CS 886
Applied Machine Learning
Learning Theory Light

Dan Lizotte

University of Waterloo

23 May 2013

Lecture 12: Learning theory

- True error vs. Training Error of a hypothesis
- VC-dimension

Binary classification: The golden goal

Given:

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

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

Determine:

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

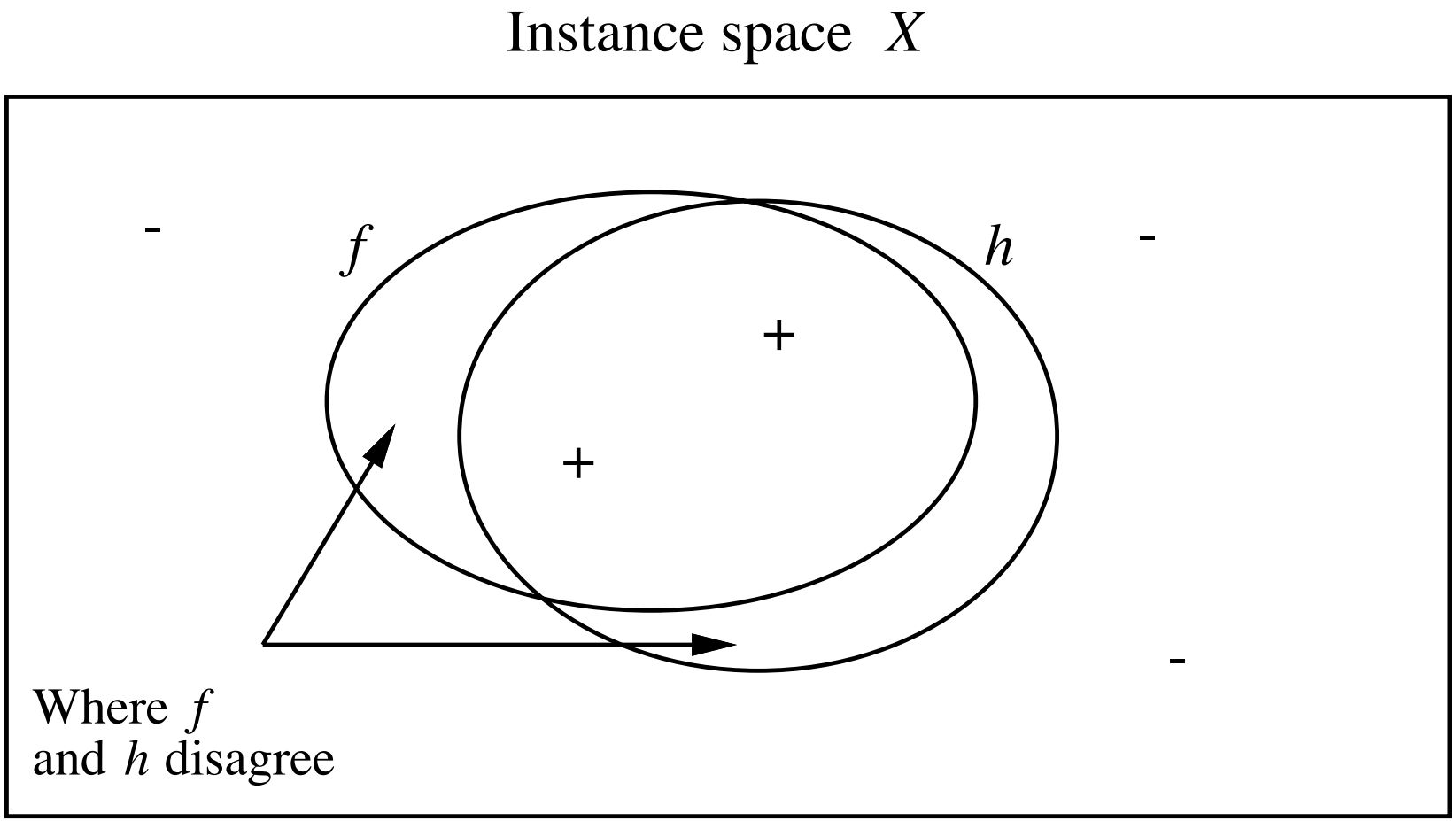
Approximate Concept Learning

- 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 *empirical error* of hypothesis h with respect to target concept f and a data sample tells how often $h(\mathbf{x}) \neq f(\mathbf{x})$ over the sample
- The *true error* of hypothesis h with respect to target concept f tells 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 empirical 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:

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

  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 $X = \{0, 1\}^p$. 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 empirical error of h on D?
- 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 $Z_1 \ldots Z_n$ be n independent identically distributed (iid) binary variables, drawn from a Bernoulli (weighted coin) distribution:

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

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

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

- 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 $H = \{h_1, \ldots h_k\}$
- Choose a hypothesis $h_i \in H$ at random.
- 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 $\hat{e}(h_i) = \frac{1}{n}\sum_{j=1}^{n} Z_j$ (this is the empirical error of h_i on the data set we have)
- Using the Hoeffding bound, we have:

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

- So, if we have *lots of data*, the *empirical error of a hypothesis h_i will be close to its true error* with high probability. NOTE WE DID NOT TRAIN ON THE DATA.

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?

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

A uniform convergence bound

- We showed that:

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

- So we have:

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

  or, in other words:

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

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

A uniform convergence bound

- We showed that:

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

- So we have:

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

  or, in other words:

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

- This is called a *uniform convergence* result because the bound holds for all hypotheses
- What is this good for? If **all** the empirical errors are close to the true errors, the h_i with the best empirical error is probably the h_i with the best true error.

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 $\delta \geq 2ke^{-2\epsilon^2 n}$
- Solving for n, we get that the number of samples should be:

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

- 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 p Boolean attributes. Then $|H| = 3^p$ (why?)
- From the previous result, we get:

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

- This is linear in p
- Hence, conjunctions are “easy to learn"

Example: Arbitrary Boolean functions

- Let H be the space of all Boolean functions on p Boolean attributes. Then $|H| = 2^{2^p}$ (why?)
- From the previous result, we get:

$$n \geq \frac{1}{2\epsilon^2} \log \frac{2|H|}{\delta} = \frac{1}{2\epsilon^2}(2^p + \log \frac{1}{\delta})$$

- This is exponential in p
- Hence, arbitrary Boolean functions are “hard to learn"

Another application: Bounding the true error

- Our inequality revisited:

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

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

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

  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:

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

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

Overfitting/Underfitting Revisited

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

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

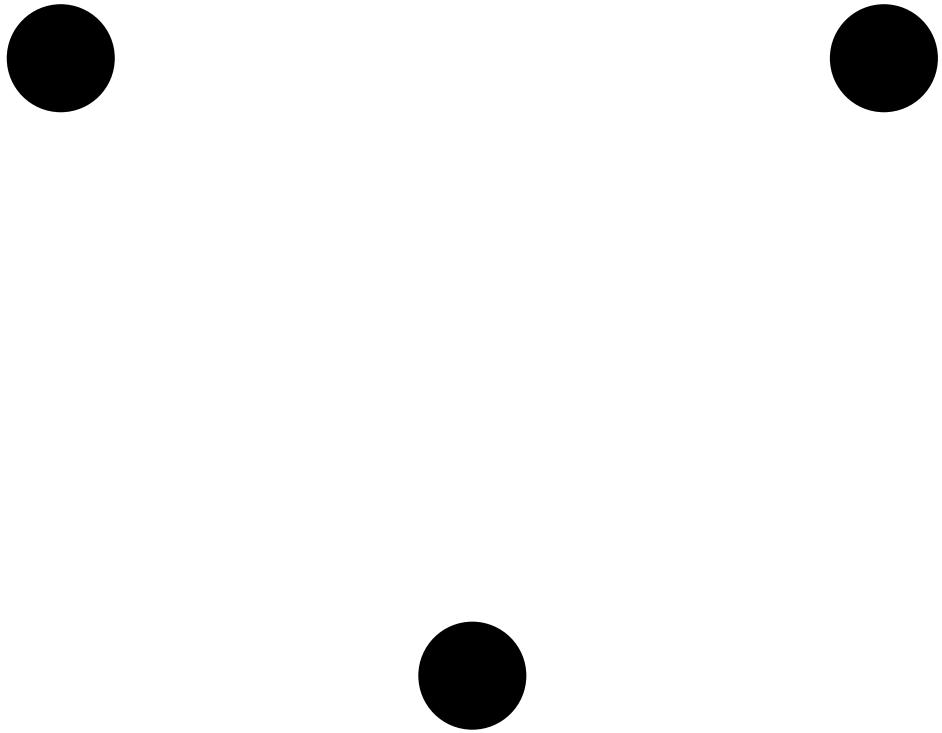
- Suppose now that we are considering two hypothesis classes $H \subseteq H'$
  - The first term would be smaller for H' (we have a larger hypothesis class, hence more able to fit the 'truth')
  - The second term would be larger (overfitting is more likely)
- Note, though, that if H is infinite, this result is not very useful...

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.
- This idea will let us measure the complexity of a hypothesis space even if it is infinite.

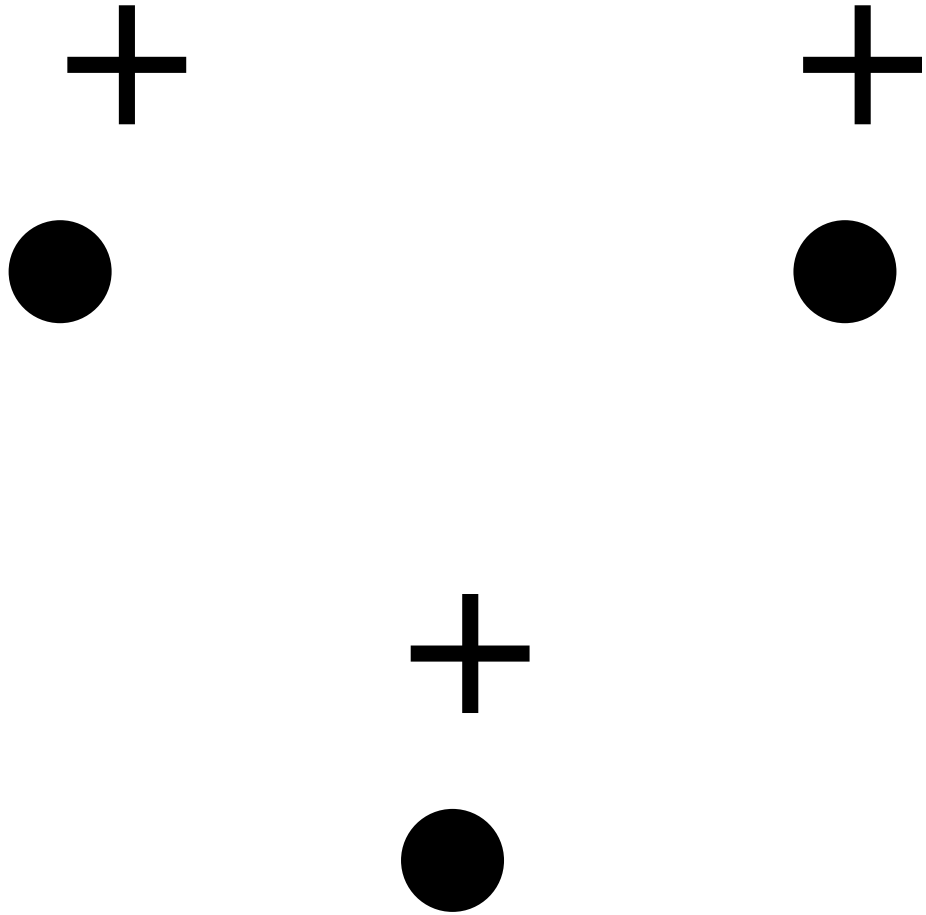
Example: Three instances

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



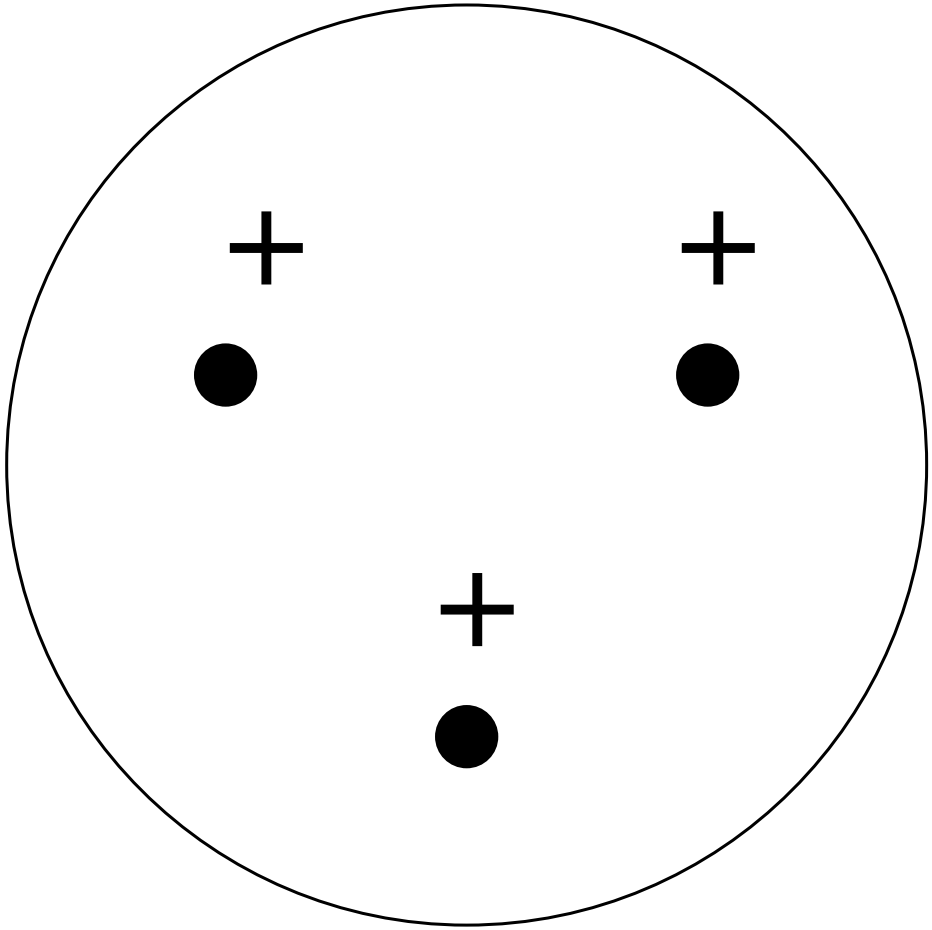
Example: Three instances

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



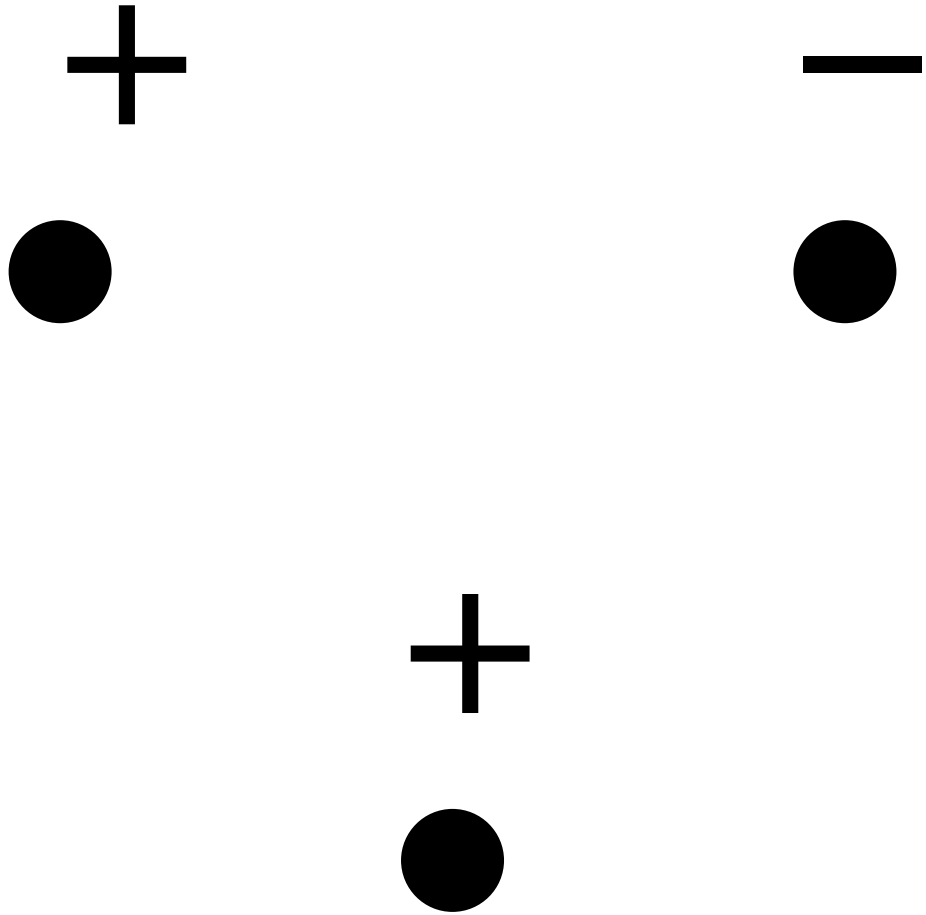
Example: Three instances

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



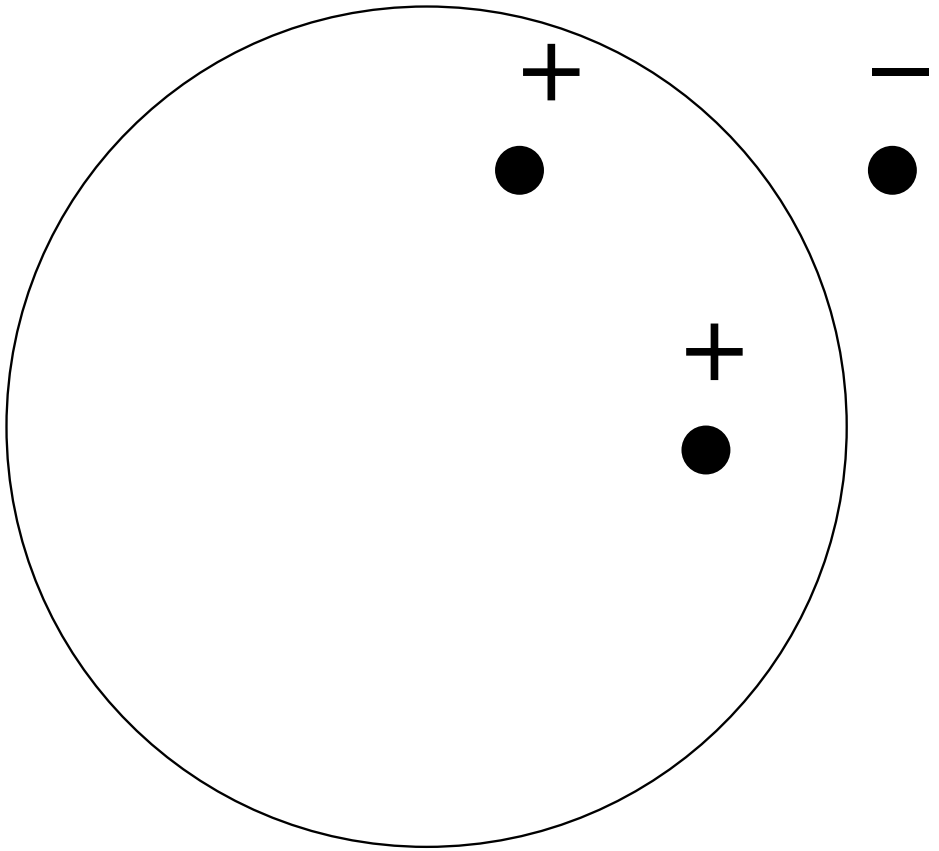
Example: Three instances

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



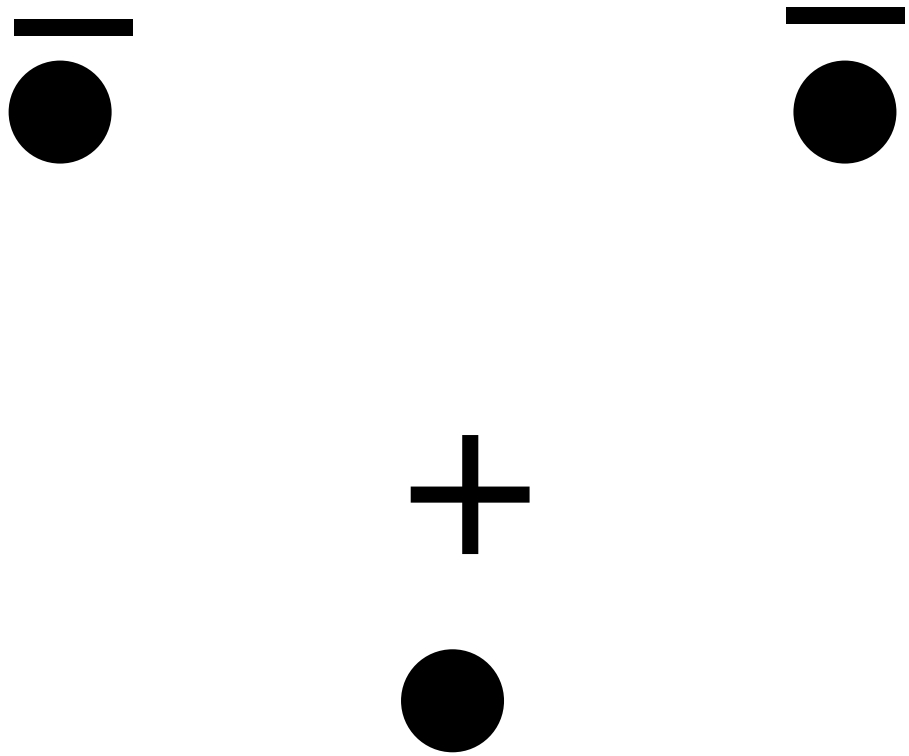
Example: Three instances

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



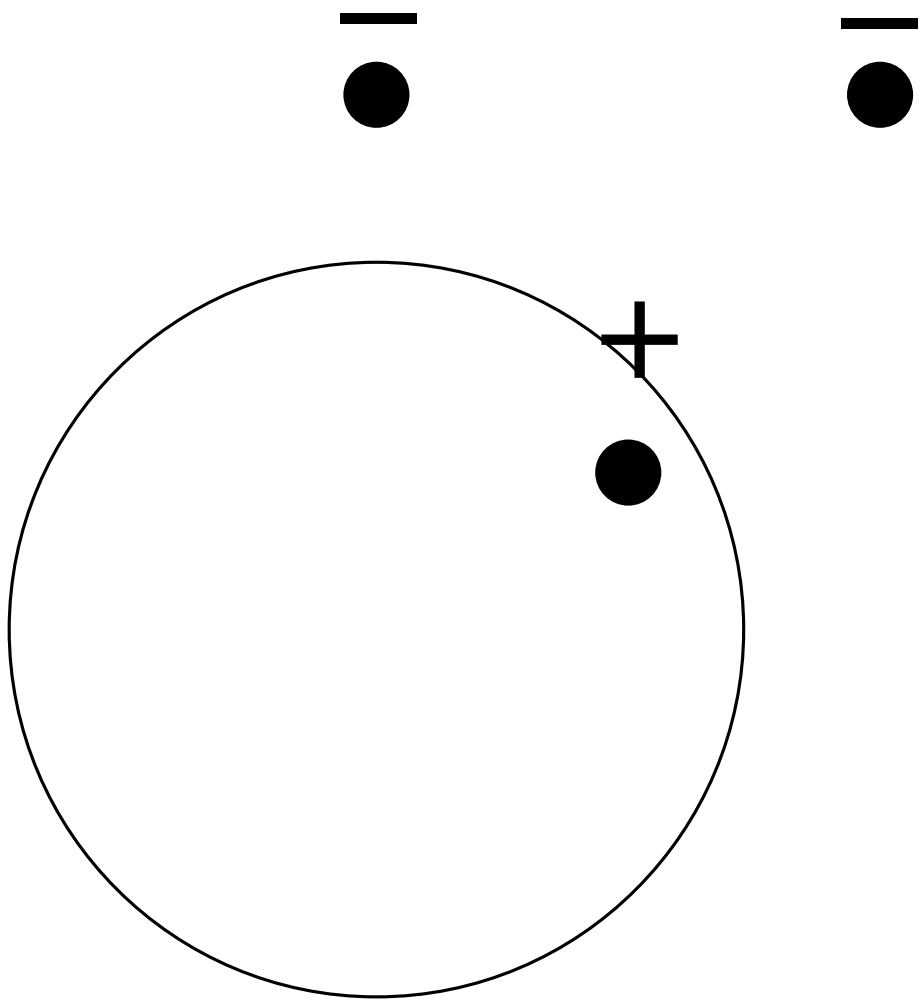
Example: Three instances

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



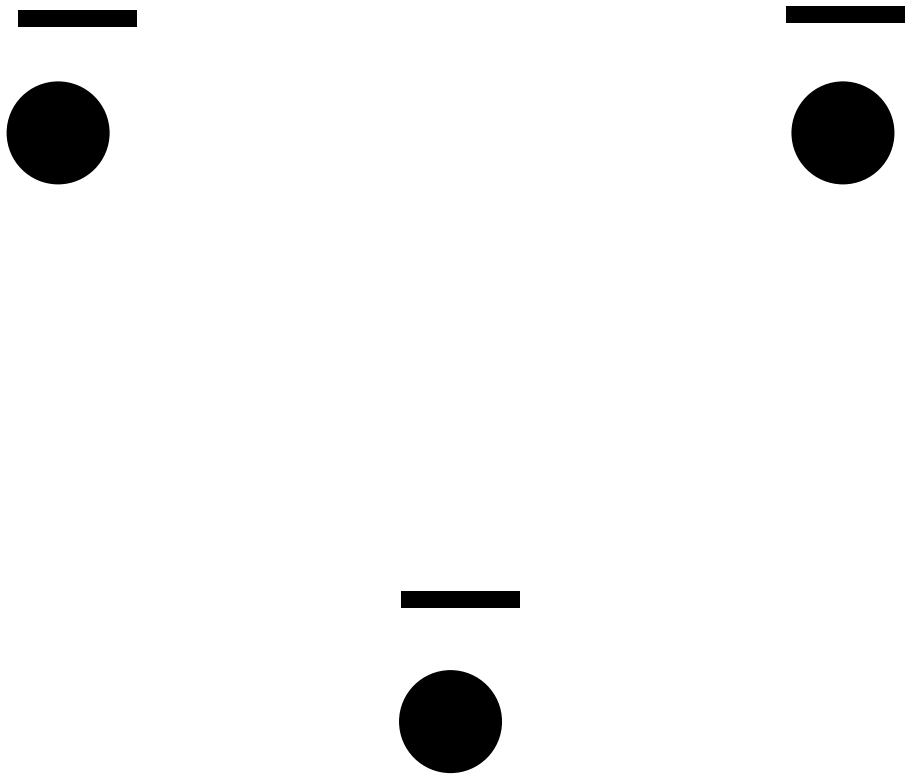
Example: Three instances

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



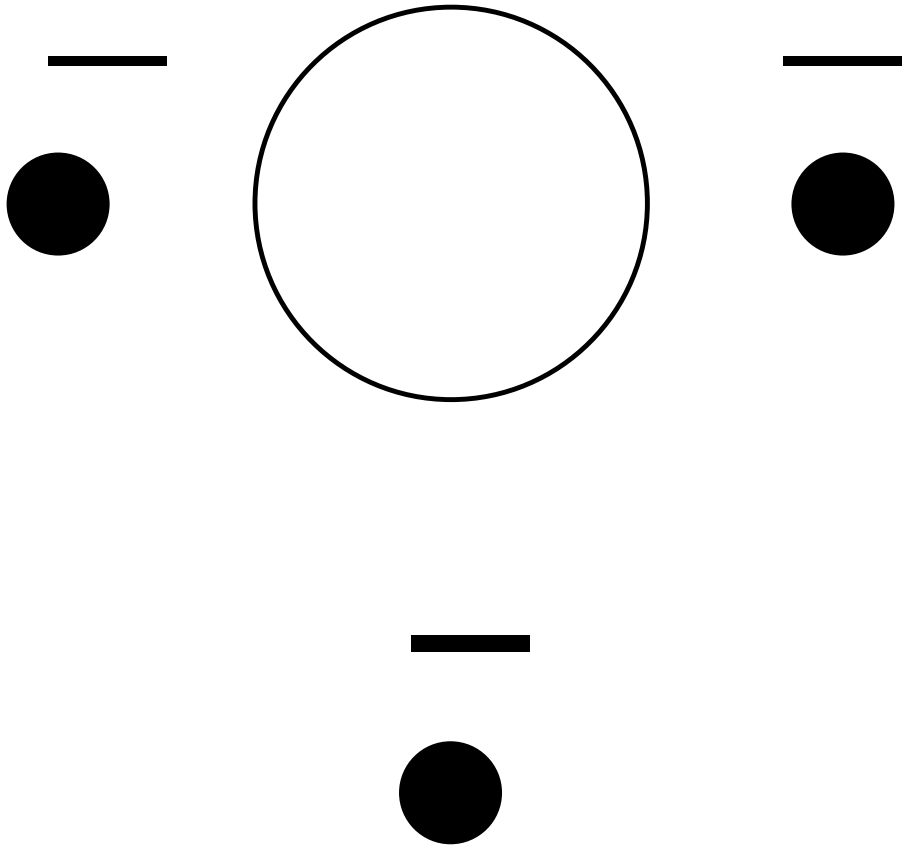
Example: Three instances

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



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 $VC(H) \equiv \infty$.
- 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 adversary:
  - You are allowed to *choose k points*. This actually gives you a lot of freedom!
  - The adversary 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 adversary can find a labeling that you cannot correctly reproduce with any of your hypotheses.

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 p-dimensional space, VC dimension of linear estimators is $p + 1$.

Error bounds using VC dimension

- Recall our error bound in the finite case:

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

- 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 n examples, with probability at least $1 - \delta$, we have:

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

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

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

- 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 cross-validation