Lectures 23-24: Introduction to learning theory

- True error of a hypothesis (classification)
- Some simple bounds on error and sample size
- Introduction to VC-dimension

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* D on the space of instances $\mathcal{X}$
- The learner is trained and tested on examples whose inputs are drawn *independently and randomly (iid)* according to D.

Generalization Error and Empirical Risk

- Given a hypothesis $h \in \mathcal{H}$, a target concept f, and an underlying distribution D, the generalization error or risk of h is defined by

$$R(h) = \mathop{\mathbb{E}}_{\mathbf{x} \sim D} [\ell(h(\mathbf{x}), f(\mathbf{x}))]$$

  where ℓ is an error function. This measures the true error of the hypothesis.

- Given training data $\mathcal{S} = \{(\mathbf{x}_i, f(\mathbf{x}_i)\}_{i=1}^{m}$, the empirical risk of h is defined by

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(\mathbf{x}_i), f(\mathbf{x}_i)).$$

  This is the average error over the sample $\mathcal{S}$, it measures the training error of the hypothesis.

Binary Classification: $0-1$ loss

- For binary classification, a natural error measure is the $0-1$ loss, which counts mismatches between $h(\mathbf{x})$ and $f(\mathbf{x})$:

$$\ell(h(\mathbf{x}), f(\mathbf{x})) = \mathrm{I}(h(\mathbf{x}) \neq f(\mathbf{x})) = \begin{cases} 1 & \text{if } h(\mathbf{x}) \neq f(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases}$$

  where I is the *indicator function*.

- In this case, the generalization error and empirical risk are given by

$$R(h) = \mathop{\mathbb{E}}_{\mathbf{x}\sim D}[\mathrm{I}(h(\mathbf{x}) \neq f(\mathbf{x}))] = \mathop{\mathbb{P}}_{\mathbf{x}\sim D}[h(\mathbf{x}) \neq f(\mathbf{x})]$$

$$\hat{R}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}(h(\mathbf{x}) \neq f(\mathbf{x})).$$

The Two Notions of Error for Binary Classification

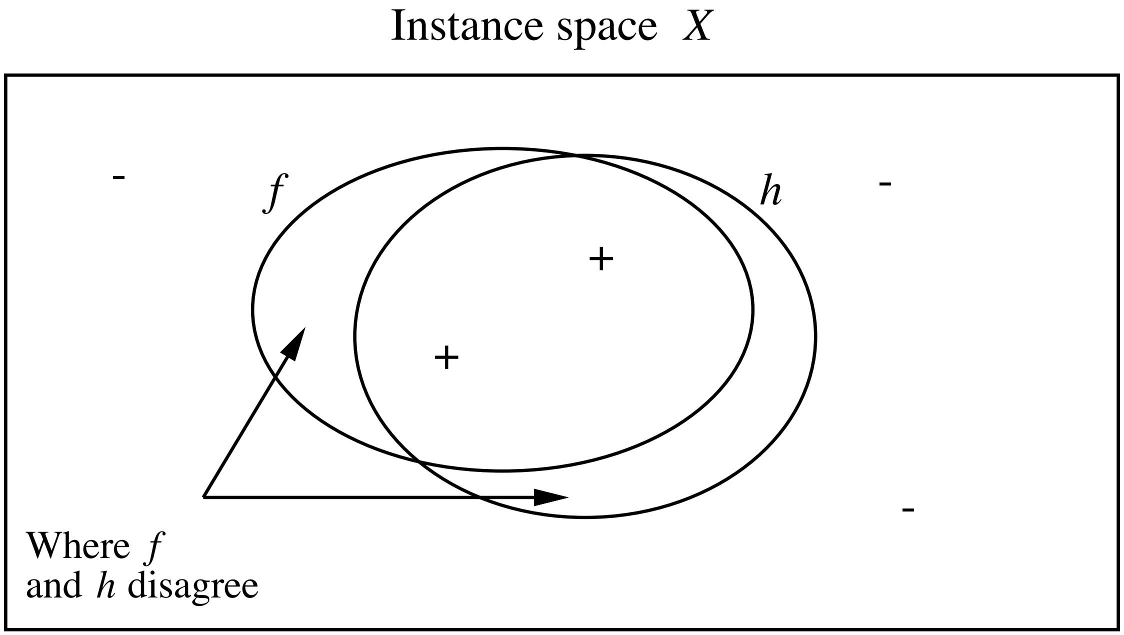
- The *training error* of hypothesis h with respect to target concept f estimates how often $h(\mathbf{x}) \neq f(\mathbf{x})$ over the training instances
- The *true error* of hypothesis h with respect to target concept f estimates how often $h(\mathbf{x}) \neq f(\mathbf{x})$ over future, unseen instances (but drawn according to D)
- Questions:
  - Can we bound the true error of a hypothesis given its training error? i.e. can we bound the generalization error by the empirical risk?

    $\hookrightarrow$ Generalization bounds
  - Can we find an hypothesis with small true error after observing a reasonable number of training points?

    $\hookrightarrow$ PAC learnability
  - How many examples are needed for a good approximation?

    $\hookrightarrow$ Sample complexity

True Error of a Hypothesis



True Error Definition

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

  $$\sum_{\mathbf{x} \in E} \mathbb{P}_{\mathbf{x} \sim D}[\mathbf{x}]$$

  This is the probability of making an error on an instance randomly drawn from $(\mathcal{X}, \mathcal{Y})$ according to D
- Let $\varepsilon \in (0, 1)$ 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 $\mathcal{X} = \{0, 1\}^n$. Let P be the uniform distribution over $\mathcal{X}$.
- Let the concept f be generated by randomly assigning a label to every instance in X.
- We assume that there is no output noise, so every instance $\mathbf{x}$ we get is labelled with the true $f(\mathbf{x})$
- Let $\mathcal{S} \subseteq \mathcal{X}$ be a set of training instances.

  The hypothesis h is generated by memorizing $\mathcal{S}$ and giving a random answer otherwise.

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

Example: Empirical and True Error

- Since we assumed that examples are labelled correctly and memorized, the empirical error is 0
- For the true error, suppose we saw m distinct examples during training, out of the total set of 2^n possible examples
- For the examples we saw, we will make no error
- For the $(2^n - m)$ examples we did not see, we will make an error with probability $1/2$
- Hence, the true error is:

$$R(h) = \mathbb{P}[h(\mathbf{x}) \neq f(\mathbf{x})] = \frac{2^n - m}{2^n}\frac{1}{2} = \left(1 - \frac{m}{2^n}\right)\frac{1}{2}$$

- Note that the true error also goes to 0 as m approaches the number of examples, 2^n
- The difference of the true and empirical error also goes to 0 as m increases

Probably Approximately Correct (PAC) Learning

- A concept class C is PAC-learnable if there exists an algorithm A such that:

  for all $f \in C$, $\varepsilon > 0$, $\delta > 0$, all distributions D, and any sample size $m \geq poly(1/\varepsilon, 1/\delta)$ the following holds:

  $$\mathop{\mathbb{P}}_{S \sim D^m}[R(h_S) \leq \varepsilon] \geq 1 - \delta$$

  If furthermore A runs in time $poly(1/\varepsilon, 1/\delta)$, C is said to be efficiently PAC-learnable.

- Intuition: the hypothesis returned by A after observing a *polynomial number of points* is *approximately correct* (error at most ε) with *high probability* (at least $1 - \delta$).

PAC learning (cont'd)

- Remarks:
  - The concept class is known to the algorithm.
  - Distribution free model: no assumption on D.
  - Both training and test examples are drawn from D.
- Examples:
  - Axis aligned rectangle are PAC learnable[1].
  - Conjunctions of boolean literals are PAC learnable but the class of disjunctions of two conjunctions is not.
  - Linear thresholds (e.g. perceptron) are PAC learnable but the classes of conjunctions/disjunctions of two linear thresholds is not, nor is the class of multilayer perceptrons.

[1] see [Mohri et al., *Fundations of Machine Learning*] section 2.1.

Empirical risk minimization

- Suppose we are given a hypothesis class $\mathcal{H}$
- We have a magical learning machine that can sift through $\mathcal{H}$ and output the hypothesis with the smallest empirical 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:

$$\mathbb{P}(E_1 \cup \cdots \cup E_k) \leq \mathbb{P}(E_1) + \cdots + \mathbb{P}(E_k)$$

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

Second tool: Hoeffding bound

- Hoeffding inequality. Let $Z_1 \dots Z_m$ be m independent identically distributed (iid) random variables taking their values in $[a, b]$. Then for any $\varepsilon > 0$

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}[Z] > \varepsilon\right] \leq \exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right)$$

Second tool: Hoeffding bound

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

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

- 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:

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

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

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

- 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 $|R(h_i) - \hat{R}(h_i)| > \varepsilon$
- Can we guarantee this is true for all hypothesis?

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

A uniform convergence bound

- We showed that:

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

- So we have:

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

  or, in other words:

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

- 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:

$$P(\forall h_i \in H, |R(h_i) - \hat{R}(h_i)| < \varepsilon) \geq 1 - \delta$$

- From the previous result, it would be sufficient to have: $1 - 2ke^{-2\varepsilon^2 m} \geq 1 - \delta$
- We get $\delta \geq 2ke^{-2\varepsilon^2 m}$
- Solving for m, we get that the number of samples should be:

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

- So the *number of samples needed is logarithmic in the size of the hypothesis space* and *depends polynomially on $1/\varepsilon$ and $1/\delta$*

Example: Conjunctions of Boolean Literals

- Let $\mathcal{H}$ be the space of all pure conjunctive formulas over n Boolean attributes.
- Then $|\mathcal{H}| = 3^n$, because for each of the n attributes, we can include it in the formula, include its negation, or not include it at all.
- From the previous result, we get:

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

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

Example: Arbitrary Boolean functions

- Let $\mathcal{H}$ be the space of all Boolean formulae over n Boolean attributes.
- Every Boolean formula can be written in canonical form as a disjunction of conjunctions
- We have seen that there are 3^n possible conjunctions over n Boolean variables, and for each of them, we can choose to include it or not in the disjunction, so $|\mathcal{H}| = 2^{3^n}$
- From the previous result, we get:

$$m \geq \frac{1}{2\varepsilon^2} \log \frac{2|H|}{\delta} = \frac{3^n}{2\varepsilon^2} \log \frac{2}{\delta}$$

- This is exponential in n!
- Hence, arbitrary Boolean functions are "hard to learn"
- A similar argument can be applied to show that even restricted classes of Boolean functions, like parity and XOR, are hard to learn

Bounding the True Error by the Empirical Error

- Our inequality revisited:

$$\mathbb{P}(\forall h_i \in H, |R(h_i) - \hat{R}(h_i)| < \varepsilon) \geq 1 - 2|\mathcal{H}|e^{-2\varepsilon^2 m} \geq 1 - \delta$$

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

$$|R(h_i) - \hat{R}(h_i)| \leq \sqrt{\frac{1}{2m} \log \frac{2|\mathcal{H}|}{\delta}}$$

  inside the probability term.

- We are now ready to see how good empirical risk minimization is

Analyzing 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:

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

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

Types of error

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

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

- The first term is a characteristic of the hypothesis class $\mathcal{H}$, also called *approximation error*
- For a hypothesis class which is consistent (can represent the target function exactly) this term would be 0
- The second term decreases as the number of examples increases, but increases with the size of the hypothesis space
- This is called *estimation error* and is similar in flavour to variance
- Large approximation errors lead to “under fitting”, large estimation errors lead to overfitting

Controlling the complexity of learning

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

- Suppose now that we are considering two hypothesis classes $\mathcal{H} \subseteq \mathcal{H}'$
- The approximation error would be smaller for $\mathcal{H}'$ (we have a larger hypothesis class) but the second term would be larger (we need more examples to find a good hypothesis in the larger set)
- We could try to optimize this bound directly, by measuring the training error and adding to it the rightmost term (which is a penalty for the size of the hypothesis space)
- We would then 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 or other types of regularization
- Note, though, that if $\mathcal{H}$ 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:

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

- Simple learning algorithm: Observe m samples, and return $[0, b]$, where b is the largest positive example seen
- 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 (because our hypothesis would be less than ε away form the true target function)
- What is the probability of seeing m iid examples *outside* of $[c, a]$?

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

- If we want

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

Example continued

- Fact:

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

- Hence, it is sufficient to have

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

- Using this fact, we get:

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

- 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 hypotheses! (in fact, about an infinite number of them)
- What if there are more “degrees of freedom”?

Example: Learning two-sided intervals

- Suppose the target concept is positive (i.e. has value 1) inside some unknown interval $[a, b]$ and negative outside of it
- The hypothesis class consists of all closed intervals (so the target can be represented exactly.
- Given a data set D, a "conservative" hypothesis is to guess the interval: $[\min_{(x,1)\in D} x, \max_{(x,1)\in D} x]$
- We can make errors on either side of the interval, if we get no example within ε of the true values a and b respectively.
- The probability of an example outside of an ε-size interval is $1 - \varepsilon$
- The probability of m examples outside of it is $(1 - \varepsilon)^m$
- The probability this happens on either side is $\leq 2(1 - \varepsilon)^m \leq 2e^{-\varepsilon m}$, and we want this to be $< \delta$

Example (continued)

- If we extract the number of samples we get:

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

  This is just like the bound for 1-sided intervals, but with a 2 instead of a 1!

- Compare this with the bound in the finite case:

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

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

Infinite hypothesis class

- For any set of points $C = \{x_1, \cdots, c_m\} \subset \mathcal{X}$ we define the *restriction of* $\mathcal{H}$ *to* C by

$$\mathcal{H}_C = \{(h(x_1), h(x_2), \cdots, h(x_m)) \ : \ h \in \mathcal{H}\}.$$

- We showed that, given m examples, for any $h \in \mathcal{H}$ with probability at least $1 - \delta$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2m} \log \frac{2|\mathcal{H}|}{\delta}}$$

- Even if $\mathcal{H}$ is infinite, it is its effective size that matters: since $|\mathcal{H}_C| \leq 2^m$ when C has size m we can actually get

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{1}{2m} \log \frac{2^{m+1}}{\delta}}$$

- But this is too loose: the second term doesn't converge to 0...

VC dimension

- $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ is a set of hypothesis.
- For any set of points $C = \{x_1, \cdots, c_m\} \subset \mathcal{X}$ we define the *restriction of* $\mathcal{H}$ *to* C by

$$\mathcal{H}_C = \{(h(x_1), h(x_2), \cdots, h(x_m)) \ : \ h \in \mathcal{H}\}.$$

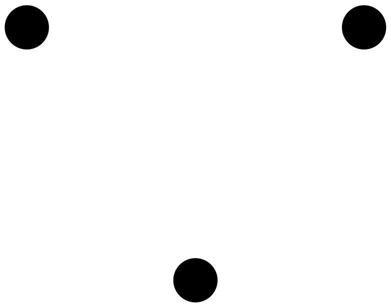
- We say that $\mathcal{H}$ shatters C if $|\mathcal{H}_C| = 2^{|C|}$.

→ *If someone can explain everything, his explanations are worthless*

- The VC dimension of $\mathcal{H}$ is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$.

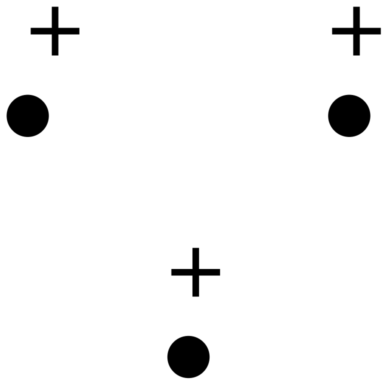
Example: Three instances

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



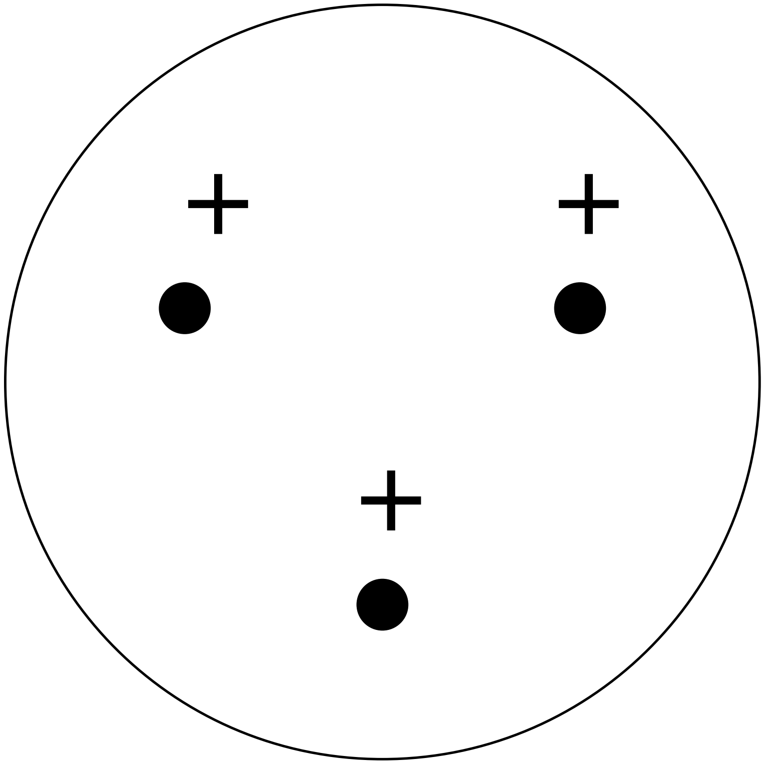
Example: Three instances dichotomy

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



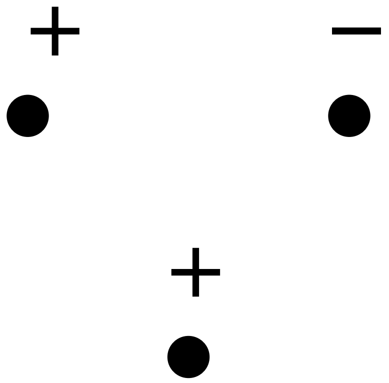
Example: Three instances

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



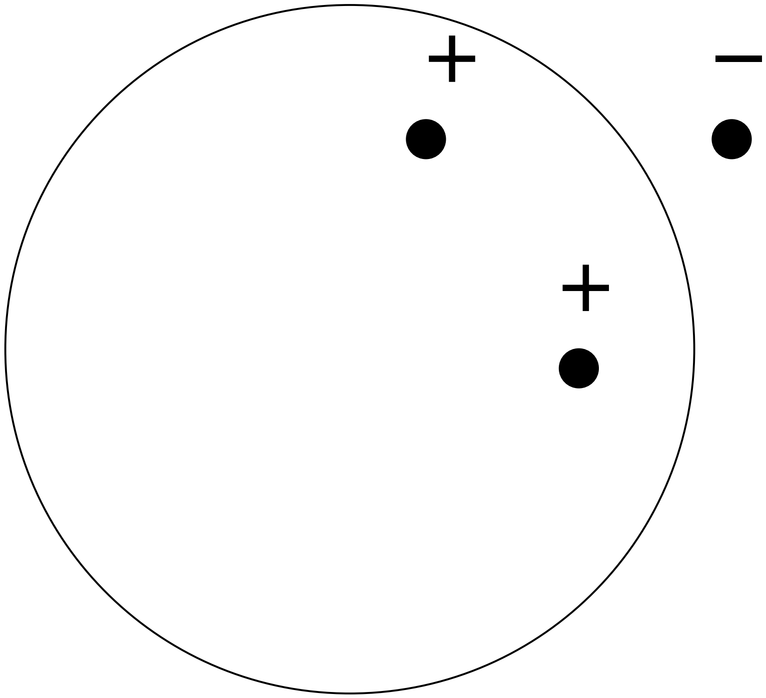
Example: Three instances dichotomy

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



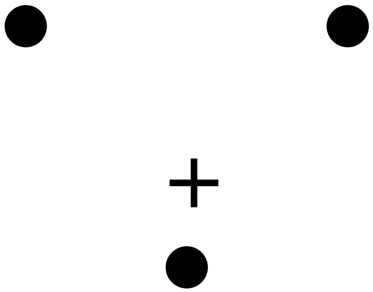
Example: Three instances

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



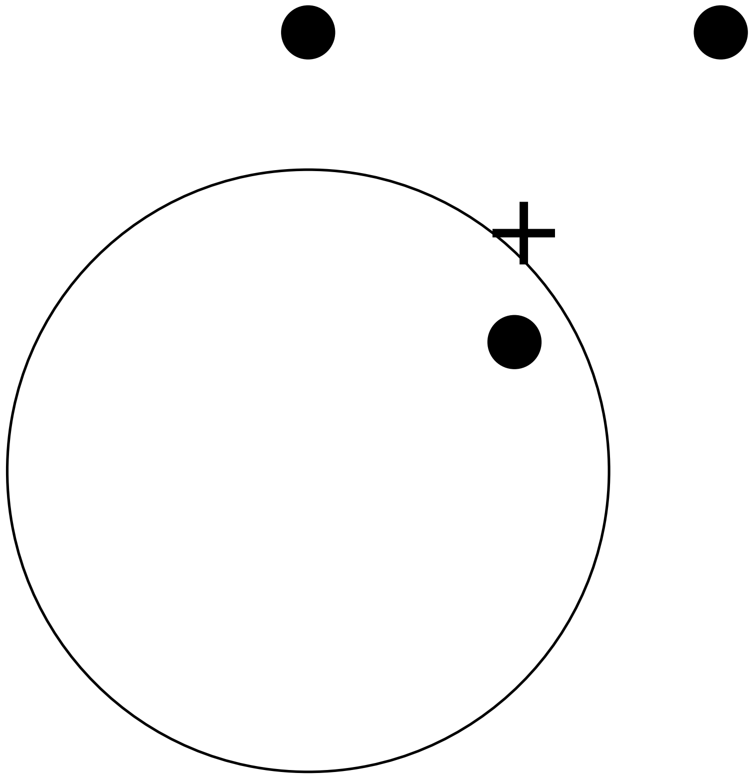
Example: Three instances dichotomy

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



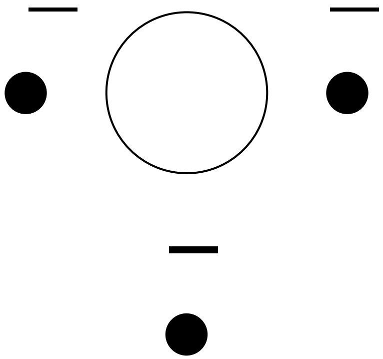
Example: Three instances

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



Example: Three instances dichotomy

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! but there is at least one set of 3 points that can be shattered (as we showed above)
- The *VC dimension of circles is 3*

Other examples of VC dimensions

- The VC dimension of 1-sided intervals is 1 and the one of 2-sided intervals is 2.
- The VC dimension of axis-aligned rectangles is 4.
- The VC dimension of halfspaces in $\mathbb{R}^d$ is $d+1$.
- Even though a pattern seems to emerge the *VC dimension is not related to the number of degrees of freedom*...
- The hypothesis space $\{x \mapsto \mathrm{sgn}(\sin(\theta x)) \ : \ \theta \in \mathbb{R}\}$ has one degree of freedom but its VC dimension is infinite.
- The VC dimension of convex polygons in $\mathbb{R}^2$ is infinite.

Growth function

- For any set of points $C = \{x_1, \cdots, x_m\} \subset \mathcal{X}$ we define the *restriction of $\mathcal{H}$ to C* by

$$\mathcal{H}_C = \{(h(x_1), h(x_2), \cdots, h(x_m)) \ : \ h \in \mathcal{H}\}.$$

- The growth function of $\mathcal{H}$ with m points is

$$\Pi_{\mathcal{H}}(m) = \max_{C=\{x_1,\cdots,x_m\}\subset\mathcal{X}} |\mathcal{H}_C|$$

- Thus the VC dimension is the largest m such that $\Pi_{\mathcal{H}}(m) = 2^m$.
- If $\mathcal{H}$ has VC dimension d_{VC} then $\Pi_{\mathcal{H}}(m) = 2^m$ for all $m \leq d_{VC}$ and $\Pi_{\mathcal{H}}(m) < 2^m$ if $m > d_{VC}$...

Sauer Lemma

- Sauer Lemma: If $\mathcal{H}$ has VC dimension d_{VC} then for all m we have

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d_{VC}} \binom{m}{i}$$

and for all $m \geq d$ we have

$$\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d_{VC}}\right)^{d_{VC}}$$

→ Up to d_{VC} the growth function is exponential (in m) and becomes polynomial afterward.

Growth function and VC dimension bounds

- For any δ, with probability at least $1-\delta$ over the choice of a sample of size m, for any $h \in \mathcal{H}$

$$R(h) \leq \hat{R}(h) + 2\sqrt{2\frac{\log \Pi_{\mathcal{H}}(2m) + \log\left(\frac{2}{\delta}\right)}{m}}.$$

If d is the VC dimension of $\mathcal{H}$, using Sauer lemma we get

- For any δ, with probability at least $1-\delta$ over the choice of a sample of size m, for any $h \in \mathcal{H}$

$$R(h) \leq \hat{R}(h) + 2\sqrt{2\frac{d_{VC}\log\left(\frac{em}{d_{VC}}\right) + \log\left(\frac{2}{\delta}\right)}{m}}.$$

Symmetrization Lemma

- One way to prove the previous bounds relies on the *symmetrization lemma*...

For any $t > 0$ such that $mt^2 \geq 2$,

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left(\mathop{\mathbb{E}}_{x\sim D}[h(x)] - \frac{1}{m}\sum_{i=1}^{m} h(x_i)\right) \geq t\right]$$
$$\leq 2\,\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m} h(x'_i) - \frac{1}{m}\sum_{i=1}^{m} h(x_i)\right) \geq \frac{t}{2}\right]$$

where $\{x_1, \cdots, x_m\}$ and $\{x'_1, \cdots, x'_m\}$ are two samples drawn from D (the latter is called a ghost sample).

Hoeffding inequality v2

- ...and a corollary of Hoeffding inequality

If $Z_1, \cdots, Z_m, Z'_1, \cdots, Z'_m$ are $2m$ iid random variables drawn from a Bernoulli, then for all $\varepsilon > 0$ we have

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i - \frac{1}{m}\sum_{i=1}^{m} Z'_i > \varepsilon\right] \leq \exp\left(\frac{-m\varepsilon^2}{2}\right)$$

Proof of the growth function bound

$$
\begin{aligned}
\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left(R(h)-\hat{R}(h)\right)\geq 2\varepsilon\right] &\leq 2\,\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left(\hat{R}'(h)-\hat{R}(h)\right)\geq\varepsilon\right]\\
&= 2\,\mathbb{P}\left[\max_{h\in\mathcal{H}_{\{x_1,\cdots,x_m,x'_1,\cdots,x'_m\}}}\left(\hat{R}'(h)-\hat{R}(h)\right)\geq\varepsilon\right]\\
&\leq 2\Pi_{\mathcal{H}}(2m)\max_{h}\mathbb{P}[\hat{R}'(h)-\hat{R}(h)\geq\varepsilon]\\
&\leq 2\Pi_{\mathcal{H}}(2m)\exp\left(\frac{-m\varepsilon^2}{2}\right)
\end{aligned}
$$

and the result follows by solving $\delta = 2\Pi_{\mathcal{H}}(2m)\exp\left(\frac{-m\varepsilon^2}{2}\right)$ for ε.

VC entropy

- The VC dimension is distribution independent, which is both good and bad (the bound may be loose for some distributions).
- For all m, the VC (annealed) entropy is defined by

$$H_{\mathcal{H}}(m) = \log \mathop{\mathbb{E}}_{C \sim D^m} |\mathcal{H}_C|.$$

- VC entropy Bound. For any δ, with probability at least $1-\delta$ over the choice of a sample of size m, for any $h \in \mathcal{H}$

$$R(h) \leq \hat{R}(h) + 2\sqrt{2\frac{H_{\mathcal{H}}(2m) + \log\left(\frac{2}{\delta}\right)}{m}}.$$

Rademacher complexity

- Given a fixed sample $S = \{(x_1, \cdots, x_m\}$ and a hypothesis class $\mathcal{H} \subset \{-1, 1\}^{\mathcal{X}}$, the empirical Rademacher complexity is defined by

$$\hat{\mathcal{R}}_S(\mathcal{H}) = \mathop{\mathbb{E}}_{\sigma_1, \cdots, \sigma_m} \left[\sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i) \right]$$

  where $\sigma_1, \cdots, \sigma_m$ are iid Rademacher RVs uniformly chosen in $\{-1, 1\}$.
- This measures how well $\mathcal{H}$ can fit a random labeling of S.
- The Rademacher complexity is the expectation over samples drawn from the distribution D:

$$\mathcal{R}_m(\mathcal{H}) = \mathop{\mathbb{E}}_{S \sim D^m} \hat{\mathcal{R}}_S(\mathcal{H})$$

Rademacher complexity and data dependent bounds

- For any δ, with probability at least $1-\delta$ over the choice of a sample S of size m, for any $h \in \mathcal{H}$

$$R(h) \leq \hat{R}(h) + \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}} \quad \text{and}$$

$$R(h) \leq \hat{R}(h) + \hat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}$$

- The second bound is data dependent: $\hat{\mathcal{R}}_S(\mathcal{H})$ is a function of the specific sample S drawn from D. Hence this bound can be very informative if we can compute $\hat{\mathcal{R}}_S(\mathcal{H})$ (which can be hard).

Rademacher bounds for kernels

- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a *bounded* kernel function: $\sup_x k(x,x) = B < \infty$.
- Let $\mathcal{F}$ be the associated RKHS.
- Let $M > 0$ and let $B(k,M) = \{f \in \mathcal{F} \ : \ \|f\|_{\mathcal{F}} \leq M\}$.
- Then for any $S = (x_1, \cdots, x_m)$,

$$\hat{\mathcal{R}}_S(B(k,M)) \leq \frac{MB}{\sqrt{m}}.$$

Conclusion

- PAC learning framework: analyze effectiveness of learning algorithms.
- Bias/complexity trade-off: sample complexity depends on the *richness* of the hypothesis class.
- Different measures for this notion of richness: cardinality, VC dimension/entropy, Rademacher complexity.
- The bounds we saw are worst-case and can thus be quite loose.

References to go further

- Books
  - *Understanding Machine Learning*, Shai Shalev-Shwartz and Shai Ben-David (freely available online)
  - *Foundations of Machine Learning*, Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar
- Lecture slides
  - Mehryar Mohri's lectures at NYU `http://www.cs.nyu.edu/~mohri/mls/`
  - Olivier Bousquet's slides from MLSS 2003 `http://ml.typepad.com/Talks/pdf2522.pdf`
  - Alexander Rakhlin's slides from MLSS 2012 `http://www-stat.wharton.upenn.edu/~rakhlin/ml_summer_school.pdf`