Lecture 14: VC dimension

- Some examples
- VC dimension
  - Definition
  - Examples for some classes of algorithms studied so far
- Error bounds using VC dimension
- Structural risk minimization
- PAC-learning

Recall: Learning scenario

- We assume that data is drawn iid from a given, unknown probability distribution
- Hypotheses have a *true error*, which is the expected error when data is drawn from the distribution
- But we can only measure the *training error* over the data points that we have
- Many learning algorithms attempt to minimize the training error - a process known as *empirical risk minimization*

Recall: Bounding the true error

- Using the union bound and concentration inequalities (e.g., Hoeffding, Chernoff) we can bound the true error of the hypothesis with the smallest training error

$$e(h_{emp}) \leq \left(\min_{h \in H} e(h)\right) + 2\sqrt{\frac{1}{2m} \log \frac{2|H|}{\delta}}$$

  The first term corresponds roughly to "bias" and the second term to "variance"
- The number of data points needed so that the training error is within ϵ of the true error, with probability at least $1 - \delta$, for a finite hypothesis space, is:

$$m \geq \frac{1}{2\epsilon^2} \log \frac{2|H|}{\delta}$$

Example: Learning an interval on the real line

- "Treatment plant is ok iff Temperature $\leq a$" for some unknown $a \in [0, 100]$
- Consider the hypothesis set:

$$H = \{[0, a] | a \in [0, 100]\}$$

- Simple learning algorithm: Observe m samples, and return $[0, b]$, where b is the largest positive example seen
- Clearly the processing time per example is polynomial. But how many examples do we need to find a good approximation of the true hypothesis?
- Our previous result is useless, since the hypothesis class is infinite.

Sample complexity of learning an interval

- Let a correspond to the true concept and let $c < a$ be a real value s.t. $[c, a]$ has probability ϵ.
- If we see an example in $[c, a]$, then our algorithm succeeds in having true error smaller than ϵ
- What is the probability of seeing m iid examples *outside* of $[c, a]$?

$$P(\text{failure}) = (1 - \epsilon)^m$$

- If we want

$$P(\text{failure}) < \delta \Longrightarrow (1 - \epsilon)^m < \delta$$

Example continued

- Fact:

$$(1 - \epsilon)^m \leq e^{-\epsilon m} \text{ (you can check that this is true)}$$

- Hence, it is sufficient to have

$$(1 - \epsilon)^m \leq e^{-\epsilon m} < \delta$$

- Using this fact, we get:

$$m \geq \frac{1}{\epsilon} \log \frac{1}{\delta}$$

- You can check empirically that this is a fairly tight bound.

Why do we need so few samples?

- Our hypothesis space is simple - there is only one parameter to estimate!
- In other words, there is one "degree of freedom"
- As a result, every data sample gives information about LOTS of hypothesis!
- What if there are more "degrees of freedom"?

Example: Learning two-sided intervals

- Suppose the target concept and hypothesis class are positive inside $[a, b]$.
- Our guess interval is $[\min_{(x,+)} x, \max_{(x,+)} x]$
- We can make errors on either side of the interval, if we get no example within ϵ of the true value.
- The probability of an example outside of an ϵ-size interval is $1 - \epsilon$
- The probability of m examples outside of it is $(1 - \epsilon)^m$
- The probability this happens on either side is $\leq 2(1 - \epsilon)^m \leq 2e^{-\epsilon m}$, and we want this to be $< \delta$

Example continued

- If we extract the number of samples we get:

$$m \geq \frac{1}{\epsilon} \ln \frac{2}{\delta}$$

- Compare this with the bound in the finite case:

$$m \geq \frac{1}{2\epsilon^2} \log \frac{2|H|}{\delta}$$

- But for us, $|H| = \infty$!
- We need a way to characterize the "complexity" of infinite-dimensional classes of hypotheses

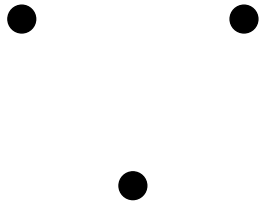
Shattering a set of instances

Definition: A **dichotomy** of a set S is a partition of S into two disjoint subsets.

Definition: A set of instances D is **shattered** by hypothesis space H if and only if *for every dichotomy of* D there exists some hypothesis in H consistent with this dichotomy.

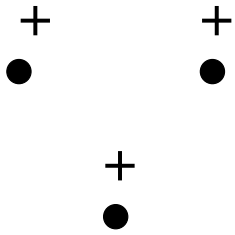
Example: Three instances

Can these three points be shattered by the hypothesis space consisting of a set of circles?



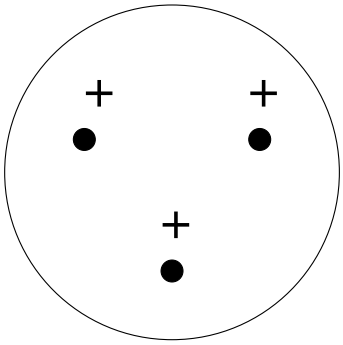
Example: Three instances

Can these three points be shattered by the hypothesis space consisting of a set of circles?



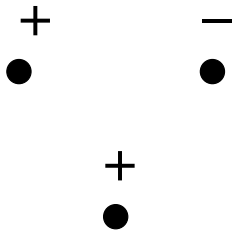
Example: Three instances

Can these three points be shattered by the hypothesis space consisting of a set of circles?



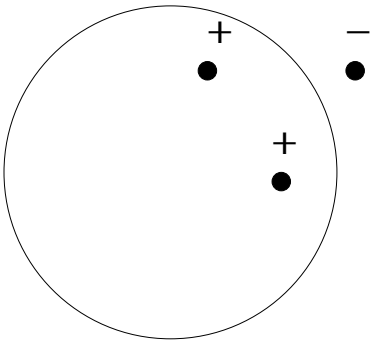
Example: Three instances

Can these three points be shattered by the hypothesis space consisting of a set of circles?



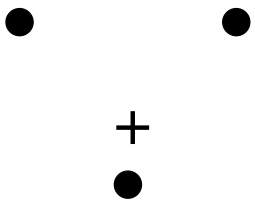
Example: Three instances

Can these three points be shattered by the hypothesis space consisting of a set of circles?



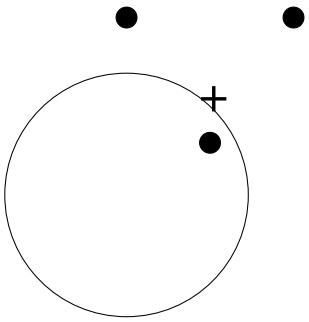
Example: Three instances

Can three points be shattered by the hypothesis space consisting of a set of circles?



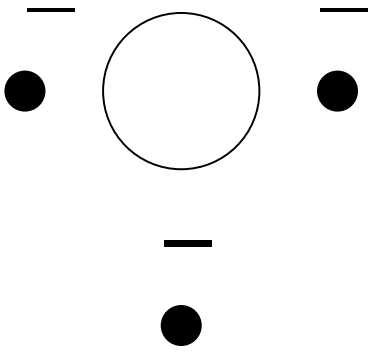
Example: Three instances

Can three points be shattered by the hypothesis space consisting of a set of circles?



Example: Three instances

Can three points be shattered by the hypothesis space consisting of a set of circles?



What about 4 points?

Example: Four instances

- These cannot be shattered, because we can label the farther 2 points as $+$, and the circle that contains them will necessarily contain the other points
- So circles can shatter one data set of three points (the one we've been analyzing), but there is no set of four points that can be shattered by circles (check this by yourself!)
- Note that not all sets of size 3 can be shattered!
- We say that the *VC dimension of circles is 3*

The Vapnik-Chervonenkis Dimension

Definition: The **Vapnik-Chervonenkis dimension**, $VC(H)$, of hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of X can be shattered by H, then $VC(H) \equiv \infty$.

- In other words, the VC dimension is the maximum number of points for which H is unbiased.
- VC dimension measures how many distinctions the hypotheses from H are able to make
- This is, in some sense, the number of "effective degrees of freedom"

A game with the "enemy"

- You are allowed to choose k points. *This actually gives you a lot of freedom!*
- The enemy then labels these points any way it wants
- You now have to produce a hypothesis, out of your hypothesis class, which correctly produces these labels.

If you are able to succeed at this game, the VC dimension is at least k.

To show that it is no greater than k, you have to show that *for any* set of $k+1$ points, the enemy can find a labeling that you cannot correctly reproduce with any of your hypotheses.

Example revisited: VC dimension of intervals

- Can we shatter 2 points on a line with an interval?
- Can we shatter 3 points on a line with one interval?
- What is the VC dimension of intervals?

Example revisited: VC dimension of intervals

- Can we shatter 2 points on a line with an interval?
  Yes!
- Can we shatter 3 points on a line with one interval?
  No! The enemy can label the most distant points $+$ and the middle one $-$
- What is the VC dimension of intervals?
  VC dimension is 2

VC dimension of linear decision surfaces

- Consider a linear threshold unit in the plane.
- First, show there exists a set of 3 points that can be shattered by a line $\Longrightarrow$ VC dimension of lines in the plane is at least 3.
- To show it is at most 3, show that NO set of 4 points can be shattered.
- For an n-dimensional space, VC dimension of linear estimators is $n + 1$.

Applying VC theory to feed-forward networks

- Let H_G be the class of functions that can be computed by feed-forward networks of perceptrons (also known as multi-layer perceptrons) defined on a fixed underlying graph G with E edges and $N \geq 2$ linear threshold nodes.
- Then it can be shown that $VC(H_G) \leq 2(E+N)\log(eN)$.

And the bad news...

Sigmoid-like functions *can* have infinite VC dimension! E.g.

$$\frac{1}{1+e^{-x}} + cx^3 e^{-x^2} \sin x$$

(see Macintyre and Sontag, 1993).

However: the usual sigmoid function, as well as the hyperbolic tangent, have finite VC dimension! :-)

But: it is doubly exponential... :-(

However, in practice, neural networks seem to approximate well even with a lot fewer examples (sometimes fewer than the number of weights).

Alternative analyses (see, e.g. Bartlett, 1996) suggest that the error may be related to the *magnitude* of the weights, rather than the number of weights, if the nodes are kept in their linear regions.

Error bounds using VC dimension

- Recall our error bound in the finite case:

$$e(h_{emp}) \leq \left(\min_{h \in H} e(h)\right) + 2\sqrt{\frac{1}{2m} \log \frac{2|H|}{\delta}}$$

- Vapnik showed a similar result, but using VC dimension instead of the size of the hypothesis space:
- For a hypothesis class H with VC dimension $VC(H)$, given m examples, with probability at least $1 - \delta$, we have:

$$e(h_{emp}) \leq \left(\min_{h \in H} e(h)\right) + O\left(\sqrt{\frac{VC(H)}{m} \log \frac{m}{VC(H)} + \frac{1}{m} \log \frac{1}{\delta}}\right)$$

Remarks on VC dimension

- The previous bound is tight up to log factors. In other words, for hypotheses classes with large VC dimension, we can show that there exists some data distribution which will produce a bad approximation.
- For many reasonable hypothesis classes (e.g. linear approximators) the VC dimension is linear in the number of "parameters" of the hypothesis. This shows that to learn "well", we need a number of examples that is linear in the VC dimension (so linear in the number of parameters, in this case).
- An important property: if $H_1 \subseteq H_2$ then $VC(H_1) \leq VC(H_2)$.

Structural risk minimization

$$e(h_{emp}) \leq \left(\min_{h \in H} e(h)\right) + O\left(\sqrt{\frac{VC(H)}{m} \log \frac{m}{VC(H)} + \frac{1}{m} \log \frac{1}{\delta}}\right)$$

- We have used this bound to measure the true error of the hypothesis with the smallest training error
- Why not use the bound directly to get the best hypothesis?
- We can measure the training error, and add to that the quantity suggested by the rightmost term
- We pick the hypothesis that is best in terms of this sum!
- This approach is called structural risk minimization, and can be used instead of crossvalidation or MDL to pick the best hypothesis class

Probably Approximately Correct (PAC) Learning

Let F be a concept (target function) class defined over a set of instances X in which each instance has length n. An algorithm L, using hypothesis class H is a **PAC learning algorithm** for F if:

- for any concept $f \in F$
- for any probability distribution P over X
- for any parameters $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$

the learner L will, with probability at least $(1 - \delta)$, output a hypothesis with true error at most ϵ.

A class of concepts F is **PAC-learnable** if there exists a PAC learning algorithm for F.

Computational vs Sample Complexity

- A class of concepts is **polynomial-sample PAC-learnable** if it is PAC learnable using a number of examples at most polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$ and n.
- A class of concepts is **polynomial-time PAC-learnable** if it is PAC learnable in time at most polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$ and n.
- *Sample complexity is often easier to bound than time complexity!*
- Sometimes there is a trade-off between the two (if there are more samples, less work is required to process each one and vice versa)

Bird Eye View of Computational Learning Theory

1. How hard is it to learn (in terms of the computation required)?
   Difficult to answer in general, but results have been established for simple problems (e.g. learning CNF and DNF formulae)
2. How many examples are required for a good approximation?
   A lot of results here, regarding sample complexity bounds for different algorithms
3. What problems can be solved by a given algorithm?
   Little work done here so far.

Different Models of Learning

- Examples come randomly from some fixed distribution (the case usually considered in supervised learning)
- The learner is allowed to ask questions to the teacher (active learning)
- Examples are given by an opponent (on-line learning, mistake-bound model)

Most of the time assumes that the examples are noise-free. However, results do exist for particular kinds of noise (e.g. noise in the target value).

Summary

- The complexity results for binary classification show trade-offs between the desired degree of precision ϵ, the number of samples m and the complexity of the hypothesis space H
- The complexity of H can be measured by the VC dimension
- For a fixed hypothesis space, minimizing the training set error is well justified (empirical risk minimization)
- We have not talked about
  - Relationship between margin and VC dimension (better bounds than the results discussed)
  - Lower bounds
  - ...