Points, Probabilities, and p-Adics: Determining VC Dimension/Density With Statistical Sampling

Matthew Jordan • jordanml@mcmaster.ca
Supervised by Dr. Deirdre Haskell

McMaster University

Abstract

Every collection of mathematical objects has a pair of properties called the VC dimension and VC density. VC dimension/density come from machine learning theory and measure the complexity of a set. This project looks specifically at the VC dimension/density of sets in an alternate number system called the p-adics. The p-adic numbers arise by redefining the notion of distance in an unintuitive way, leading to some unusual properties. For example, in the 2 -adics,

$$1 + 2 + 4 + 8 + \ldots = -1$$

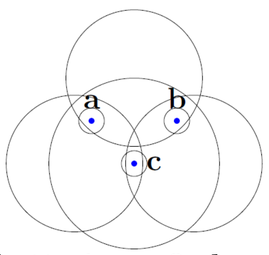
Our approach uses statistical methods and tests hypotheses to estimate an optimal bound on the VC dimension/density of p-adic sets.

VC Dimension

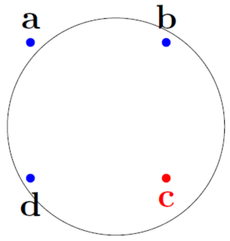
The VC Dimension of a family of sets – whether they be intervals in the line, disks in $\mathbb{R}^2$, or p-adic annuli (more to come there) – measures its capacity to *shatter* a set of points. A set A is shattered by a family of sets $\mathcal{F}$ if:

$$\text{For every } B \subset A \text{ there is } F \in \mathcal{F} \text{ such that } B = F \cap A.$$

That is, the set is shattered if every subset can be *singled out* by an element of $\mathcal{F}$. Below is an example with disks in the plane. Every subset of the 3-element set can be singled out, while the "corners" of the 4-element set cannot.



Shattering a 3-element set with disks



4-element set cannot be shattered by disks

We thus define the VC dimension of a family of sets as follows:

$$\dim_{VC}\mathcal{F} = \begin{cases} \max\{|A| : A \text{ is shattered by } \mathcal{F}\} & \text{if it exists.} \\ \infty & \text{otherwise.} \end{cases}$$

So, using this definition, the VC dimension of disks is 3.

VC Density

The shatter function, $\pi_{\mathcal{F}}(n)$, measures how many subsets of a set of size n are singled out by $\mathcal{F}$. Notice that since all sets have 2^n subsets, then $\pi_{\mathcal{F}}(n) = 2^n$ when n is less than the VC dimension of $\mathcal{F}$. By the fundamental Sauer-Shelah lemma:

$$\pi_{\mathcal{F}}(n) \leq \sum_{i=0}^{d} \binom{n}{i} \in O(n^d).$$

This polynomial bound is not necessarily optimal, however. So, the VC density of a family is:

$$\text{dens}_{VC}\mathcal{F} = \begin{cases} \inf\{r \in \mathbb{R}^+ : \pi_{\mathcal{F}}(n) \in O(n^r)\} & \text{if it exists.} \\ \infty & \text{otherwise.} \end{cases}$$

The p-Adics

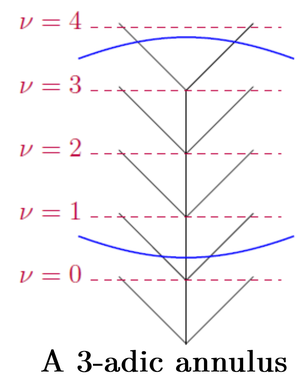
The p-Adic numbers are obtained by an alternative interpretation of numerical size and distance. Here's how: First, pick a prime number p and define a unary function called the *valuation* as follows:

$$\nu_p(x) = \max\{n \in \mathbb{N} : p^n \mid x\}$$

For example, $\nu_3(36) = 2$, because $3^2 = 9$ divides 36, but $3^3 = 27$ does not.
In our familiar Euclidean space, the *norm or* size of a number is simply its magnitude or absolute value. In the p-adics, we assign a new norm based on the valuation:

$$|x|_p = p^{-\nu_p(x)}$$

This new interpretation of size means that a number is "small" if a large power of p divides it. To see why this yields strange results, consider the a geometric series $\sum_{n=0}^{\infty} r^n$, which we know converges to $\frac{1}{1-r}$ when $|r| < 1$. But given our new definition of size, $|2|_2 = \frac{1}{2}$, which means that $\sum_{n=0}^{\infty} 2^n = 1 + 2 + 4 + 8 + \ldots = \frac{1}{1-2} = -1$. Unusual indeed!
The p-adics can be visualized with a tree structure, branching off at each valuation.



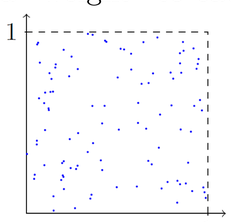
A 3-adic annulus
$\{x \in \mathbb{Q}_3 : 1 \leq \nu_3(x) \leq 3\}$

Using ε-Nets

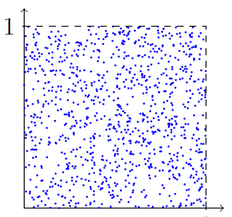
It's generally quite difficult to directly calculate the VC dimension/density of p-adic sets, since they are tough to visualize and behave somewhat erratically.
Fortunately, there exists an estimation tool called ε-nets. We call a set an ε-net for $\mathcal{F}$ when:

$$\text{For every } F \in \mathcal{F}, \text{ if } \mu(F) > \epsilon, \text{ then } F \cap S \neq \emptyset.$$

The function μ is called a *measure*, and simply assigns a "weight" to each member of the family of sets.



A non-ε-net



An ε-net

The disk cannot fit in the set on the right without intersecting it.

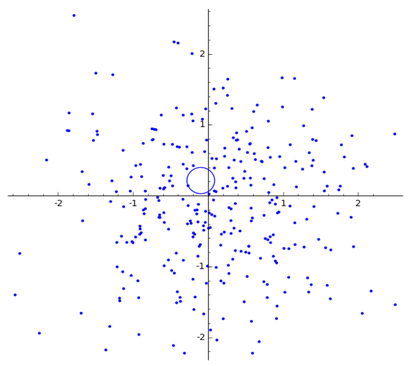
Statistical Sampling

We use hypothesis testing to determine the VC dimension of a family of sets:

1. Hypothesis: d $= \dim_{VC}\mathcal{F}$
2. Fix $\varepsilon > 0$ and determine the measure of the set of non-ε-nets.
3. Use a series of ε-net theorems to determine the probability of seeing that proportion of non-ε-nets
4. If the probability is less than 5%, reject the hypothesis

The success of this method is largely dependent on the type of measure used. The most effective is a binormal measure, in which the "weight" of a set is greatest toward the mean of the distribution.

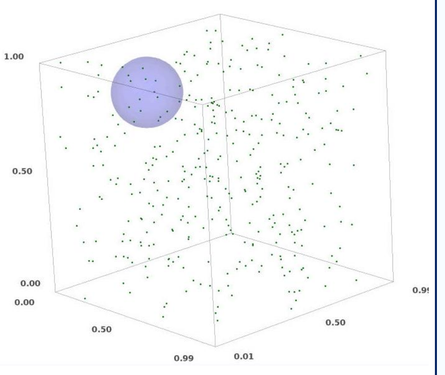
Hunting for Non-ε-Nets



A 250-Point Non-ε-Net

In order to test hypotheses, it's necessary to find non-ε-nets for various families of sets using different measures. These include uniformly distributed and binormal points in both the plane, and three-space. This involves tailoring an optimization algorithm to find a disk/ball that can squeeze into randomly generated sets.

Among the other interesting sets that can be explored using this method are multi-dimensional balls, p-adic annuli, and linear combinations of p-adic-defined functions.



A 300-Point Non-ε-Net in $\mathbb{R}^3$

Acknowledgements

I would like to thank the McMaster Arts & Science Program for the funding of this USRA project, and Dr. Deirdre Haskell and Dr. David Lippel for their supervision, support, and guidance throughout the research process. I'm also especially grateful to Kaitlyn Chubb, with whom I collaborated on the entirety of this project.