}}(h) < \varepsilon$ since $(R' \subseteq R)$. If a hypothesis h is consistent with the sample S and the probability of hitting the shaded region is greater that ε then clearly this event is bounded by

$$\Pr[\text{err}_S(h) = 0 \text{ and } \text{err}_{\mathcal{D}}(h) > \varepsilon] \leq (1-\varepsilon)^{\ell}$$

that is, the probability of hitting the non-hatched region is at most $(1-\varepsilon)$ and doing so ℓ times $(1-\varepsilon)^{\ell}$ (the conjunction of independent events is simply the product of the probabilities of the independent events).

However, this is only one hypothesis of many which are consistent with *some* sample and have $\text{err}_{\mathcal{D}}(h) > \varepsilon$. Think of the rectangle R' being free to move within rectangle R depending on where the samples may fall. If we split the hatched region into 4 rectangular strips (left, right, top, bottom) we can limit the number of consistent hypothesis. The top rectangular strip is shown hatched:



Let the weight of a strip be the probability, with respect to $\mathcal{D}$, of falling into it. If the weight[2] of the four strips is at most $\varepsilon/4$ then the error of R' is at most $4(\varepsilon/4) = \varepsilon$. However, if $\text{err}_S(h) = 0$ and $\text{err}_{\mathcal{D}} > \varepsilon$, then the weight of one of these strips must be greater than $\varepsilon/4$. Without the loss of generality, consider the top strip marker by T.

[2] This is of course a pessimistic view since we count the overlapping corners twice.

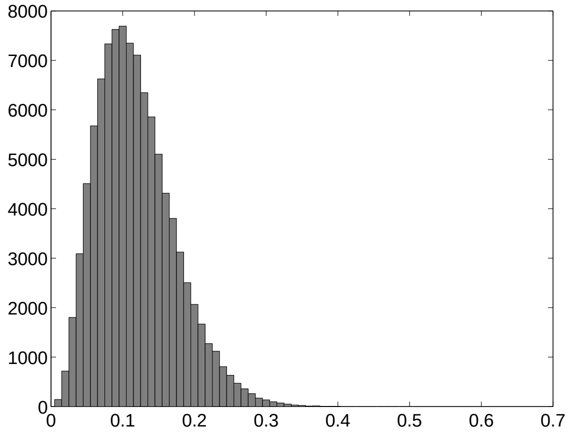
By definition of T, the probability of a single draw from the distribution $\mathcal{D}$ misses the region T is exactly $1-\varepsilon/4$. The probability of ℓ independent draws from $\mathcal{D}$ all miss the region T is exactly $(1-\varepsilon/4)^{\ell}$. The same analysis holds for the other three strips. Therefore, the probability that any of the four strips has a weight greater than $\varepsilon/4$ is at most $4(1-\varepsilon/4)^{\ell}$, by the union bound[3].

> If ℓ is chosen to satisfy $4(1-\varepsilon/4)^{\ell} \leq \delta$, then with probability $1-\delta$ over the ℓ random examples, the weights of the error region will be bounded by ε. Using the inequality
>
> $$(1-z) \leq \exp(-z)$$
>
> the above inequality may be written as $4\exp(-\varepsilon\ell/4) \leq \delta$ or equivalently $\ell \geq (4/\varepsilon)\ln(4/\delta)$, also known as the **sample complexity**.

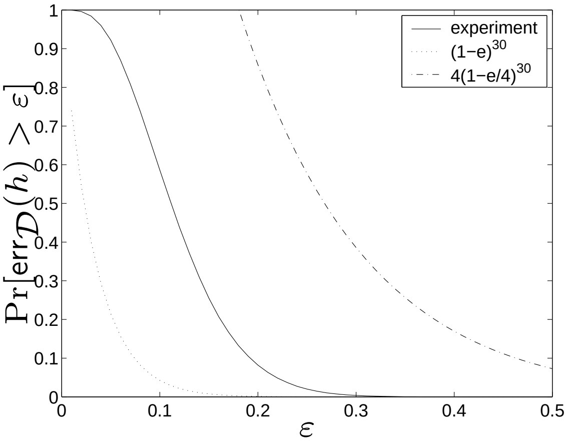
[3] The union bound: if A and B are any two events (subsets of a probability space), then $\Pr[A \cup B] \leq \Pr[A] + \Pr[B]$.

Numerical example

Let $\mathcal{D}$ be uniformly distributed over $[0\ \ 1]^2$ and let the rectangle $R = \boldsymbol{x}_1 \times \boldsymbol{x}_2 = (0.2, 0.8) \times (0.2, 0.9)$. The error in this case is simply $R - R'$. We now perform 100.000 independent experiments with $\ell = 30$ and record each time the error. Here is a histogram of the error:



and the probability that $\Pr[\mathsf{err}_{\mathcal{D}}(h) > \varepsilon]$ for different ε:



Remarks

- The analysis holds for any fixed probability distribution (only the independence of successive points are needed).
- The sample size bound behave like one might expect: for greater accuracy (decrease ε) or greater confidence (decrease δ), then the algorithm requires more examples ℓ.
- It may be that the distribution of samples is concentrated in one region of the plain, creating a distorted picture of the rectangle R' with respect to the true rectangle R. However, the hypothesis is still good, since the learner is tested using the same distribution that generated the samples. That is, there may be large regions which belong to R but are not in R', however, this is not important as the learner is unlikely to be tested in this region.
- The algorithm analyzed is efficient, the sample size required is a slowly growing function of $1/\varepsilon$ (linear) and $1/\delta$ (logarithmic).
- The problem readily generalizes to n-dimensional rectangles for $n > 2$, with a bound of $2n/\varepsilon \ln(2n/\delta)$, on the sample complexity. It is not clear how to generalize the example (even in two dimensions) to other concept such as, circles, half-planes or rectangles of arbitrary orientation.

Vapnik Chervonenkis (VC) Theory

Let us return to the hypothesis h which is *consistent with its training set* S, i.e. $\mathsf{err}_S(h) = 0$. We would like to do know the upper bound δ on the probability that the error is greater than some value ε, that is

$$\Pr[\mathsf{err}_S(h) = 0 \text{ and } \mathsf{err}_{\mathcal{D}}(h) > \varepsilon] < \delta$$

Because h consistent with the sample S the probability that a point $\boldsymbol{x}_i$ has $h(\boldsymbol{x}_i) = y_i$ is at most $(1 - \varepsilon)$, and the probability of $y_i = h(\boldsymbol{x}_i)$ for all $i = 1, \ldots, \ell$ is at most $(1 - \varepsilon)^{\ell}$. Hence

$$\Pr[\mathsf{err}_S(h) = 0 \text{ and } \mathsf{err}_{\mathcal{D}}(h) > \varepsilon] \leq (1 - \varepsilon)^{\ell} \leq \exp(-\varepsilon\ell).$$

Now the probability that a single hypothesis is consistent with a sample S and has an error greater than ε is at most $|H| \exp(-\varepsilon\ell)$ by the union bound on the probability that one of several events occur, i.e.

$$\Pr[\mathsf{err}_S(h) = 0 \text{ and } \mathsf{err}_{\mathcal{D}}(h) > \varepsilon] \leq |H| \exp(-\varepsilon\ell) < \delta$$

where $|H|$ is the cardinality of the hypothesis space H.

To ensure that the rights hand side is less than δ, we set

$$\varepsilon = \varepsilon(\ell, \delta, H) = \frac{1}{\ell} \ln \frac{|H|}{\delta}$$

This shows how the complexity of the function class H measured here by its cardinality has a direct effect on the error bound.

> What do we then do when H has an infinite cardinality ($|H| \equiv \infty$, e.g.: linear machines characterized by real valued weight vectors), are there any infinite concept classes that are learnable from a finite sample?

The key idea here is that what matters is not the cardinality of H, but rather H's *expressive power*. This notion will be formalized in terms of the Vapnik-Chervonenkis dimension – a general measure of complexity for concept classes of infinite cardinality.

Growth function

Suppose the hypothesis space H is defined on the instance space $\mathcal{X}$, and $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell\}$ is a sample of size ℓ from $\mathcal{X}$.

Define $B_H(\ell)$ as the *number of classifications of* $\boldsymbol{X}$ *by* H, to be the *number of distinct vectors* of the form

$$B_H(\boldsymbol{X}) = (h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), \ldots, h(\boldsymbol{x}_\ell))$$

as h runs through all hypothesis of H. Note that $h(\boldsymbol{x}) \in \{-1, +1\}$ and so for a set of ℓ points $\boldsymbol{X}$, $B_H(\ell) \leq 2^\ell$ (the maximum number of distinct vectors of length ℓ made up of minus and plus ones).

Then the **growth function** is defined by,

$$\begin{aligned} B_H(\ell) &= \max\{\big|B_H(\boldsymbol{X})\big| : \boldsymbol{X} \in \mathcal{X}^\ell\} \\ &= \max_{\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell\} \in \mathcal{X}^\ell} |\{(h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), \ldots, h(\boldsymbol{x}_\ell)) : h \in H\}| \end{aligned}$$

> The growth function $B_H(\ell)$ can be thought of as a measure of the complexity of H: the faster this function grows the more behaviors $(B_H(\boldsymbol{X}))$ on sets of ℓ points that can be realized by H as ℓ increases.

Shattering

A set of points $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell\}$ for which the set

$$\{(h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_\ell)) : h \in H\} = \{-1, +1\}^\ell$$

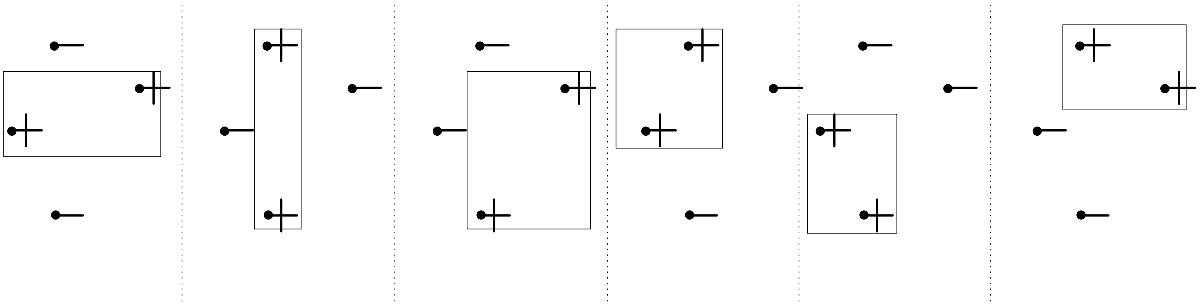
i.e. the set includes all possible ℓ-vectors, whose elements are -1 and 1, is said to be *shattered* by H. If there are sets of any size which can be shattered the growth function is equivalent to 2^ℓ for all ℓ.

Alternatively, a sample $\boldsymbol{X}$ of length ℓ is *shattered* by H, or H *shatters* $\boldsymbol{X}$, if $B_H(\ell) = 2^\ell$. That is, if H gives all possible classifications of $\boldsymbol{X}$. The samples in $\boldsymbol{X}$ must be distinct for this to happen.

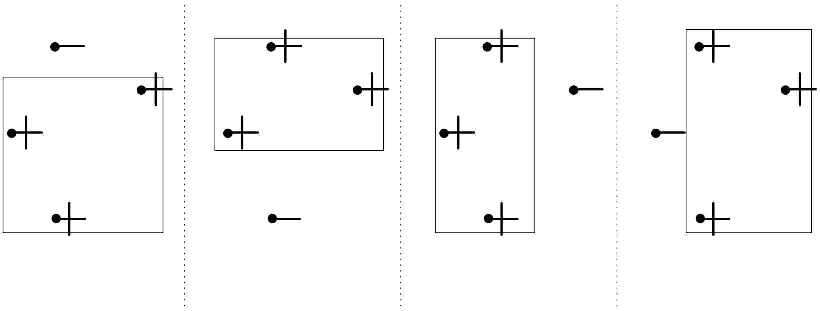
Shattering example: axis aligned rectangles

The shattering of four points by a rectangle classifier is illustrated.

1. In case all points are labeled $-$, take an empty rectangle, $(\binom{4}{0} = 1)$.
2. In the case where only one point is labeled $+$, take a rectangle enclosing this point only, there are $\binom{4}{1} = 4$ of them.
3. In the case two points are labeled $+$ and two points are $-$, take the following $\binom{4}{2} = 6$ rectangles:



4. In the case three points are labeled $+$ and one point is $-$, take the following $\binom{4}{3} = 4$ rectangles:



5. In the case all points are labeled $+$, take a rectangle enclosing all points. There is only one of them $\binom{4}{4} = 1$.

The cardinality of the hypothesis space is $|H| = \sum_{i=0}^{\ell} \binom{\ell}{i} = 16$ where $\ell = 4$.

Note that **not all** *sets of four points can be shattered! Still the existance of a single shattered set of size four is sufficient to lower bound the VC dimension.*

Vapnik-Chervonenkis Dimension

The Vapnik-Chervonenkis (VC) dimension of H is the cardinality of the largest finite set of points $\boldsymbol{X} \subseteq \mathcal{X}$ that is shattered by H. If arbitrary large finite sets are shattered, the VC dimension (VCdim) of H is infinite. Thus,

$$\text{VCdim}(H) = \max\{\ell : B_H(\ell) = 2^\ell\},$$

where we take the maximum to be the infinite if the set is unbounded. Note that the empty set is always shattered, so the VCdim is well defined.

> In general one might expect $B_H(\ell)$ to grow as rapidly as an exponential function of ℓ. *Actually, it is bounded by a polynomial in ℓ of degree $d =$ VCdim(H).* That is, depending on whether the VCdim is finite or infinite, the growth function $B_H(\ell)$ is either eventually polynomial or forever exponential.

The following simple result on *finite* hypothesis spaces is useful[4],

$$\text{VCdim}(H) \leq \log_2 |H|.$$

[4] Note that $\log_2 |H| = \ln |H| / \ln(2)$.

VCdim example: axis aligned rectangles

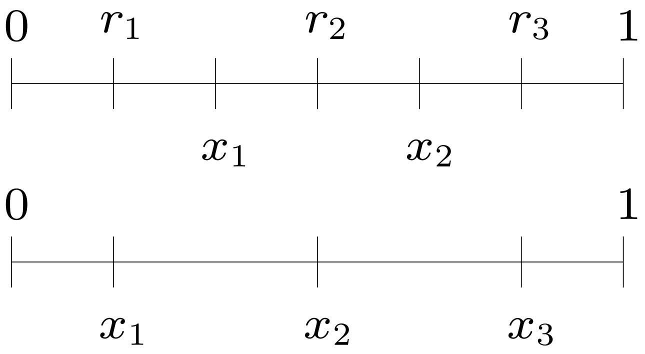
If we had five points, then at most four of the points determine the minimal rectangle that contains the whole set. Then no rectangle is consistent with the labeling that assigns these four boundary points "$+$" and assigned the remaining point (which must reside inside any rectangle covering these points) a "$-$". Therefore,

$$\text{VCdim}(\textit{axis-aligned rectangles} \in \mathbb{R}^2) = 4$$

VCdim example: intervals on the real line

The concepts are intervals on a line, where points lying on or inside the interval are positive, and the points outside the interval are negative.

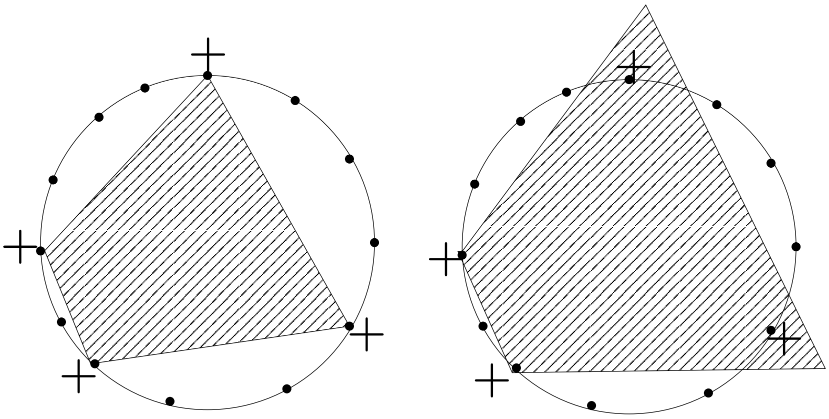
Consider the a set $\boldsymbol{X} = \{x_1, x_2\}$ a subset of X and the hypothesis $h_1 = [0, r_1]$, $h_2 = [r_1, r_2]$, $h_3 = [r_2, r_3]$, $h_4 = [0, 1]$. Let the hypothesis class be $H = \{h_1, h_2, h_3, h_4\}$. Then, $h_1 \cap S = \emptyset$, $h_2 \cap S = \{x_1\}$, $h_3 \cap S = \{x_2\}$ and $h_4 \cap S = S$. Therefore, $\boldsymbol{X}$ is shattered by H.



Consider now the set of three, $\boldsymbol{X} = \{x_1, x_2, x_3\}$. There is no concept which contains x_1 and x_3 and does not contain x_2. This, $\boldsymbol{X}$ is not shattered by H and thus VCdim$(H) = 2$.

VCdim example: where VCdim$(H) = \infty$

Any set of points on a circle in $\mathbb{R}^2$, as shown in the figure below, may be shattered by a set of *convex polygons in* $\mathbb{R}^2$. This implies that the VCdim of a set of polygons is infinity. This does no imply that you cannot learn using polygons, i.e. the converse of the VCdim theory is not true in general.



VCdim example: hyperplanes (half-spaces)

Classroom exercise.

VC lemma

If VCdim$(H) = d$ then

$$B_H(\ell) \leq \sum_{i=0}^{d} \binom{\ell}{i}$$

for any ℓ. For $\ell = d$, $\sum_{i=0}^{\ell} \binom{\ell}{i} = 2^{\ell}$. For $d < \ell$, since $0 \leq d/\ell \leq 1$, we can write:

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

and by dividing both sides by $\left(\frac{d}{\ell}\right)^d$ yields:

$$B_H(\ell) \leq \sum_{i=0}^{d} \binom{\ell}{i} \leq \left(\frac{e\ell}{d}\right)^d$$

giving polynomial growth with exponent d.

The fact that $B_H(X) = H$, implies that $|H| = B_H(|X|)$. This really does not help us with the bound:

$$\Pr[\text{err}_S(h) = 0 \text{ and } \text{err}_{\mathcal{D}}(h) > \varepsilon] \leq |H| \exp(-\varepsilon\ell)$$

since $|X|$ may also be infinite. For this reason we would like to carry out the analysis on a small random subset S instead of the entire domain X.

> The important property here is that the sample complexity depends on the VC dimension d and ε and δ, but is independent of $|H|$ and $|X|$.

VC theorem

Suppose we draw a multiset S of ℓ random examples from $\mathcal{D}$ and let A denote the event that S fails to fall in the error region r, but the probability of falling into an error region is at least ε. We want to find an upper bound for this event.

We now introduce another sample $\hat{S}$, a so called ghost sample. Since the probability of hitting the error region is at least ε, then assuming that $\ell > 2/\varepsilon$ the probability $\hat{S}$ hits the region r at least $\varepsilon\ell/2$ times is at least[5] $1/2$.

Now let B be the combined event that A occurs on the draw S and the draw $\hat{S}$ has at least $\varepsilon\ell/2$ hits in the error region. We have argued that $\Pr[B|A] \geq 1/2$ and we also have that $\Pr[B] = \Pr[B|A]\Pr[A]$, therefore

$$\Pr[A] \leq 2\Pr[B]$$

or equivalently

$$\Pr[\text{err}_S(h) = 0 \text{ and } \text{err}_{\mathcal{D}}(h) > \varepsilon] \leq$$

$$2\Pr[\text{err}_S(h) = 0 \text{ and } \text{err}_{\hat{S}}(h) > \varepsilon\ell/2]$$

where $\ell > 2/\varepsilon$. This is sometimes called the “two sample trick”.

[5] The probability $\hat{S}$ hits r at least $\varepsilon\ell/2$ times (i.e. any integer number of times greater than or equal to $\varepsilon\ell/2$) is greater then $1/2$ because if $\ell > 2/\varepsilon$ then $\lceil\varepsilon\ell/2\rceil < \varepsilon\ell$ (the expected value).

Bounding the ghost sample

From above we now have that:

$$\Pr[\text{err}_S(h) = 0 \text{ and } \text{err}_{\mathcal{D}}(h) > \varepsilon] \leq 2 \Pr[\text{err}_S(h) = 0 \text{ and } \text{err}_{\hat{S}}(h) > \varepsilon\ell/2]$$

what we need to to do now is bound

$$\Pr[\text{err}_S(h) = 0 \text{ and } \text{err}_{\hat{S}}(h) > \varepsilon\ell/2]$$

from above. First of all notice that this probability depends only on two samples of size ℓ.

1. Instead of having two samples we now draw a multiset[6] S' of size 2ℓ and randomly divide them into S and $\hat{S}$. The resulting distribution is the same since each draw from $\mathcal{D}$ is independent and identically distributed.
2. Now fix the sample S' and a error region r satisfying $|r| \geq \varepsilon\ell/2$.

   The task is now reduced to this: *we have 2ℓ balls (the multiset S'), each coloured red and blue, with exactly $m \geq \varepsilon\ell/2$ red balls (these are the instances of S' that fall in the region r). We divide the balls randomly into*

[6] A set-like object in which order is ignored, but multiplicity is explicitly significant. Therefore, multisets $\{1, 2, 3\}$ and $\{2, 1, 3\}$ are equivalent, but $\{1, 1, 2, 3\}$ and $\{1, 2, 3\}$ differ.

two groups of equal size S and $\hat{S}$, and we are interested in bounding the probability that all m of the red balls fall in $\hat{S}$ (that is, the probability that $r \cap S = \emptyset$).

3. Equivalently, we can first divide 2ℓ uncoloured balls into S and $\hat{S}$, and then randomly choose m of the balls to be marked red, and the rest blue. Then the probability that all m of the red marks fall on balls $\hat{S}$ is exactly[7] :

$$\binom{\ell}{m} / \binom{2\ell}{m} = \prod_{i=0}^{m-1} \frac{\ell - i}{2\ell - i} \leq \prod_{i=0}^{m-1} \left(\frac{1}{2}\right) = \frac{1}{2^m}$$

and since $|r| \geq \varepsilon\ell/2$ the probability of the random partitioning of S' resulting in $r \cap S = \emptyset$ is at *most*

$$2^{-\varepsilon\ell/2}$$

4. Finally, the number of possible ways this can happen is bounded by $B_H(2\ell)$, therefore, by the union of the bound,

$$\begin{aligned} \Pr[\mathsf{err}_S(h) = 0 \text{ and } \mathsf{err}_{\hat{S}}(h) > \varepsilon\ell/2] &\leq B_H(2\ell)2^{-\varepsilon\ell/2} \\ &= \left(\frac{2e\ell}{d}\right)^d 2^{-\varepsilon\ell/2} \end{aligned}$$

[7] Initially the probability of a red mark falling on a ball in $\hat{S}$ is $\ell/2\ell$, in the next try it is $(\ell-1)/(2\ell-1)$ and in the third $(\ell-2)/(2\ell-2)$, and so on.

Vapnik and Chervonenkis Theorem

Finally now we combine all the results above and obtain:

$$\begin{aligned}
&\Pr[\mathsf{err}_S(h) = 0 \text{ and } \mathsf{err}_{\mathcal{D}}(h) > \varepsilon] \\
&\qquad \leq \quad 2\Pr[\mathsf{err}_S(h) = 0 \text{ and } \mathsf{err}_{\hat{S}}(h) > \varepsilon\ell/2] \\
&\qquad \leq \quad 2\left(\frac{2e\ell}{d}\right)^d 2^{-\varepsilon\ell/2}
\end{aligned}$$

resulting in a PAC bound for any consistent hypothesis h of

$$\mathsf{err}_{\mathcal{D}}(h) \leq \varepsilon(\ell, \delta, H) = \frac{2}{\ell}\left(d \log\frac{2e\ell}{d} + \log\frac{2}{\delta}\right)$$

where $d = \mathsf{VCdim}(H) \leq \ell$ and $\ell > 2/\varepsilon$. See also (Vapnik and Chervonenkis) Theorem 4.1 page 56 in book.

Structural risk minimization (SRM)

In some cases it may not be possible to achieve find a hypothesis which is consistent with the training sample S. The VC theory can in this case be adapted to allow for a number of errors on the training set by counting the permutations which leave no more errors on the left hand side:

$$\text{err}_{\mathcal{D}}(h) \leq 2\left(\frac{\text{err}_S(h)}{\ell} + \frac{2}{\ell}\left(d \log \frac{2e\ell}{d} + \log \frac{4}{\delta}\right)\right)$$

provided $d \leq \ell$.

- **empirical risk minimization**: the above suggest that for a fixed choice of hypothesis class H one should seek to minimize the number of training errors ($\text{err}_S(h)$).
- One could apply a nested sequence of hypothesis classes:

  $$H_1 \subset H_2 \subset \ldots \subset H_i \subset \ldots \subset H_M$$

  where the VC dimensions $d_i = VCdim(H_i)$ form a non-decreasing sequence.
- The minimum training error $\text{err}_S(h_i) = k_i$ is sought for each class H_i.
- Finding the h_i for which the bound above is minimal is known as **structural risk minimization**.

Benign distributions and PAC learning

Theorem 4.3 Let H be a hypothesis space with finite VC dimension $d \geq 1$. The for any learning algorithm there exist distributions such that with probability at least δ over ℓ random samples, the error of the hypothesis h returned by the algorithm is at least

$$\max\left(\frac{d-1}{32\ell}, \frac{1}{\ell}\ln\frac{1}{\delta}\right)$$

- The VCdim($\mathfrak{L}$) of a linear learning machine is $d = n + 1$, this is the largest number of examples that can be classified in all 2^d possible classifications by different linear functions.
- In a high dimensional feature space Theorem 4.3 implies that learning is not possible, in the distribution free sense.
- The fact that SVMs can learn must therefore derive from the fact that the distribution generating the examples is not worst case as required for the lower bound of Theorem 4.3.
- The margin classifier will provide a measure of how helpful the distribution is in indentifying the target concept, i.e.

  $$\text{err}_{\mathcal{D}}(h) \leq \varepsilon(\ell, \mathfrak{L}, \delta, \gamma)$$

  and so does not involve the dimension of the feature space.

Margin-Based Bounds on Generalization

Recall the (page 11) the *(functional) margin of an example* $(\boldsymbol{x}_i, y_i)$ with respect to hypeplance $\boldsymbol{w}, b)$ to be the quantity

$$\gamma_i = y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b) = y_i f(\boldsymbol{x}_i)$$

where $f \in \mathfrak{F}$, a real-valued class of functions, and $\gamma_i > 0$ implies correct classification of $(\boldsymbol{x}_i, y_i)$.

The margin $m_S(f)$ is the minimum of the margin distributions of f w.r.t. training set S:

$$m_S(f) = \min M_S(f) = \min\{\gamma_i = y_i f(\boldsymbol{x}_i) : i = 1, \ldots, \ell\}.$$

The margin of a training set S with respect to the function class $\mathfrak{F}$ is the maximum margin over all $f \in \mathfrak{F}$.

Fat shattering dimension and γ-shattering

- The **fat shattering dimension** of a function class $\mathfrak{F}$ is a straightforward extension of the VC dimension. It is defined for real valued functions as the maximum number of points that can be γ-shattered and is denoted by $\text{fat}_{\mathfrak{F}}(\gamma)$.

- A set $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell\}$ is γ-**shattered**, is there exists some $r_i \in \mathbb{R}$ such that for all sets $y_i\{\pm 1\}$ there exists an $f \in \mathfrak{F}$ with $y_i(f(\boldsymbol{x}_i) - r_i) \geq \gamma$.

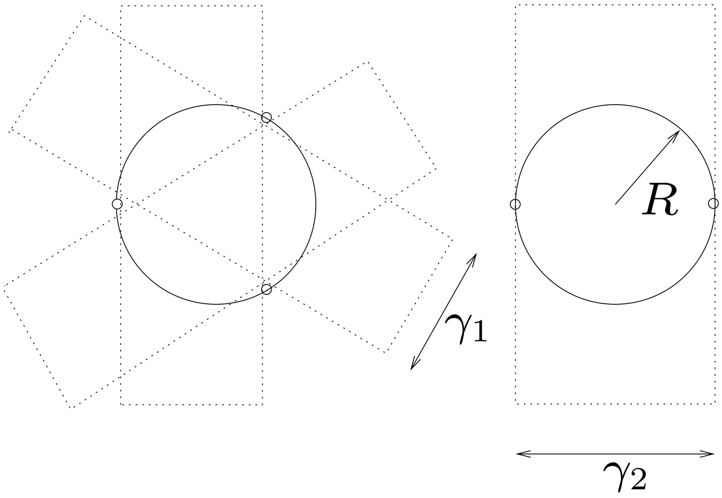
- A more restrictive definition is the *level fat shattering dimension*, where a set is γ-shattered if $y_i(f(\boldsymbol{x}_i) - r) \geq \gamma$ for a common value of r.

- The real numbers r_i can be thought of as individual thresholds for each point. This may be seem to introduce flexibility over using a single r for all points, however, in the case of linear functions it does not increase the fat shattering dimension.

Maximal margin bounds

The larger the value of γ, the smaller the size of a set that can be γ shattered. This can be illustrated by the following **example**:

> Assume that the data points are contained in a ball of radius R. Using a hyperplane with margin γ_1 it is possible to separate 3 points in all possible ways. Using a hyperplane with the larger margin γ_2, this is only possible for 2 points. Hence the VC dimension is 2 rather than 3.



A large margin of separation amounts to limiting the VC dimension!

VC dimension of large margin linear classifiers

Suppose that $\mathcal{X}$ is the ball of radius R in a Hilbert space, $\mathcal{X} = \{\boldsymbol{x} \in H : \|\boldsymbol{x}\| \leq R\}.$, and consider the set

$$\mathfrak{F} = \{\boldsymbol{x} \mapsto \big\langle \boldsymbol{w} \cdot \boldsymbol{x} \big\rangle : \|\boldsymbol{w}\| \leq 1, \boldsymbol{x} \in \mathcal{X}\}.$$

Then

$$\mathsf{fat}_{\mathfrak{F}}(\gamma) \leq \left(\frac{R}{\gamma}\right)^2.$$

There is a constant c such that for all probability distributions, with probability at least $1 - \delta$ over ℓ independently generated training examples, then the error of sgn(f) is no more than

$$\frac{c}{\ell}\bigg(\frac{R^2}{\gamma^2}\log^2 \ell + \log(1/\delta)\bigg).$$

Furthermore, with probability at least $1 - \delta$, every classifier sgn$(f) \in$ sgn$(\mathfrak{F})$ has error no more than

$$\frac{k}{\ell} + \sqrt{\frac{c}{\ell}\bigg(\frac{R^2}{\gamma^2}\log^2 \ell + \log\frac{1}{\delta}\bigg)}$$

where $k < \ell$ is the number of labelled training examples with margin less than γ. See also chapter 4.3.2 in book.

Generalization demo

We digress now to illustrate some key concepts discussed so far. In this demonstrations we cover the following:

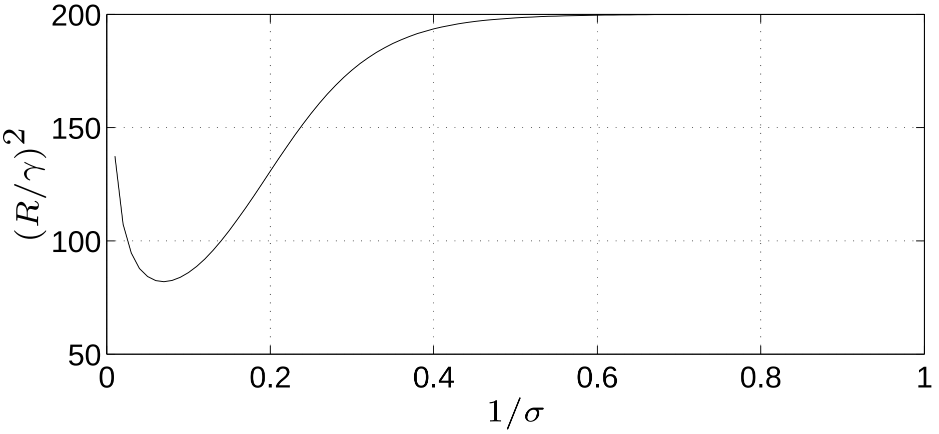
- compiling and setting up the LIBSVM package,
- adding your own kernel to LIBSVM,
- reading and writing LIBSVM files to and from MATLAB
- computing the bound on the fat shattering dimension,
- and finally how this dimension is used for model selection.

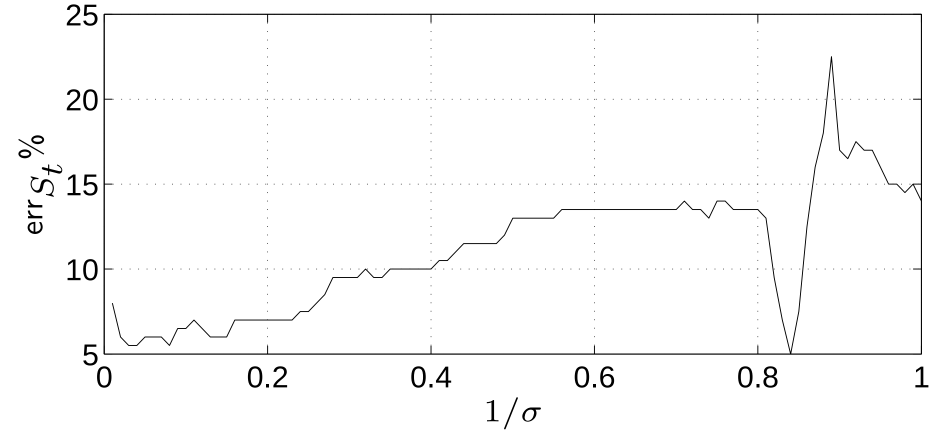
Our data is the Copenhagen chromosome set used in assignment 6 (see also lecture notes for chapter 3). One of the kernels used was the following:

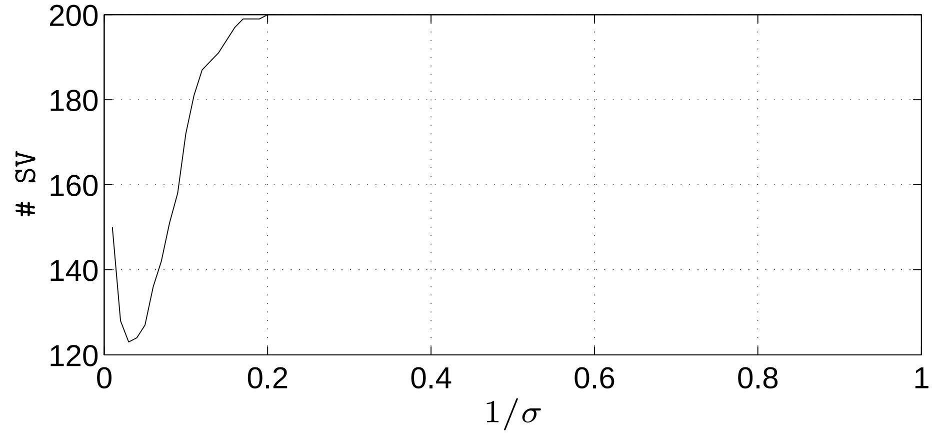
$$K(s,t) = \exp\left(- \frac{d(s,t)}{\sigma} \right)$$

where $d(s,t)$ is the string edit distance.

The question here is what should $1/\sigma$ be?


$(R/\gamma)^2$
200
150
100
50
0
0.2
0.4
0.6
0.8
1
$1/\sigma$


$\text{err}_{S_t}\%$
25
20
15
10
5
0
0.2
0.4
0.6
0.8
1
$1/\sigma$


# SV
200
180
160
140
120
0
0.2
0.4
0.6
0.8
1
$1/\sigma$

Results

`See course web page for instructions.`

Let the different hypothesis spaces H_i correspond to the different values of σ_i.

- The $1/\sigma_i$ for which the upper bound on $\text{fat}_{\mathfrak{L}}(h_i) \leq (R/\gamma)^2$ is minimal correspond approximately to the minimum error on the test set.
  Here
  $$R^2 = \max_{1 \leq i \leq \ell} K(\boldsymbol{x}_i, \boldsymbol{x}_i)$$
  and
  $$\gamma^2 = 1/\|\boldsymbol{\alpha}\|_1$$
  which is the *geometric* margin (see also p. 99 in book).

- The error on the test set $\text{err}_{S_t}(h_i)$ is noisy and may in one instance be misleading (around 0.85).

- The number of support vectors appears to be a good estimator for the generalization error, **why**?

Cross-validation

Cross-validation is another method of estimating the generalization error. It is based on the training set:

> Split the training data into p parts of equal size, say $p = 10$. Perform p independent training runs each time leaving out one of the p parts for validation. The model with the minimum average validation error is chosen.

The extreme case is the *leave-one-out* cross-validation error:

$$LOO = \frac{1}{\ell} \sum_{i=1}^{\ell} \big| y_i - h_i(\boldsymbol{x}_i) \big|$$

where h_i is the hypothesis found using the training set for which the sample $(y_i, \boldsymbol{x}_i)$ was left out.

Note that if an example is left out which is not a support vector then the LOO will not change! In fact the LOO error is bound[8] by the number of support vectors, i.e.

$$\mathsf{E}\big(|y_i - h_i(\boldsymbol{x}_i)|\big) < \frac{\#\ \mathsf{SV}}{\ell}$$

[8] The use of an expected generalization bound gives no guarantee about the variance and hence its reliability.

Remarks

- Radius-margin bound (Vapnik 1998):

$$\mathsf{E}\big(|y_i - h_i(\boldsymbol{x}_i)|\big) < \frac{1}{\ell}\left(\frac{R}{\gamma}\right)^2$$

- *Littlestone and Warmuth*[9]: According to $\mathcal{D}$ there are at most #SV examples that map via the maximum-margin hyperplane to a hypothesis that is both consistent with all ℓ examples and has error larger than ε is at most

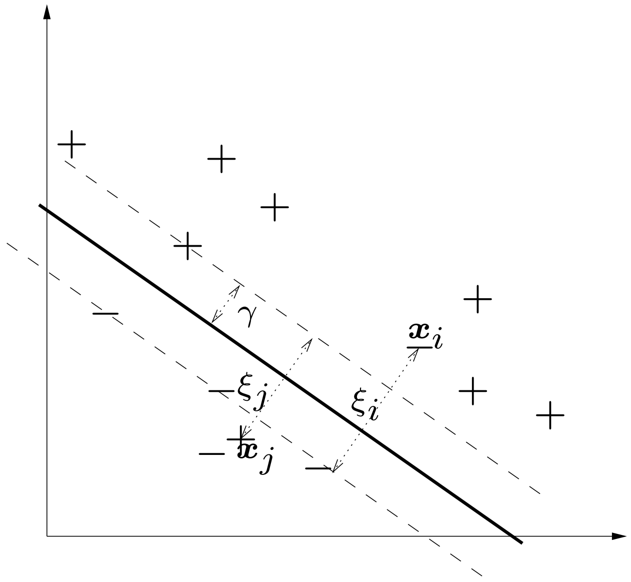
$$\sum_{i=0}^{\#\mathsf{SV}} \binom{\ell}{i} (1-\varepsilon)^{\ell-i}$$

  This implies that the generalization error of a maximal margin hyperplane with #SV support vectors among a sample of size ℓ can with confidence $1-\delta$ be bounded by

$$\frac{1}{\ell - \#\mathsf{SV}}\left(\#\mathsf{SV}\log\frac{e\ell}{\#\mathsf{SV}} + \log\frac{\ell}{\delta}\right).$$

[9] See also chapter 4.4 in book, here #SV playes the role of the VC dimension in the bound (see also Shawe-Taylor and Bartlett, 1998, *Structural Rist Minimization over Data-Dependent Hierarchies*).

Margin slack vector



The slack margin variable measures the amount of "non-separability" of the sample. Recall the *margin slack variable* of an example $(\boldsymbol{x}_i, y_i)$ with respect to the hyperplane $(\boldsymbol{w}, b)$ and target margin γ:

$$\xi_i = \max\Big(0 \ \ , \ \ \gamma - y_i(\langle\, \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b)\Big).$$

If $\xi_i > \gamma$, then $\boldsymbol{x}_i$ is misclassified by $(\boldsymbol{w}, b)$. The *margin slack vector* is simply a vector containing all the ℓ margin slack variables:

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_\ell).$$

Soft margin bounds

Theorem 4.2.4 *Consider thresholding real-valued linear functions $\mathfrak{L}$ with unit weight vectors on an inner product space $\mathcal{X}$ and fix $\gamma \in \mathbb{R}^+$. There is a constant c, such that for any probability distribution $\mathcal{D}$ on $\mathcal{X} \times \{-1, 1\}$ with support in a ball of radius R around the origin, with probability $1-\delta$ over ℓ random examples S, any hypothesis $f \in \mathfrak{L}$ has error no more than*

$$err_{\mathcal{D}}(f) \leq \frac{c}{\ell}\left(\frac{R^2 + \|\boldsymbol{\xi}\|_1^2 \log(1/\gamma)}{\gamma^2} \log^2 \ell + \log \frac{1}{\delta}\right),$$

where $\boldsymbol{\xi} = \boldsymbol{\xi}(f, S, \gamma)$ is the margin slack vector with respect to f and γ.

Remarks:

- The generalization error bound takes into account the amount of points failing to meet the target margin γ.
- The bound does not require the data to be linearly separable, just that the norm of the slack variable vector be minimized.
- Note that minimizing $\|\boldsymbol{\xi}\|$ does **not** correspond to *empirical risk minimization*, since this does not imply minimizing the number of misclassifications.
- Optimizing the norms of the margin slack vector has a diffuse effect on the margin, for this reason it is referred to as a **soft margin**.

Regression and generalization

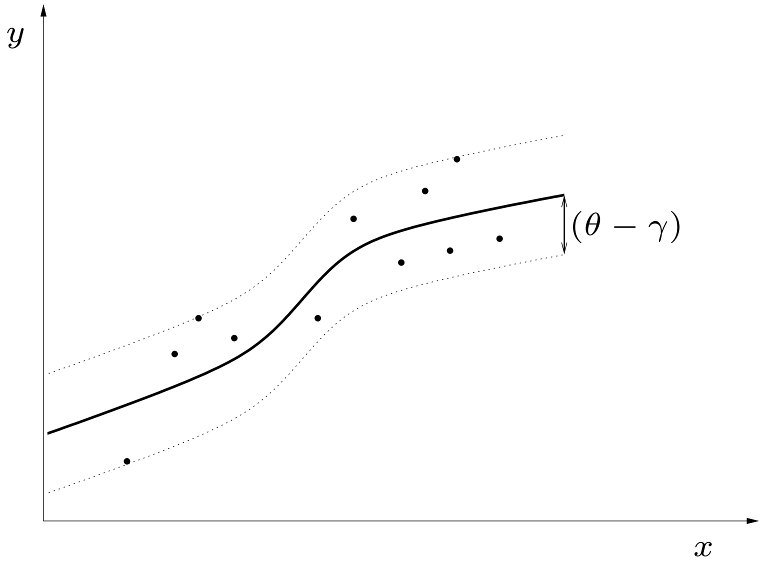
Regression:

- No binary output, instead have a *residual*, the real difference between the target and hypothesis.
- Small residuals may be inevitable and we wish to avoid large residuals.

Generalization:

- *In order to make use of the* **dimension free bounds** *one must allow a margin in the regression accuracy that corresponds to the margin of the classifier.*
- This is accomplished by introducing a target accuracy θ. Any test point outside the band $\pm\theta$ is a mistake.
- Recall the ϵ-insensitive regression (assignment 5), this ϵ is equivalent to θ.
- Now introduce the margin $\gamma \leq \theta$.

Consistent within $\theta - \gamma$ on training set



Theorem 4.26 *Consider performing regression estimation with linear function $\mathfrak{L}$ with unit weight vector on an inner product space $\mathcal{X}$ and fix $\gamma \leq \theta \in \mathbb{R}^+$. For any probability distribution $\mathcal{D}$ on $\mathcal{X} \times \mathbb{R}$ with support in a ball of radius R around the origin, with probability $1 - \delta$ over ℓ random examples S, any hypothesis $f \in \mathfrak{L}$, whose output is within $\theta - \gamma$ of the training value for all of the training set S, has residual greater than θ on a randomly drawn test point with probability at most*

$$err_{\mathcal{D}}(f) \leq \frac{2}{\ell}\left(\frac{64R^2}{\gamma^2}\log\frac{e\ell\gamma}{4R}\log\frac{128\ell R^2}{\gamma^2} + \log\frac{4}{\delta}\right)$$

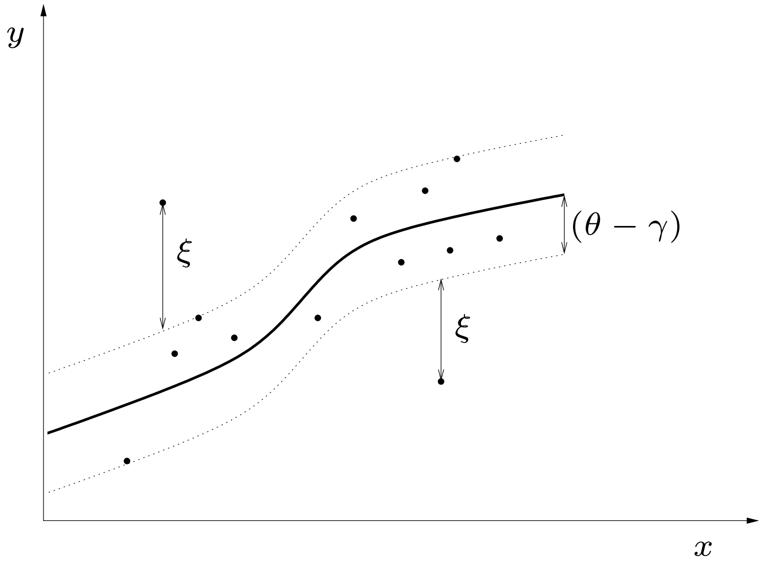
provided $\ell > 2/\varepsilon$ and $64R^2/\gamma^2 < \ell$.

Margin slack variables

The *margin slack variable* of an example $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathbb{R}$ with respect to a function $f \in \mathfrak{F}$, target accuracy θ and loss margin γ is defined as follows:

$$\xi_i = \max\big(0, |y_i - f(\boldsymbol{x}_i)| - (\theta - \gamma)\big)$$

Note that $\xi_i > \gamma$ implies an error on $(\boldsymbol{x}_i, y_i)$ of more than θ.



Theorem 4.30 *Consider performing regression with linear functions $\mathfrak{L}$ on an inner product space $\mathcal{X}$ and fix $\gamma \leq \theta \in \mathbb{R}^+$. There is a constant c, such that for any probability distribution $\mathcal{D}$ on $\mathcal{X} \times \mathbb{R}$ with support in a ball of radius R around the origin, with probability $1 - \delta$ over ℓ random examples S, the probability that a hypothesis $\boldsymbol{w} \in \mathfrak{L}$ has output more than θ away from its true value is bounded by*

$$\mathit{err}_{\mathcal{D}}(f) \leq \frac{c}{\ell}\left(\frac{\|\boldsymbol{w}\|_2^2 R^2 + \|\boldsymbol{\xi}\|_1^2 \log(1/\gamma)}{\gamma^2} \log^2 \ell + \log \frac{1}{\delta} \right)$$

where $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{w}, S, \theta, \gamma)$ is the margin slack vector with respect to $\boldsymbol{w}, \theta$ and γ.

Note that we do not fix the norm of the weight vector.

The above theorem applies to the 1-norm regression see p. 73-74 in the book a similar theorem for the 2-norm regression is given.

If we optimize the 1-norm bound the resulting regressor takes less account of points that have large residuals and so can handle outliers better than by optimizing the 2-norm bound[10].

[10] Replace $\|\boldsymbol{\xi}\|_1^2 \log(1/\gamma)$ by $\|\boldsymbol{\xi}\|_2^2$ in the bound above.