VC Dimension Review

The purpose of this document is to review VC dimension and PAC learning for infinite hypothesis spaces.

Previously, in discussing PAC learning, we were trying to answer questions about how difficult it might be to learn a particular concept, and how long it would take a learner to do so. In that discussion, we had one sort of big problem: the hypothesis space H had to be finite. This time we hope to discuss PAC learning for infinite hypothesis spaces. To do this we will need to introduce the concept of VC dimension, along with a bunch of other definitions to support it.

In what follows, we are considering a binary classification problem from the space of instances X. So, each hypothesis $h \in H$ should split X into two sets:

$$\{x \in X \mid h(x) = 1\} \quad \text{and} \quad \{x \in X \mid h(x) = 0\}.$$

In this situation, we will say that a dichotomy has been imposed on X by h. Note that either one of the sets above might be the empty set -- this is still acceptable.

Definition: A set of instances $S \subset X$ is *shattered* by H iff for every possible dichotomy of S, there exists some hypothesis $h \in H$ that is consistent with this dichotomy.

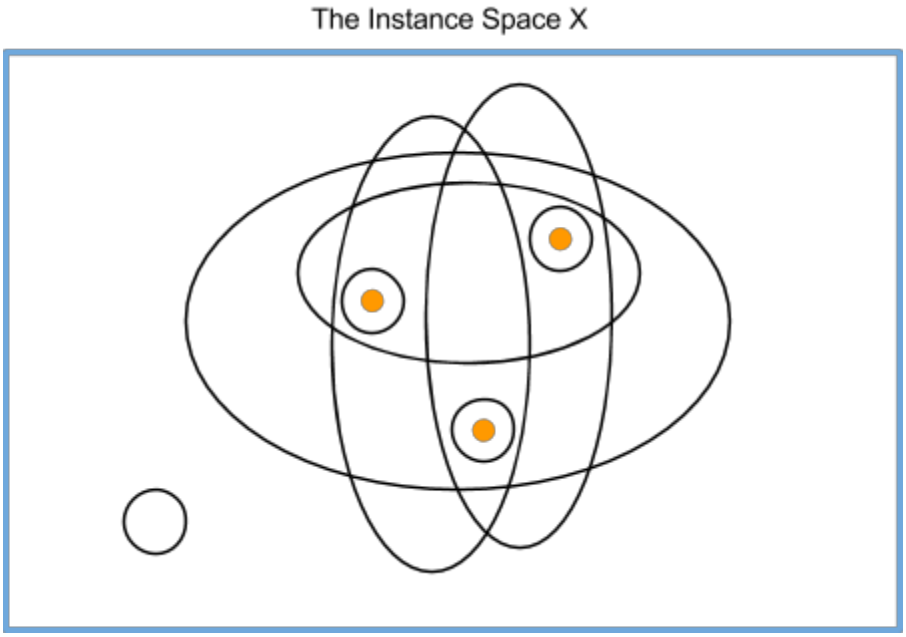
In other words, S is shattered by H if there are enough hypotheses in H to agree with every possible labeling of S. To get a better understanding of this, let's look at some examples.

We will start with the simplest example where S is one point, x_1. In this case S has only two possible labelings: $x_1 = 0$ and $x_1 = 1$. So now we can easily come up with a hypothesis space with a hypothesis for each of the labelings, i.e. a hypothesis space that shatters S. Let's try

$$H = \{h_1(x) = 1,\ h_2(x) = 0\}$$

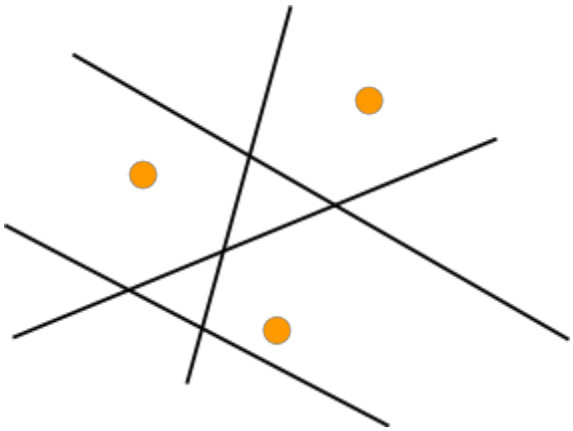
for example. In this H the first hypothesis labels everything as 1, and the second labels everything as 0. The first hypothesis is correct for the labeling $x_1 = 1$, and the second hypothesis is correct for the labeling $x_1 = 0$.

Here is a figure showing a set of three instances shattered by eight hypotheses.



Here the hypotheses are indicated by by the ellipses, and classification is determined by being inside or outside of an ellipse. For every possible labeling of the points with 0's and 1's, there is an ellipse that agrees with that labeling.

There are also several good examples from the lectures. Michael asks us to find the largest number of points in the plane that can be shattered by the hypothesis space of lines. In this example, the lines divide the plane into 0's and 1's. It turns out that we can find an arrangement of 3 points that can be shattered by lines, but no arrangement of 4 points that can be.



The image above shows a three point set shattered by lines in the plane.

This leads naturally to the following question: for a given instance space X, and hypothesis space H, what is the largest subset of X that can be shattered by H? The size of this subset has special significance, and is termed the VC dimension of H, which is denoted as $VC(H)$.

Definition: Given an instance space X, the Vapnik-Chervonenkis dimension of H over X is the size of the largest finite subset of X that can be shattered by H. If arbitrarily large subsets of X can be shattered, then $VC(H) = \infty$.

Despite what is written in the lectures, it is best not to think of VC dimension as a measure of the power of a hypothesis space. Instead, it is better to think of VC dimension as a measure of the *complexity* of a hypothesis space. This distinction is important for heuristic reasons later on.

It is important to remember that $VC(H)$ is the size of the largest subset of X that can be shattered by H. In the example above with the lines in the plane, even though $VC(H) = 3$, we can cook up an example of 3 points in the plane that can not be shattered by H. For example, the three colinear points in the plane shown below can not be shattered by lines.

In general, it is easy to find a lower bound m for $VC(H)$ since we only need to find one example of set of m points that can be shattered by H. It is harder to find an upper bound n for $VC(H)$, because we need to *prove* that *no* set of n points can be shattered by H.

Example: Let's suppose the hypothesis space H is the set of single intervals in the real line R, and the sample space X is the set of points on the line. Then we can shatter sets of two points with H, so $VC(H) \geq 2$, but it can be shown that no set of three points can be shattered by H, so we have $VC(H) = 2$.

Example: Now let's suppose that the hypothesis space H is the set of all convex polygons in the plane, with the sample space consisting of points in the plane. Then for any number n, we can choose a polygon with points being the n vertices. We can deform the polygon slightly to leave out any of the vertices, giving an example of an n point set shattered by H. Since n can be arbitrarily large, we see that $VC(H) = \infty$.

We also note that if the hypothesis space is finite, then there is a relationship between the size of the hypothesis space H and $VC(H)$. Suppose $VC(H) = d$. For d points from X, there are 2^d possible labelings of these points. Each of these labelings requires a separate hypothesis in order for the subset of d points to be shattered. Thus $2^d \leq |H|$. It follows that

$$VC(H) = d \leq \log_2|H|.$$

Now, returning to PAC-learning and ε-exhaustion, we note that it is possible (but apparently very difficult) to derive a bound for the number of training samples needed to ε-exhaust the version space of

H with probability $(1-\delta)$. This bound is given by:

$$m \geq \frac{1}{\varepsilon}(8\, VC(H) \log_2(13/\varepsilon) + 4 \log_2(2/\delta)) \quad ,$$

which is analogous to the bound we saw in the case where H was finite (recall, that bound was $m \geq \frac{1}{\varepsilon}(\ln(H) + \ln(1/\delta))$). In the expression above, we see that $VC(H)$ must be finite for us to bound the number of training examples needed to ε-exhaust the version space of H. This is consistent with the theorem Michael provides in the lecture, which states:

Theorem: A concept class C is PAC learnable if and only if $VC(C) < \infty$.

Finally, we recall the heuristic that $VC(C)$ measures (in this case) the complexity of the concept class C. This heuristic is also consistent with the theorem above -- if the concept class is too complex, i.e. $VC(C) = \infty$, then roughly speaking, we will have trouble choosing with probability $(1-\delta)$ a hypothesis that has low error, and so the concept class C will not be PAC learnable.