Lecture 11: VC Dimension

- PAC learning in continuous spaces
- VC dimension; examples
- VC dimension of linear approximators
- VC dimension of neural networks

Reminder: PAC-learning in finite hypotheses spaces

We established a lower bound on the number of examples needed to learn a concept with error at most ϵ and probability at leats $(1-\delta)$:

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

What if $|H|$ is infinite?

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 example seen
- Clearly the processing time per example is polynomial. but how many examples do we need?
- Our previous result is useless, since the hypothesis class is infinite.

Sample complexity of learning an interval

- 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!
- What is the probability of seeing m examples *outside* of $[c, a]$?

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

- If we want

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

- We get (using our magical inequality from last time):

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

- 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"?

Learning two-sided intervals

- We proceed just like before, but now we can make errors on two sides of the interval.
- The error of h is $P(h \neq C) = P(E)$.
- $P(E) \geq \epsilon$ only happens with probability $\delta \leq (1-\epsilon)^m$ on either side of the interval
- Thus, with probability $\delta \leq 2(1-\epsilon)^m$ the error will be $\geq \epsilon$.

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

- We want a definition of the "degrees of freedom" of a hypothesis space

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 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.

Example: Three instances

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

How about four instances?

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 hypothesis can exhibit
- This is, in some sense, the number of "effective degrees of freedom"

Example: VC dimension of circles?

Let H be the set of circles. Based on our previous analysis, $VC(H) = 3$, since 3 points can be shattered but not 4.

What if H are rectangles?

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 $n + 1$.

Sample complexity and the VC dimension

Using $VC(H)$ as a measure of the complexity of H (instead of $\ln |H|$), Blumer et. al (1989) derived an alternative bound for the number of examples required:

$$m \geq \frac{1}{\epsilon}(4 \log_2 \frac{2}{\delta} + 8VC(H) \log_2 \frac{13}{\epsilon})$$

Compare this to:

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

Lower bound on sample complexity

Theorem: Given a concept class C $|C| \geq 2$, then for any learning algorithm A, there exists a distribution D over the concept class C such that the expected error of A is $> \epsilon$ if A sees less than

$$\frac{VC(C) - 1}{8\epsilon}$$

examples

Comments on the sample complexity results

- VC dimension gives us both a lower and upper bound on the number of examples m, and both depend on $\frac{1}{\epsilon}$.
- The upper bound depends on $VC(H)$ (the VC dimension of the hypothesis space) and has an additional factor of $\log \frac{1}{\epsilon}$.
  This is a lot tighter than the previous bound that we had, which depended on $\log |H|$*!* Why?

  If $VC(H) = d$, then H can shatter d instances, which requires 2^d distinct hypotheses. Hence $d \leq \log_2 |H|$.

More comments

- The lower bound depends on the VC dimension of the concept space, which describes how hard it is to learn a given concept (given any hypothesis class).
- Does it mean that we cannot learn with hypotheses of infinite dimension?

  No! Depends on the concept class
- Are "complicated hypotheses "bad"?

  Not necessarily! But expect a need for lots of data in order to learn complex concept classes

VC theory for perceptron networks

Let G be a directed layered graph with n input nodes, s internal nodes and 1 output node, with each internal node having at most r inputs. Let C be a concept class of VC dimension d, corresponding to what can be represented by the internal nodes. Let C_G be the set of functions that can be represented by D.

Then $VC(C_G) \leq 2ds\log(es)$.

Immediate consequence: for networks of perceptrons, the VC dimension is:

$$VC(C_G) \leq 2(r+1)s\log(es).$$

Example: VC dimension of perceptrons

Consider training a network of 5 perceptrons, with 10 inputs, with error less than 10% and reliability at least 95%.

If we use the VC dimension formula above, we come up with
≈ 22000 *examples needed!*

This is way too high compared to empirical data...

Applying VC theory to feed-forward networks

Let F be the class of functions that can be computed by feed-forward nets defined on a fixed underlying graph G with E edges and $N \geq 2$ linear threshold nodes. Let $W = E + N$ be the total number of edges in the network (why?).

Then it can be shown that $VC(F) \leq 2W \log(eN)$.

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.