The VC dimension

Nuno Vasconcelos

ECE Department, UCSD
Margins and VC dimension

- we have been talking about techniques that achieve good generalization by maximizing the margin
- it turns out that the quantity of interest is the so-called VC dimension of the family of functions implemented by the learning machine
- the margin allows us to control this dimension and this is what makes it important
- in the next two lectures we are going to formalize this a little better
- disclaimer: the topic is a bit technical, we will only do a superficial review of the main results
- the idea is to get the big picture
Loss functions and Risk

- goal of the learning machine: to find the set of parameters that minimizes the risk (expected value of the loss)

\[ R(f) = E_{X,Y} \{ L[y, f(x)] \} \]

\[ = \int P_{X,Y}(x, y) L[y, f(x)] dx dy \]

- in practice it is impossible to evaluate the risk, because we do not know what \( P_{X,Y}(x,y) \) is.

- all we have is a training set

\[ D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \]

- we estimate the risk by the empirical risk on this training set

\[ R_{emp}(\alpha) = \frac{1}{n} \sum_{i=1}^{n} L[y_i, f(x_i)] \]
Empirical risk minimization

- the ERM principle recommends choosing the function $f$ that minimizes this empirical risk.

- we have already seen that this is a bad idea when $\mathcal{F}$, the set of possible $f$, is unconstrained.

- here is an $f^*$ that will always achieve zero empirical risk:

$$f^*(x) = \begin{cases} y_i, & \text{if } x = x_i \\ 0, & \text{otherwise} \end{cases}$$

- clearly has zero error, but:
  - on a different sample it is unlikely that we will get the same $x_i$.
  - outside the training set it always says “0”, which is clearly bad.

- the point is that we have to constrain $\mathcal{F}$ to have meaningful generalization.
The law of large numbers

let’s start by trying to understand the relationship between risk and empirical risk

since we are dealing with classification we use the loss

$$L[y_i, f(x_i)] = |y_i - f(x_i)|$$

if we have a sequence of iid errors $\xi_i = |f(x_i) - y_i|$ the convergence of the empirical risk

$$\frac{1}{n} \sum_{i} \xi_i$$

to the risk $E[\xi]$ is determined by the law of large numbers

there are many variations, we will do here the so-called Hoeffding bound

for this we need some intermediate results
The Markov inequality

- The basic result is the Markov inequality

**Markov:** let $\xi$ be a non-negative r.v. with distribution $P_\xi(\xi)$. Then for all $\lambda > 0$

$$P(\xi \geq \lambda E[\xi]) \leq \frac{1}{\lambda}$$

**Proof:**

$$E[\xi] = \int_0^\infty \xi P(\xi) \, d\xi \geq \int_{\lambda E[\xi]}^\infty \xi P(\xi) \, d\xi \quad (\xi > \lambda E[\xi])$$

$$\geq \lambda E[\xi] \int_{\lambda E[\xi]}^\infty P(\xi) \, d\xi = \lambda E[\xi] P(\xi \geq \lambda E[\xi])$$
The law of large numbers

we will also use the following

Lemma: let \( x \) be a r.v. with distribution \( E[x] = 0, \ a \leq x \leq b \). Then for \( s > 0 \)

\[
E[e^{sx}] \leq e^{\frac{s^2(b-a)^2}{8}}
\]

Proof:

- by convexity of the exponential

\[
e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa} = \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} + \frac{x}{b-a}(e^{sb} - e^{sa})
\]

- and, from \( E[x]=0 \)

\[
E[e^{sx}] \leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} = (1-p)e^{sa} + pe^{sb} \quad \left( p = -\frac{a}{b-a} \right)
\]
The law of large numbers

- \( E[e^{sx}] \leq (1 - p)e^{sa} + pe^{sb} = \left(1 - p + pe^{s(b-a)}\right)e^{sa} \)
  \(= \left(1 - p + pe^{s(b-a)}\right)e^{-ps(b-a)} = e^{\phi(u)} \)

with \( u = s(b-a) \) and \( \phi(u) = -pu + \log(1-p+pe^u) \)

- Taking derivatives

\[
\frac{\partial \phi}{\partial u} = -p + \frac{pe^u}{1 - p + pe^u} = -p + \frac{p}{(1 - p)e^{-u} + p} , \quad \frac{\partial \phi}{\partial u}(0) = 0
\]

and

\[
\frac{\partial^2 \phi}{\partial u^2} = \frac{p(1 - p)e^{-u}}{(p + (1 - p)e^{-u})^2} \leq \frac{1}{4}
\]

- By Taylor expansion with reminder, there is a \( \phi \) such that

\[
\phi(u) = \phi(0) + u \frac{\partial \phi}{\partial u}(0) + \frac{u^2}{2} \frac{\partial^2 \phi}{\partial u^2} \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8} \]

\[\blacksquare\]
The law of large numbers

we are now ready to show Hoeffding’s result

Theorem: if $x_i$ are iid such that $x_i - E[x_i] \in [a, b]$ and $S_n = \frac{1}{n} \sum_{i} x_i$ for all $\varepsilon > 0$ we have

$$P\left\{ \left| S_n - E[S_n] \right| \geq \varepsilon \right\} \leq 2e^{-\frac{n\varepsilon^2}{(b-a)^2}}$$

Proof:

• From Markov, for non-negative $x$, $P(x \geq \varepsilon) \leq E[x]/\varepsilon$
• hence, for all $s > 0$ and $x > 0$,

$$P(x \geq \varepsilon) = P(e^{sx} > e^{s\varepsilon}) \leq \frac{E[e^{sx}]}{e^{s\varepsilon}}$$
• and

$$P\left( S_n - E[S_n] \geq \varepsilon \right) \leq e^{-s\varepsilon} E[e^{s|S_n - E[S_n]|}]$$
The law of large numbers

\[ P\left(\left|S_n - E[S_n]\right| \geq \varepsilon\right) \leq e^{-s\varepsilon} E\left[e^{s|S_n - E[S_n]|}\right] \]

\[ = e^{-s\varepsilon} E\left[\exp\left\{\frac{s}{n} \sum_{i} (x_i - E[x_i])\right\}\right] \]

\[ \leq e^{-s\varepsilon} \prod_{i} E\left[\exp\left\{\frac{s}{n} |x_i - E[x_i]|\right\}\right] \] (\(x_i\text{ iid}\))

\[ \leq e^{-s\varepsilon} \prod_{i} E\left[\exp\left\{\frac{s^2}{n^2} \frac{(b - a)^2}{8}\right\}\right] \] (by Lemma)

\[ = e^{-s\varepsilon} \exp\left\{\frac{s^2}{n} \frac{(b - a)^2}{8}\right\} = 2e^{-\frac{2\varepsilon^2 n}{(b-a)^2}} \]

• choosing \(s = \frac{4\varepsilon n}{(b-a)^2}\)
Sidenote

- there are many variations on the law of large numbers, this is only one
- they all have the flavor that
\[ P \left( \left| S_n - E[S_n] \right| \geq \varepsilon \right) \leq O(e^{-n}) \]
- this is an amazing result when you think of it
  - estimating expectations by empirical means converges exponentially fast
- note that this is not an abstract theoretical result
  - it is the reason why it is worth studying statistics
  - why would we want to compute statistics if they did not converge to the true quantities?
Empirical risk vs risk

noting that

\[ R_{emp}[f] = \frac{1}{n} \sum_{i} \xi_i = S_n \quad R[f] = E[\xi] = E[S_n] \]

the theorem

\[ P\left( \left| S_n - E[S_n] \right| \geq \varepsilon \right) \leq e^{\frac{-2\varepsilon^2 n}{(b-a)^2}} \]

seems to indicate that

\[ P\left( \left| R_{emp}[f] - R[f] \right| \geq \varepsilon \right) \leq e^{-\varepsilon^2 n} \]

i.e. the empirical risk converges to the risk exponentially fast

this is the best that one could hope for, there seems to be no reason to use anything other than ERM
Empirical risk vs risk

- since we know that ERM is not that good, something must be wrong
- the problem is that the bounds assume independent errors $\xi_i$
- since we are choosing $f$ (by ERM) so that the mean of $\xi_i$ is as small as possible, this is not the case
- we need to look for alternative ways to understand the relationship between the two risks
- for them to be equivalent we need the ERM solution $f^*$ to converge to the lowest value of $R[f]$
- this turns out not to be possible unless we restrict $F$
Convergence

- Consider the values of the empirical risk and the risk over $F$
- They are something like this
- $f^{opt}$ minimizes the risk, $f^*$ the empirical risk for a particular sample
- Clearly

$$R[f] - R[f^{opt}] \geq 0, \quad \forall f \in \mathcal{F}$$
$$R_{emp}[f] - R_{emp}[f^*] \geq 0, \quad \forall f \in \mathcal{F}$$

- And

$$R[f^*] - R[f^{opt}] \geq 0,$$
$$R_{emp}[f^{opt}] - R_{emp}[f^*] \geq 0$$
Convergence

- \( R[f^*] - R[f^{opt}] \geq 0 \),
- or

\[
0 \leq R[f^*] - R[f^{opt}] + R_{emp}[f^{opt}] - R_{emp}[f^{*}]
= R[f^*] - R_{emp}[f^{*}]
+ R_{emp}[f^{opt}] - R[f^{opt}]
\leq \sup_{f \in \mathcal{F}} \left(R[f] - R_{emp}[f]\right)
+ R_{emp}[f^{opt}] - R[f^{opt}]
\]

- note that, because \( f^{opt} \) is a fixed function, independent of the sample, we can use Hoeffding on the last term.
Convergence

which leads to
\[ R_{emp}[f^{opt}] - R[f^{opt}] \to 0 \quad \text{as} \quad n \to \infty \]

hence, if
\[ \sup_{f^* \in \mathcal{F}} (R[f^*] - R_{emp}[f^*]) \to 0 \quad \text{as} \quad n \to \infty \] (*)&

then
\[ R[f^*] - R[f^{opt}] \to 0 \] (**)
\[ R_{emp}[f^{opt}] - R_{emp}[f^*] \to 0 \]

it turns out that (*) is also a sufficient condition for (**)
Consistency of ERM

**Theorem:** (VC) The condition

\[
\lim_{n \to \infty} \mathbb{P} \left[ \sup_{f \in \mathcal{F}} \left( R(f) - R_{\text{emp}}(f) \right) > \varepsilon \right] = 0
\]

is necessary and sufficient for consistency of ERM

- This shows that consistency depends on the class of functions \( \mathcal{F} \), but is not terribly useful in practice.
- We next look at properties of \( \mathcal{F} \) that ensure convergence.
- The first thing to do is to try to bound this quantity.
VC Bounds

- the probability

\[
P\left[ \sup_{f \in \mathcal{F}} (R[f] - R_{\text{emp}}[f]) > \varepsilon \right]
\]

is easy to bound when \( \mathcal{F} \) is finite

- let \( \mathcal{F} \) be the set \( \mathcal{F} = \{f_1, \ldots, f_M\} \) and

\[
C^i_{\varepsilon} = \left\{ (x_1, y_1), \ldots, (x_n, y_n) \mid (R[f_i] - R_{\text{emp}}[f_i]) > \varepsilon \right\}
\]

the set of samples for which the risks obtained with the \( i^{\text{th}} \) function differ by more than \( \varepsilon \)

- then, if \( M=2 \),

\[
P\left[ \sup_{f \in \mathcal{F}} (R[f] - R_{\text{emp}}[f]) > \varepsilon \right] = P\left( C^1_{\varepsilon} \cup C^2_{\varepsilon} \right)
\]

\[
= P\left( C^1_{\varepsilon} \right) + P\left( C^2_{\varepsilon} \right) - P\left( C^1_{\varepsilon} \cap C^1_{\varepsilon} \right) \leq P\left( C^1_{\varepsilon} \right) + P\left( C^2_{\varepsilon} \right)
\]
VC Bounds

- in general, 
  \[
  P \left[ \sup_{f \in \mathcal{F}} \left( R[f] - R_{\text{emp}}[f] \right) > \varepsilon \right] \leq \sum_{i=1}^{M} P(C^i_\varepsilon)
  \]

- this is called the union bound

- recalling that
  \[
  C^i_\varepsilon = \left\{ (x_1, y_1), \ldots, (x_n, y_n) \mid \left( R[f_i] - R_{\text{emp}}[f_i] \right) > \varepsilon \right\}
  \]

  and noting that the \( f_i \) are fixed, from which the errors are independent, we can now

  - just apply the LLN to each of the \( P(C^i_\varepsilon) \)
  - each of them is bounded by \( O(e^{-n\varepsilon}) \)
  - the overall bound is \( O(Me^{-n\varepsilon}) \) and convergence exponentially fast
VC Bounds

- the problem is the case when $F$ contains an infinite number of functions.
- In this case the method of proof will fail.
- one of the main VC results is the solution to this
- they showed that
  - the probability of the empirical risk on a sample of $n$ points differing from the risk by more that $\varepsilon$ can be bounded by
  - twice the probability that it differs from the empirical risk on a second sample of size $2n$ by more than $\varepsilon/2$
VC Bounds

**Theorem:** for \( n \epsilon^2 > 2 \)

\[
P\left[ \sup_{f \in \mathcal{F}}(R[f] - R_{\text{emp}}[f]) > \epsilon \right] \leq 2P\left[ \sup_{f \in \mathcal{F}}(R_{\text{emp}}[f] - R'_{\text{emp}}[f]) > \epsilon / 2 \right]
\]

where

- the 1st \( P \) refers to sample of size \( n \) and the 2nd to that of size \( 2n \).
- in the latter case, \( R_{\text{emp}} \) measures the loss on first half and \( R'_{\text{emp}} \) the loss on the second half

**this is intuitive:**

- if the \( R_{\text{emp}} \)s on two independent \( n \)-samples are close then they should also be close to the true error rate

**the practical significance is that**

- this makes \( \mathcal{F} \) effectively finite
- when we restrict the functions to \( 2n \) points there are at most \( 2^{2n} \) different elements in the set
VC dimension

- graphically

Since there are $2n$ points where the functions are either $+1$ or $-1$, the total number of distinct elements is at most $2^{2n}$.

In practice, it can, of course, be smaller.

The VC dimension is determined by this number.
VC dimension

- to formalize this, we denote the 2n point sample by
  \[ Z_{2n} = \{(x_1, y_1), \ldots, (x_{2n}, y_{2n})\} \]

- and the cardinality of the set of distinct functions by
  \[ N(\exists, Z_{2n}) \]

- the maximum cardinality over all possible samples of size 2n is the shattering coefficient (or covering number) of \( F \)
  \[ N(\exists, 2n) \]

- it is a measure of the complexity of \( F \), the number of ways in which it can separate the two classes
VC dimension

- if $N(F, n) = 2^n$, all possible separations can be implemented by functions in $F$
- in this case the functions in $F$ are said to shatter $n$ points

**note:**
- this means that there are $n$ points that can be separated in all possible ways
- does not mean that this applies to all sets of $n$ points

**example:**
- let $F$ be the set of lines in $R^2$. For 3 points in general position
  - the set of lines shatters three points on $R^2$!
VC dimension

- example:
  - find a set of four points that is shattered (that is can be separated in all possible ways) by a line in $R^2$
  - the following configuration is never possible (xor)

  ![XOR configuration](image)

  - the set of lines does not shatter four points on $R^2$!

- this brings us to the VC dimension

  $$VC(\mathcal{F}) = \max \# \text{ of points that can be shattered by functions } \in \mathcal{F}$$

- example
  - the VC dimension of the set of hyperplanes in $R^d$ is $d+1$
VC dimension

- why is the VC dimension important?
  - this will be clear next class, where we will show that

\[
R[f] \leq R_{emp}[f] + \phi[VC(F)]
\]

- by controlling the VC dimension
  - we upper bound difference between risk and empirical risk
  - effectively control the generalization ability of the learning machine

- and this is really what structural risk minimization is about

- but before we get into that, I will tie all of this to the margin, which is the quantity that we can control in practice

- there are various results relating margin and VC dimension, we next go over a simple one
The role of the margin

**Theorem:** consider hyperplanes of the form \( w^T x = 0 \), where \( w \) is normalized wrt a sample \( D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \) in the usual way

\[
\min_i |w^T x_i| = 1
\]

Then, the set of functions

\[
\mathcal{S} = \{\text{sgn}(w^T x)\}
\]

deﬁned on \( X^* = \{x_1, \ldots, x_n\} \) and satisfying

\[
\|w\| \leq \lambda
\]

has VC dimension such that

\[
\text{VC}(\mathcal{S}) \leq R^2 \lambda^2
\]

where \( R \) is the radius of the smallest sphere centered at the origin and containing \( X^* \).
The role of the margin

Proof:

- we need to show that the max number of points shattered by normalized hyperplanes with $||w|| \leq \lambda$ is upper bounded by $R^2\lambda^2$
- assume $\{x_1, ..., x_r\}$ are shattered by normalized hyperplanes of $||w|| \leq \lambda$
- then, for all $y_1, ..., y_r$ in $\{1, -1\}$, there is a $w$ s.t. $||w|| \leq \lambda$ and
  \[
y_i w^T x_i \geq 1, \forall i
  \]
- summing over all $i$
  \[
w^T \sum_i y_i x_i \geq r
  \]
- and by Cauchy-Schwarz
  \[
r \leq w^T \sum_i y_i x_i \leq ||w|| \left\| \sum_i y_i x_i \right\| \leq \lambda \left\| \sum_i y_i x_i \right\| \iff \frac{r}{\lambda} \leq \left\| \sum_i y_i x_i \right\| (\ast)
  \]
The role of the margin

- next, consider that the $y_i$ are iid r.v.s uniformly distributed on \{$1,-1$\}

$$E \left[ \left\| \sum_i y_i x_i \right\|^2 \right] = E \left[ \left( \sum_i y_i x_i \right)^T \left( \sum_j y_j x_j \right) \right]$$

$$= \sum_i E \left[ (y_i x_i)^T \left( \sum_j y_j x_j \right) \right]$$

$$= \sum_i E \left[ (y_i x_i)^T \left( \sum_{j \neq i} y_j x_j + y_i x_i \right) \right]$$

$$= \sum_i E \left[ \sum_{j \neq i} y_i y_j x_i^T x_j + \left\| y_i x_i \right\|^2 \right]$$

$$= \sum_i \left[ \sum_{j \neq i} E[y_i] E[y_j] x_i^T x_j + E \left[ \left\| y_i x_i \right\|^2 \right] \right]$$

$$= \sum_i E \left[ y_i^2 \right] \left\| x_i \right\|^2 = \sum_i \left\| x_i \right\|^2 \leq rR^2$$
The role of the margin

• hence

\[ E \left[ \left\| \sum y_i x_i \right\|^2 \right] \leq rR^2 \]

• which means that there must be at least one set of labels for which this inequality holds, i.e. there is at least one set of labels such that

\[ \left\| \sum y_i x_i \right\|^2 \leq rR^2 \]

• since (*) holds for this set, we have

\[ \left( \frac{r}{\lambda} \right)^2 \leq \left\| \sum y_i x_i \right\|^2 \leq rR^2 \quad \iff \quad r \leq R^2 \lambda^2 \]

• since this holds for any \( r \), it will also for the max number of points that can be shattered, i.e. the VC dimension and

\[ VC(\mathcal{F}) \leq R^2 \lambda^2 \]
The role of the margin

- notes:
  - there are extensions to the case where \( b \neq 0 \) but they are a bit more complicated
  - the theorem basically says that if
    \[
    \|w\| \leq \lambda
    \]
    then
    \[
    VC(\mathcal{F}) \leq R^2 \lambda^2
    \]
  - when we maximize the margin (minimize \( \|w\| \)) we are decreasing the value of \( \lambda \) and, therefore,
  - decreasing the upper bound on the VC dimension
  - note that, as long as \( \|w\| \) is finite, the VC dimension is finite even if the dimension of the space is infinite
  - since \( \gamma = 1/\|w\| \), this is always true if the margin is strictly greater than zero
Any questions?