Lecture 11: Learning theory

- True error of a hypothesis
- Probably Approximately Correct (PAC) model
- VC-dimension
- Other computational learning theory models

Binary classification: The golden goal

Given:

- The set of all possible instances X
- A target function (or concept) $f : X \rightarrow \{0, 1\}$
- A set of hypotheses H
- A set of training examples D (containing positive and negative examples of the target function)

$$\langle \mathbf{x}_1, f(\mathbf{x}_1) \rangle, \ldots \langle \mathbf{x}_m, f(\mathbf{x}_m) \rangle$$

Determine:

A hypothesis $h \in H$ such that $h(\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x} \in X$.

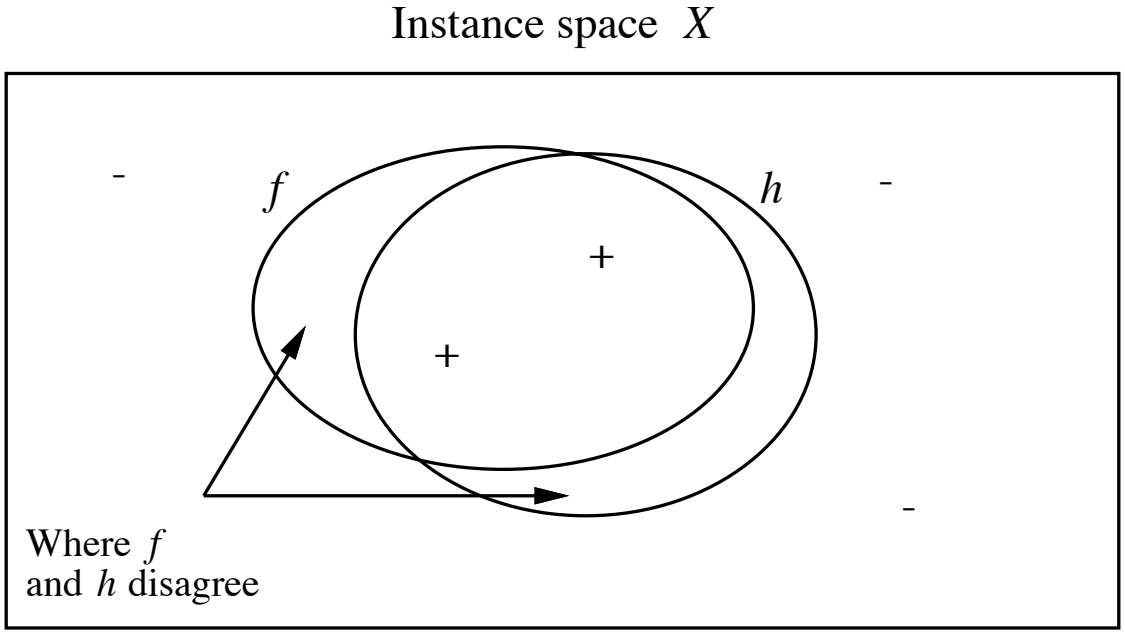
Approximate Concept Learning

- Requiring a learner to acquire the right concept is too strict
- Instead, we will allow the learner to produce a *good approximation* to the actual concept
- For any instance space, there is a non-uniform likelihood of seeing different instances
- We assume that there is a *fixed probability distribution* P on the space of instances X
- The learner is trained and tested on examples whose inputs are drawn *independently and randomly* according to P.

Recall: Two Notions of Error

- The *training error* of hypothesis h with respect to target concept f estimates how often $h(\mathbf{x}) \neq f(\mathbf{x})$ over the training instances
- The *true error* of hypothesis h with respect to target concept f estimates how often $h(\mathbf{x}) \neq f(\mathbf{x})$ over future, unseen instances (but drawn according to P)
- Questions:
  - Can we *bound the true error* of a hypothesis given only its training error?
  - How many examples are needed for a good approximation? This is called the *sample complexity* of the problem

True Error of a Hypothesis



True Error Definition

- The set of instances on which the target concept and the hypothesis disagree is denoted: $S = \{\mathbf{x}|h(\mathbf{x}) \neq f(\mathbf{x})\}$
- The *true error* of h with respect to f is:

$$\sum_{\mathbf{x} \in S} P(\mathbf{x})$$

  This is the probability of making an error on an instance randomly drawn from X according to P
- Let $\epsilon \in (0, 1)$ be an *error tolerance* parameter. We say that h is a *good approximation* of f (to within ϵ) if and only if the true error of h is less than ϵ.

Example: Rote Learner

- Let $X = \{0,1\}^n$. Let P be the uniform distribution over X.
- Let the concept f be generated by randomly assigning a label to every instance in X.
- Let $D \subset X$ be a set of training instances.

  The hypothesis h is generated by memorizing D and giving a random answer otherwise.

- What is the training error of h?
- What is the true error of h?

Empirical risk minimization

- Suppose we are given a hypothesis class H
- We have a magical learning machine that can sift through H and output the hypothesis with the smallest training error, h_{emp}
- This is process is called *empirical risk minimization*
- Is this a good idea?
- What can we say about the error of the other hypotheses in h?

First tool: The union bound

- Let $E_1 \ldots E_k$ be k different events (not necessarily independent). Then:

$$P(E_1 \cup \cdots \cup E_k) \leq P(E_1) + \cdots + P(E_k)$$

- Note that this is usually loose, as events may be correlated

Second tool: Hoeffding bound

- Let $Z_1 \ldots Z_m$ be m independent identically distributed (iid) binary variables, drawn from a Bernoulli (binomial) distribution:

$$P(Z_i = 1) = \phi \text{ and } P(Z_i = 0) = 1 - \phi$$

- Let $\hat{\phi}$ be the mean of these variables: $\hat{\phi} = \frac{1}{m} \sum_{i=1}^{m} Z_i$
- Let ϵ be a fixed error tolerance parameter. Then:

$$P(|\phi - \hat{\phi}| > \epsilon) \leq 2e^{-2\epsilon^2 m}$$

- In other words, if you have lots of examples, the empirical mean is a good estimator of the true probability.
- Note: other similar concentration inequalities can be used (e.g. Chernoff, Bernstein, etc.)

Finite hypothesis space

- Suppose we are considering a finite hypothesis class $H = \{h_1, \ldots h_k\}$ (e.g. conjunctions, decision trees,...)
- Take an arbitrary hypothesis $h_i \in H$
- Suppose we sample data according to our distribution and let $Z_j = 1$ iff $h_i(\mathbf{x}_j) \neq y_j$
- So $e(h_i)$ (the true error of h_i) is the expected value of Z_j
- Let $\hat{e}(h_i) = \frac{1}{m}\sum_{j=1}^{m} Z_j$ (this is the empirical training error of h_i on the data set we have)
- Using the Hoeffding bound, we have:

$$P(|e(h_i) - \hat{e}(h_i)| > \epsilon) \leq 2e^{-2\epsilon^2 m}$$

- So, if we have *lots of data*, the *training error of a hypothesis h_i will be close to its true error* with high probability.

What about all hypotheses?

- We showed that the empirical error is "close" to the true error for one hypothesis.
- Let E_i denote the event $|e(h_i) - \hat{e}(h_i)| > \epsilon$
- Can we guarantee this is true for all hypothesis?

$$
\begin{aligned}
P(\exists h_i \in H, |e(h_i) - \hat{e}(h_i)| > \epsilon) &= P(E_1 \cup \ldots E_k) \\
&\leq \sum_{i=1}^{k} P(E_i) \text{ (union bound)} \\
&\leq \sum_{i=1}^{k} 2e^{-2\epsilon^2 m} \text{ (shown before)} \\
&= 2ke^{-2\epsilon^2 m}
\end{aligned}
$$

A uniform convergence bound

- We showed that:

$$P(\exists h_i \in H, |e(h_i) - \hat{e}(h_i)| > \epsilon) \leq 2ke^{-2\epsilon^2 m}$$

- So we have:

$$1 - P(\exists h_i \in H, |e(h_i) - \hat{e}(h_i)| > \epsilon) \geq 1 - 2ke^{-2\epsilon^2 m}$$

  or, in other words:

$$P(\forall h_i \in H, |e(h_i) - \hat{e}(h_i)| < \epsilon) \geq 1 - 2ke^{-2\epsilon^2 m}$$

- This is called a *uniform convergence* result because the bound holds for all hypotheses
- What is this good for?

Sample complexity

- Suppose we want to guarantee that with probability at least $1 - \delta$, the sample (training) error is within ϵ of the true error.
- From our bound, we can set $\delta \geq 2ke^{-2\epsilon^2 m}$
- Solving for m, we get that the number of samples should be:

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

- So the *number of samples needed is logarithmic* in the size of the hypothesis space

Example: Conjunctions of Boolean Literals

- Let H be the space of all pure conjunctive formulae over n Boolean attributes. Then $|H| = 3^n$ (why?)
- From the previous result, we get:

$$m \geq \frac{1}{2\epsilon^2} \log \frac{2|H|}{\delta} = n \frac{1}{2\epsilon^2} \log \frac{6}{\delta}$$

- This is linear in n!
- Hence, conjunctions are "easy to learn"

Another application: Bounding the true error

- Our inequality revisited:

$$P(\forall h_i \in H, |e(h_i) - \hat{e}(h_i)| < \epsilon) \geq 1 - 2ke^{-2\epsilon^2 m} = 1 - \delta$$

- Suppose we hold m and δ fixed, and we solve for ϵ. Then we get:

$$|e(h_i) - \hat{e}(h_i)| \leq \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$$

  inside the probability term.
- Can we now prove anything about the generalization power of the empirical risk minimization algorithm?

Empirical risk minimization

Let h^* be the best hypothesis in our class (in terms of true error). Based on our uniform convergence assumption, we can bound the true error of h_{emp} as follows:

$$
\begin{aligned}
e(h_{emp}) &\leq \hat{e}(h_{emp}) + \epsilon \\
&\leq \hat{e}(h^*) + \epsilon \text{ (because } h_{emp} \text{ has better training error} \\
&\quad \text{than any other hypothesis)} \\
&\leq e(h^*) + 2\epsilon \text{ (by using the result on } h^*) \\
&\leq e(h^*) + 2\sqrt{\frac{1}{2m}\log\frac{2|H|}{\delta}} \text{ (from previous slide)}
\end{aligned}
$$

This bounds how much worse h_{emp} is, wrt the best hypothesis we can hope for!

Bias and variance revisited

- We showed that, given m examples, with probability at least $1 - \delta$,

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

- Suppose now that we are considering two hypothesis classes $H \subseteq H'$
  - The first term would be smaller for H' (we have a larger hypothesis class, hence less "bias")
  - The second term would be larger (the "variance" is increasing)
- Note, though, that if H is infinite, this result is not very useful...

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:

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

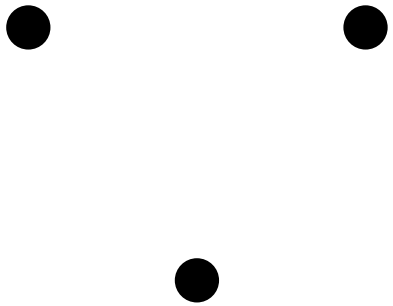
- But for us, $|H| = \infty$!
- We need a way to characterize the "complexity" of infinite-dimensional classes of hypotheses

Shattering a set of instances

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

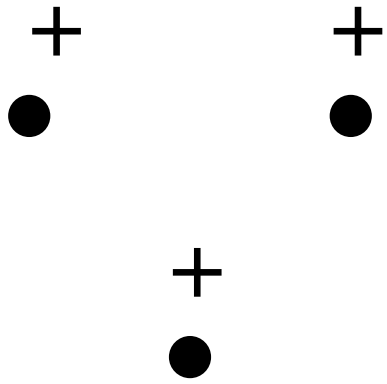
Example: Three instances

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



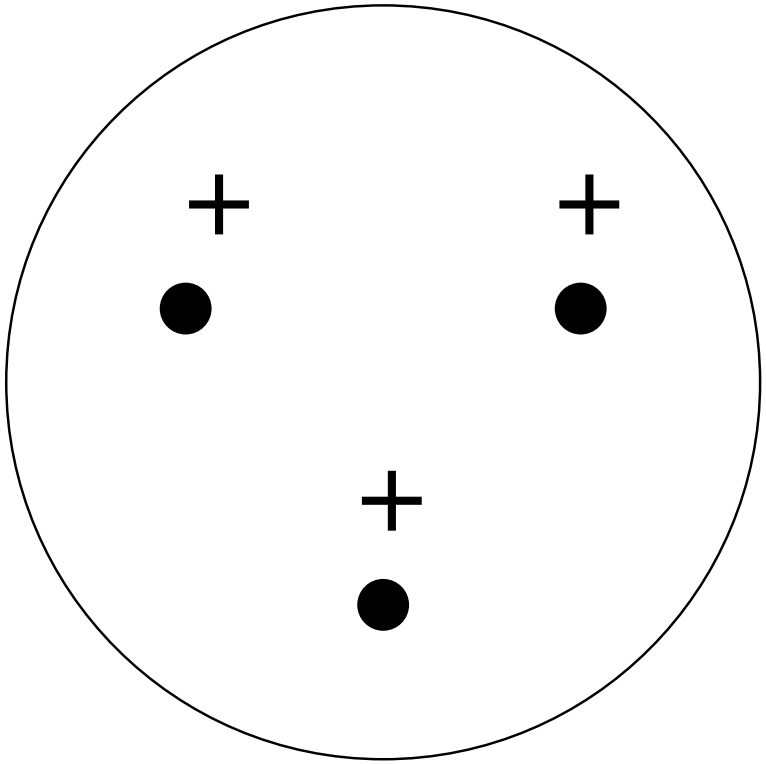
Example: Three instances

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



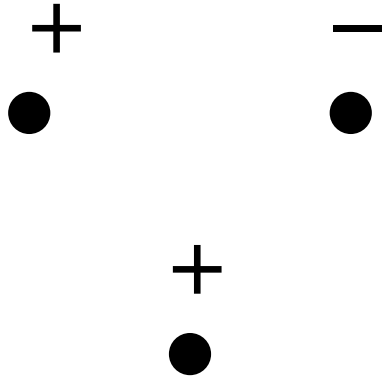
Example: Three instances

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



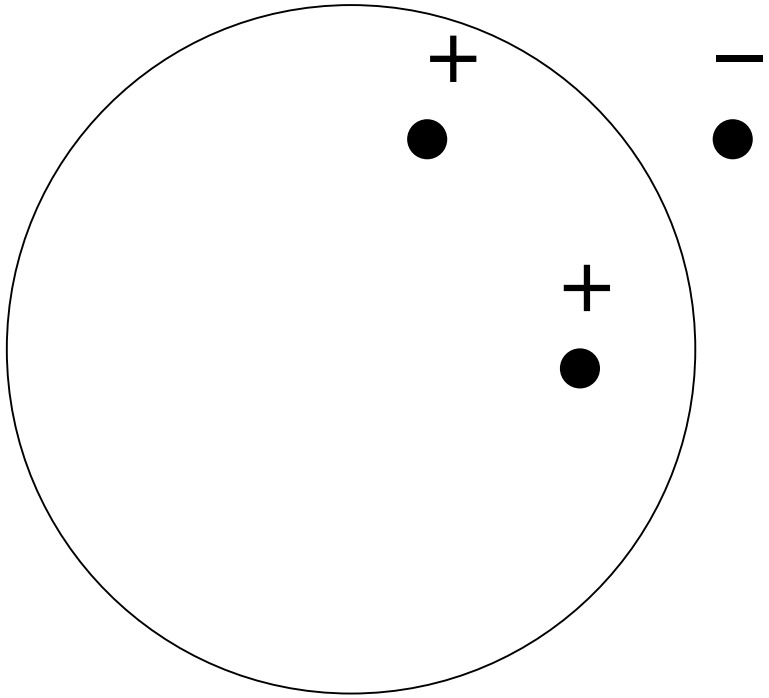
Example: Three instances

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



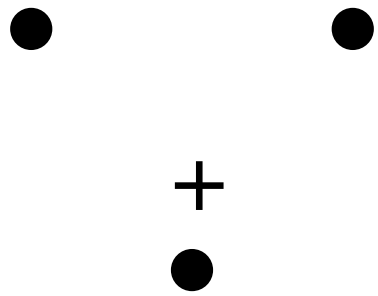
Example: Three instances

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



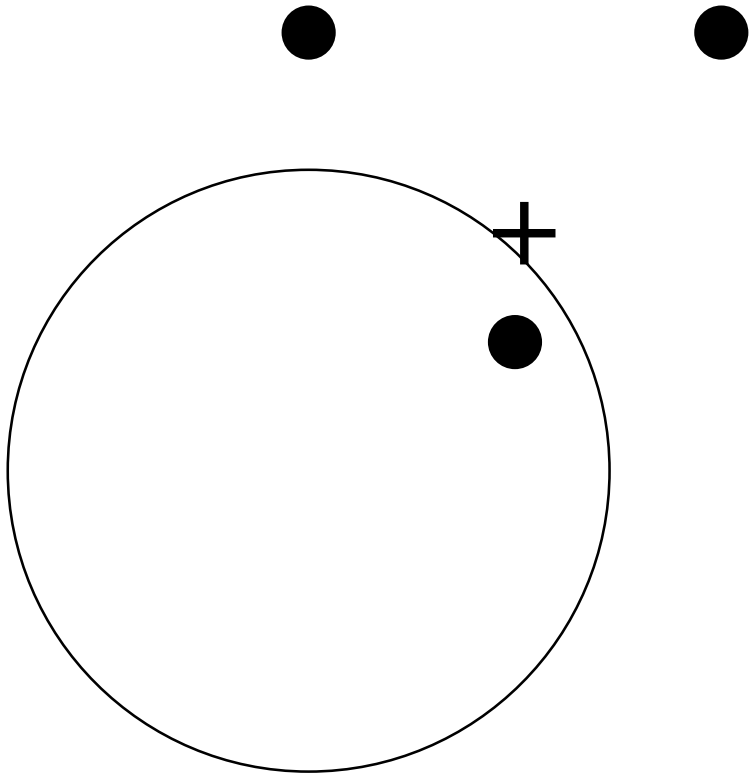
Example: Three instances

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



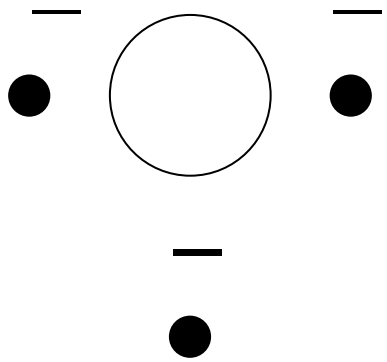
Example: Three instances

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 (VC) Dimension

- 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"

Establishing the VC dimension

- Play the following 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 some 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) - we will look at this again later
- Examples are given by an opponent (on-line learning, mistake-bound model)

Most of the time the results assume 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
  - ...