Outline for today

- PAC learning for finite H
- What if don't assume $c \in H$? (Agnostic learning)
- A new measure of H complexity: VC dimension
- Upper and lower bounds on sample complexity

Example: CNF

Suppose $C = H =$ conjunctions of boolean literals and their negations (e.g., $Tall \wedge \neg Young$)

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

$|H| =$?

How many examples if we have 12 boolean literals, and want 90% probability that we've have a hypothesis with accuarcy $\geq$ 85%

How does this compare with our 20-questions bound?

PAC learning CNF

Is there an algorithm to PAC learn CNF?

We showed we need only $m = poly(1/\epsilon, 1/\delta, n, size(c))$ examples

Must show we need only $poly(1/\epsilon, 1/\delta, n, size(c))$ time

So, just show that we need only $poly(...)$ time per example

Consider the algorithm $FindS$:

Initialize S to $l_1 \wedge \neg l_1 \wedge l_2 \wedge \neg l_2 \ldots$

For each positive training instance x

1. Remove any literal inconsistent with x

Example: Pairwise disjunctions of CNF

Suppose $H =$ pairwise disjunctions of h from H_{CNF}

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

$|H| =$?

How many examples if we have 12 boolean literals, and want 90% probability that we've have a hypothesis with accuarcy $\geq$ 85%

Example: Arbitrary disjunctions of CNF

Suppose $H =$ pairwise disjunctions of h from H_{CNF}

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

$|H| =$?

Summary

Useful general fact:

Any consistent learner can probably $(1-\delta)$ approximately (error $\leq \epsilon$) learn concept class C using $H \supset C$ from m examples, where

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

- This is an overestimate ($|H|e^{-\epsilon m}$ can be > 1)
  - Recall proof assumed $Pr(fail(h_1) \vee fail(h_2)\ldots) \leq \sum_i Pr(fail(h_i))$
  - Must handle all possible instance distributions $\mathcal{D}$
- Assumes error-free training data
- Assumes $H \supset C$
- What's a really bad distribution?

Agnostic Learning

Agnostic learning = don't assume $H \supset C$

Simply ask that learner output the hypothesis with *lowest* error over the training examples.

Suppose this hypothesis h has training error $r(h) =$ % of training examples misclassified.

How many training examples m suffice to assure (with probability $(1-\delta)$) that $r(h) \leq error(h) + \epsilon$?

$$m \geq \frac{1}{2\epsilon^2}(\ln|H| + \ln(1/\delta))$$

An Alternate Measure of Hypothesis Complexity

So far, we've used $|H|$ to characterize complexity of H

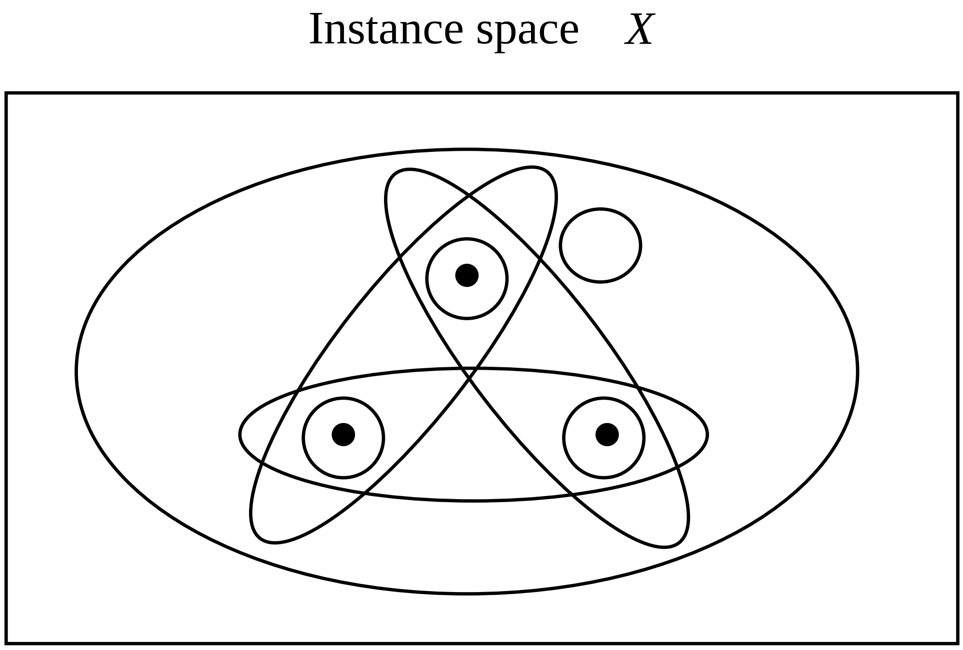
We can do better: Vapnik-Chervonenkis (VC) dimension

- Gives tighter bounds in many cases
- Handles some infinite H

Key idea:

- Don't count number of distinct hypotheses
- Count instead number of distinct instances that can be separated by H

Shattering a Set of Instances



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

H shatters S if it can distinguish every subset in the power set of S

VC 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 $VC(H) = \infty$.

$VC(H) = |S|$, where S is the largest subset of X that H can shatter

VC Dimension Examples

What is VC dimension of

- Intervals over the reals?
- Lines in the plane
- CNF over n boolean literals?

Sample Complexity using VC Dimension

Any consistent learner can probably $(1-\delta)$ approximately (error $\leq \epsilon$) learn concept class C using $H \supset C$ from m examples, where

$$m \geq \frac{1}{\epsilon}(4\log(2/\delta) + 8VC(H)\log(13/\epsilon))$$

Interesting to compare with our old bound:

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

Consider $H =$ lines in the plane, for instance.

Bound on Minimum Number of Examples

Lower bound on sample complexity. Consider any concept class C such that $VC(C) \geq 2$, any learner L, and any $0 < \epsilon < \frac{1}{8}$, and $0 < \delta < \frac{1}{100}$. Then there exists a distribution $\mathcal{D}$ and target concept in C such that if L observes fewer examples than

$$\max\left[\frac{1}{\epsilon}\log(1/\delta), \frac{VC(C)-1}{32\epsilon}\right]$$

then with probability at least δ, L outputs a hypothesis h having $error(h) > \epsilon$.

Interesting: with fewer examples, *no* learner can PAC learn non-trivial concept class C