Epsilon Nets and VC Dimension

Subhash Suri

November 21, 2019

1 Sampling

- Sampling is popular and powerful idea applied widely in many disciplines, including CS. There are at least two important uses of sampling: *Estimation and Detection.*
- You have almost surely encountered polls reported by CNN or NYT etc where sampling is used to estimate the size of a particular group in the larger population, such as likely support for a presidential candidate (with high confidence). Similar polls are used by industry to gauge potential market share of a many new product.
- Similarly, random traffic checks are a form of sampling, where the goal is to sample (by observing cars during some time window at some particular spot) traffic in the hope of catching speeders or drunk drivers. This is sampling for *detection*; target here are frequent speeders or drinkers.
- In this lecture, we will discuss the concept of VC dimension and ε-nets, which provide theoretical basis for sampling in multi-dimensional data with an abstract and therefore widely applicable framework.
- The mathematical framework uses the idea of a *range space*, which is a pair of sets $(A, \mathcal{R})$, where members of A are "points" and members of $\mathcal{R}$ are subsets of A.
- To build some intuition for our framework and the kind of questions we are interested in, let us first consider a toy example.
- Suppose A is a population of consumers, and $G \subset A$ be a consumer group that frequently eats a certain food, say, chocolates. In general, there may be multiple (overlapping) such groups $G_1, G_2, \ldots, G_k$, which *hidden* habits or preferences we are interested in learning.

- In order to estimate what fraction of A belongs to a group G_1, we can draw a random sample $S \subset A$, and count how many of consumers in S eat that food or have that preference.. Our belief is that

$$\frac{|G_1 \cap S|}{|S|} \approx \frac{|G_1|}{|A|}$$

and statistics tells us to what extent that is justified. Many consumer surveys are indeed based on this.

1.1 ε-Nets

- But suppose we are interested in a slightly different estimation problem. *We want to know the* **most popular** *food products, say, the ones frequently eaten by more than ε fraction of consumers, for some $0 < \varepsilon < 1$.*

- Of course, we could draw a sample for each group G_i *separately*, estimate its size, and identify those that are large. But when there are many groups, this requires too many (infeasibly many) samples.

- Instead, what we want to know whether there exists a *single sample* N of consumers that represents *all popular groups* simultaneously! In other words, is there a *universal* sample that can be used to estimate all target groups? Formally, we want a subset $N \subset A$ so that

$$\forall i,\ \frac{|G_i|}{|A|} > \varepsilon \quad \Longrightarrow \quad G_i \cap N \neq \emptyset$$

- Such a subset is called an *epsilon net*, or ε-net. Obviously, $N = A$ is such an epsilon net, for all $eps < 0$, but we want something much smaller.

- With this semi-informal motivation, let us now begin to formalize the concepts more rigorously.

- Let $\mathcal{X}$ be a possibly infinite set, and $\mathcal{R} \subset 2^{\mathcal{X}}$. The pair $(\mathcal{X}, \mathcal{R})$ is called a *range space*, with $\mathcal{X}$ its points and $\mathcal{R}$ its ranges.

- Given a range space $(\mathcal{X}, \mathcal{R})$, let $A \subset \mathcal{X}$ be a finite subset, and $0 < eps < 1$. Then a subset $N \subset A$ is called an ε-net of A w.r.t. to $\mathcal{R}$ if

$$\forall\ r \in \mathcal{R}, |r \cap A| > \varepsilon |A| \quad \Longrightarrow \quad r \cap N \neq \emptyset$$

- The definition is not so easy to appreciate so let us do a few examples.

  1. (R^1, J): R^1 is 1-dim line, and $J = \{[a, b] \mid a \leq b \in R^1\}$
  2. (R^d, H_d): H_d are (closed) halfspace in R^d.
  3. (R^d, B_d): B_d are (closed) balls in R^d.
  4. (R^d, C_d): C_d are convex subsets in R^d.
  5. $R^2, RECT)$: RECT are axis-aligned rectangles in R^2.
  6. So, for instance, ε-net for (R^d, B_d) asks: given any set of n points in R^d, what is the smallest sample so that any ball containing at least εn points also contains one of the samplr points?

1.2 Basic Sampling

- Let us start with a simple exercise showing that large ranges are easy to catch. Assume that we have a range r s.t. $|r \cap A| > \varepsilon |A|$, for some $\varepsilon < 1$.

- Draw a random sample $S \subseteq A$ of size $s = |S|$, by drawing s items uniformly at random (with replacements).

- What is the probability S fails to *detect* r, that is, the prob. that $r \cap S = \emptyset$.

- Define $p = |r \cap A|/|A|$ where we know that $p > \varepsilon$. Then, in each draw of A, we have prob. p that a point from range r is chosen. Thus,

$$Pr[S \cap r = 0] \;=\; (1-p)^s \;<\; (1-\varepsilon)^s \;\leq\; e^{-\varepsilon s}$$

- Thus, if $s = 1/\varepsilon$, this prob. is at most $e^{-1} \approx 0.368$. And if we choose $s = c/\varepsilon$, for some $c > 1$, the failure prob. decreases exponentially as e^{-c}.

- As an example, if $|A| = 10^5$, and $|r \cap A| < 100$, that is, r contains at least 1% of the population, a sample of size $s = 300$ has failure prob. of $\leq e^{-3} \approx 0.05$. In conclusion, if we draw a random sample of size about $O(1/\varepsilon)$, it catches at least one point of the range r almost surely.

- The important question is: *what does this sample S say about some other range $r' \neq r$?* Recall our original motivation: we want to "hit" all big ranges, not just one specific one.

- It seems reasonable to claim that since S was drawn at random, any other range should also behave just like r w.r.t. to these samples. After all, samples in S didn't know which specific range r we are interested in. Thus, as long as r' is also large, meaning $|r' \cap A| \geq \varepsilon n$, why shouldn't we expect $r' \cap S \neq \emptyset$?

- This common sense reasoning however is completely wrong. The trouble is that when we analyze the performance of a sampling method, we need to be careful about in which order we choose our objects.

- The sampling we set out to do worked as follows:

    1. First, we fixed a range r, which we want to hit.
    2. Then, we select a sample, where we choose items of A uniformly at random, one at a time, until s items have been selected.
    3. Lastly, we argue that at least one of these chosen items also lies in r.

- But if we *first fix the sample* S, how do we argue about some other range r'? In particular,

    1. What process do we use to choose r'?
    2. Should it be chosen at random from $\mathcal{R}$?
    3. Or, should an adversary decide which r' to target?
    4. In fact, if our goal is to satisfy *all large ranges*, then an adversarial such range is perfectly logical.
    5. But here is the kicker: S only has $O(1/\varepsilon) \ll n$ points, so vast majority of elements of A are not in S. Thus, we can easily choose a set r' of size εn that is disjoint from S. In fact, there are a huge number of such subsets!
    6. Therefore, we should not expect S to hit another range r' that we may be interested in.

- So, the sample S we prepared is only a good ε-net for one specific range r, and we cannot afford to maintain a separate sample for each r since there are way too many ranges.

- Hopefully, this helps explain why ε-net is not a trivial idea: a single (tiny size) sample will hit every large range in $\mathcal{R}$. **But** to be able to do this, the range space must have some nice structure: they can't just be arbitrary subsets of A.

- Indeed, if we go back and look at the range spaces mentioned earlier, they all have quite a bit of structure. For instance, in the case of rectangles in 2D: the number of different subsets of A that can be realized is much much smaller than the full power set 2^A. In fact, intuitively, there are only $O(n^4)$ subsets. Similarly, for ranges created by circles, halfspaces..

- **Connection to Machine Learning.** $\varepsilon-$net and VC dimension plays an important role in machine learning. ML algorithms learn a *model* of data from a particular class, called its *hypothesis class*, which is essentially the range space. The VC dimension will be the expressive power of the hypothesis class. For instance, perceptron uses hyperplanes (linear function) range space. More complex forms use boolean functions over hyperplanes etc.

2 Shattering and VC Dimension

- We now formalize what makes range spaces such as hyperplanes or balls amenable to small epsilon nets. The key is the *combinatorial complexity* of how many subsets of A can be generarated by a range. Turns out that there is a very natural quantity attached to each range space, which is called its *VC Dimension*, named after their its inventors Vapnik and Chervonenkis.

- **Projection:** Given a range space $(\mathcal{X}, \mathcal{R})$, and a set $A \subset \mathcal{X}$, define

$$\mathcal{R}_A \quad = \quad \{r \cap A \mid r \in \mathcal{R}\}$$

  That is, $\mathcal{R}_A$ is the family of subsets of A that can be produced by ranges of $\mathcal{R}$. Notice that *even if $\mathcal{R}$ is infinite, $\mathcal{R}_A$ has size at most 2^n, if $|A| = n$.*

- Many simple ranges cannot produce the full power set. For instance, if $\mathcal{X}$ is the real line, and ranges are *intervals*, then for any set A of n points we have $|\mathcal{R}_A| = O(n^2)$.

- **Shattering:** We say that A is *shattered* by $\mathcal{R}$ if $\mathcal{R}_A$ has size 2^n. (That is, all possible subsets of A can be produced by the family of ranges $\mathcal{R}$.)

- VC**-Dimension.** The VC dimension of $(\mathcal{X}, \mathcal{R})$ is the cardinality of the largest subset $A \subset \mathcal{X}$ shattered by $\mathcal{R}$. If sets of arbitrarily large size can be shattered, then VC dim is ∞.

- Examples.

1. Range space (R, J) of intervals. Clearly any two points can be shattered, since each of the four subsets $\emptyset, \{a\}, \{b\}, \{a, b\}$ can be realized by intersecting A with some interval.

   On the other hand, we claim that no 3-point set can be shattered. In particular, if $a < b < c$, then there is no interval that contains a and c but not b. So, (R, J) has VC dim 2.

2. Similarly, VC dim of halfplanes in 2D is 3. Any three non-colliear points can be shattered, but no 4-point set can be shattered.

   To see that consider the convex hull of the four points. If CH is a triangle, say, (a, b, c), with fourth point d inside, then no range can include a, b, c but not d.

   If all four are on the hull, then no range can include a, c and exclude b, d.

3. What about the VC dimension of $(R^2, CONVEX - POLYGON)$? Supriringly, it is ∞. For any n, consider a convex n-gon. Then, any subset can be generated by a convex polygon whose vertices are at those points.

- VC Dimension shows that set systems can be defined in very general terms, and as long as they do not shatter large subsets, their "effective" dimension is small, for the purposes of ε-nets. In constrast, using simple geometric shapes such as rectangles or hyperplanes to form ranges can be very limiting in some applications. For rectangular ranges, it is easy to see that n points can define only $O(n^4)$ distinct ranges, based on the points that constrain the four sides. But for more complex shapes, it is difficult to derives such bounds. In those cases, it is easier to argue using VC dimension—we just need to bound their shattering.

3 Sauer's Lemma and ε-Net Size

- Sauer's Lemma shows the intimate connection between VC dim of a range space and the number of different subsets it can produce.

  Sauer's Lemma: Let $(\mathcal{X}, \mathcal{R})$ be a range space of VC dim d. Then, for all $A \subset \mathcal{X}$ with $|A| = n$,

  $$|\mathcal{R}_A| \quad \leq \quad \sum_{i=0}^{d} nchoosei$$

  That is, if a range space has VC dim d, then it can only generate polynomially many distinct subsets of A as projections of $\mathcal{R}$.

- **Epsilon Net Theorem.** Let $(\mathcal{X}, \mathcal{R})$ be a range space of VC dim d, and let A be any finite subset of $\mathcal{X}$. Then, any random sample S of A of size

$$O\left(\frac{d}{\varepsilon}\log\frac{1}{\varepsilon} \;+\; \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$$

  is an ε-net for $(A, \mathcal{R})$, with prob. $\geq 1-\delta$. Assuming d and δ are constant, the ε-net has size $O(1/\varepsilon \log 1/\varepsilon)$.

- Proof uses standard sampling techniques, and Chernoff bounds, but has quite a few non-trivial technical details. See Haussler-Welzl paper, or ...

4 Applications of ε-Net

- Used in many divide-and-conquer algorithms in geometry.
- Geometric set covering and hitting sets.
- r-cuttings.
  1. Let H be a set of n hyperplanes in R^d. We want to divide R^d into simplicies, none of which is cut by too many planes of H. We choose a parameter $\varepsilon > 0$.
  2. A set C of simplicies is called an ε-cutting for H if
     - the union of the simplicies is R^d and their interiors are disjoint
     - the interior of any simplex is intersected by at most εn hyperplanes of H.
     - example use of cuttings in point location in the arrangements of hyperplanes.
  3. **Thm.** There exist ε-cuttings of size $O((1/\varepsilon)^d)$, which is optimal.

5 VC Dimension of Hyperplanes and Radon's Theorem

- It is easy to argue that the VC dim of 2-dim halfplanes is 3. Any 3 points shattered, and no four shattered, using the same argument as rectangles. What is the VC dim of d-dim Euclidean halfspaces? A direct argument gets tedious quickly, but fortunately the following geometric lemma gives an elegant proof.
- **Radon's Theorem.** Any set A of $n+2$ points in R^d can be partitioned into A_1, A_2 s.t. $CH(A_1) \cap CH(A_2) \neq \emptyset$.
- Assuming Radon'e Theorem, it is much easier to prove that $VCdim(R^d, H_d) = d+1$.

1. First, any set of $d+1$ points forming vertices of a simplex can be easily shattered by halfspaces, so the $VCdim \geq d+1$.
2. To show that no set of $d+2$ points can be shattered, perform a Radon partition of the points into A_1 andn A_2. Since their CHs intersect, any halfspace containing A_1 must contain at least one points of A_2, and therefore there can be no h with $h \cap A = A_1$.
3. In fact, for this particular range space, the bound of Sauer's Lemma is exactly achieved (recall our claim about the size of hyperplane arrangements).

- In fact, we also have $VCdim(\mathcal{R}^d, B_d) = d+1$.

6 Small Epsilon-Nets with $\log n$ factor size, but a simpler proof

- One sample of size s fails to hit a particular range r with prob. $\leq e^{-\varepsilon s}$. Since there are only n^d ranges, the union bound requires that we just need to ensure

$$n^d e^{-\varepsilon s} < 1 \quad \Longrightarrow \quad n^d \ < \ e^{\varepsilon s} \quad \Longrightarrow \quad s \ > \ \frac{d \log n}{\varepsilon}$$

- Thus, we can have a small ε-net, but it depends on the size of input range space. By constrast, the ε-net theorem guarantees a net *independent* of input range space size.