Lecture 23   11/23/2025

Sampling in Geometric Range Spaces

Set systems arise in many applications. A set system consists of a pair (P, R) where P is a set and R is a collection of subsets of P. When P is finite we have a finite set system. Sometimes finite set systems are thought of as hypergraphs with P as vertices and each $e \in R$ as a hyperedge.

Here we will be concerned with set systems that arise in geometric settings where P is typically all of $\mathbb{R}^d$ for some dimension d, or a finite subset of $\mathbb{R}^d$. We also consider $\mathcal{R}$ to be sets that are induced by "structured" shapes that intersect with P.

Examples of shapes include convex sets such as intervals, disks, half spaces, convex polygons etc.

In the geometric setting $(P, \mathcal{R})$ are often called range spaces and each $r \in \mathcal{R}$ is called a range. Typically we associate r with a shape such as an interval I and ~~the~~ then

$r = I \cap P.$

Geometric range spaces have additional properties that lead to a number of applications and the notion of VC-dimension, ε-sample theorem, ε-net theorem and others have had striking influence on many areas, in particular machine learning where the notion of VC-dimension arose.

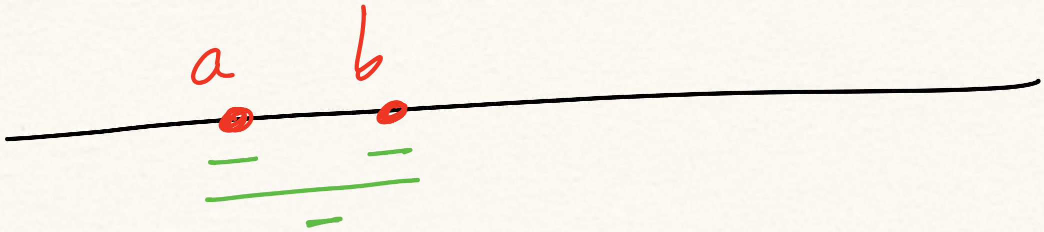
VC-Dimension

VC-dimension of a set system is one important measure of the complexity of a set system.

<u>Defn:</u> Let $(P, \mathcal{R})$ be a range space. A finite subset $Q \subseteq P$ is said to be <u>shattered</u> by $\mathcal{R}$ if $\forall\, Q' \subseteq Q$ $\exists\, r \in \mathcal{R}$ such that $Q' = Q \cap r$.

In other words $\{ Q \cap r \mid r \in \mathcal{R} \}$ $= 2^{Q}$ the powerset of Q.

<u>Ex:</u> Suppose P is the real line and $\mathcal{R}$ is the collection of all closed intervals.



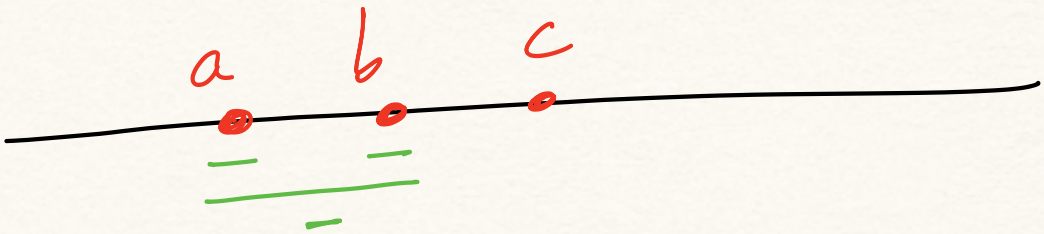
It can be seen from the figure that $Q = \{a, b\}$ can be shattered by collection of intervals.

<u>Defn</u>: The VC-dimension of a set system (P, R) is the maximum cardinality of a finite set $Q \subseteq P$ such that Q is shattered by R.

<u>Ex</u>: Let $P = \mathbb{R}$ and R be the collection of intervals. Then VC-dim $= 2$.

Why. We saw that it is at least 2. Can it be ≥ 3.

Suppose $Q = \{a, b, c\}$ where $a < b < c$



Can we get the set $\{a, c\}$ as an intersection of $\{a, b, c\}$ and an interval? No.

Ex: $P = \mathbb{R}^2$ the 2-d plane and

$$\mathcal{R} = \{ D \mid D \text{ is a closed disk in the plane} \}$$

VC-dimension is 3.

3 points can be shattered but not 4.

Ex: $P = \mathbb{R}^d$ and R = set of half-spaces.

Recall a half space is defined by an inequality $\sum_{i=1}^{1} a_i x_i \le b$ for some $a_1, a_2, \ldots, a_d, b \in \mathbb{R}$

Claim VC-dim $= d+1$

It is easy to see that VC-dim $\geq d+1$

Take the $d+1$ points $(0,0,\ldots,0)$ and $(1,0,\ldots,0)$, $(0,1,0,\ldots,0)\ldots(0,0,\ldots,0,1)$.

This set can be shattered. Why?

However $d+2$ points cannot be shattered and this follows from Radon's theorem.

Theorem [Radon's theorem]

Let Q be a set of $d+2$ points in $\mathbb{R}^d$. Then one can partition Q into S_1 and S_2 such that

$$\text{convexhull}(S_1) \cap \text{convexhull}(S_2) \neq \phi.$$

The preceding theorem $\Rightarrow$ Q cannot be shattered by half spaces.

Now that we have seen the definition of VC-dimension we state and prove a key technical lemma about set systems with bounded VC-dimension.

Sauer's lemma

Theorem [Sauer's Lemma]

Suppose a set system (P, R) has VC-dimension at most d. Let $Q \subseteq P$ be a finite set of cardinality n.

Then $|\{Q \cap r \mid r \in R\}|$

$$\leq \sum_{i=0}^{d} \binom{n}{i} \leq n^d.$$

Proof: By induction on n.

If $n=0$ it is trivial.

Let Q be a set of n points $n>0$.

We can restrict attention to $r \cap Q$ $\forall r \in R$. Hence we can work with a finite range space. Now all we need to do is count $|R|$.

Fix some $p \in Q$.

Let $R_1 = \{ r - \{p\} \mid r \in R \}$ be the set of all ranges obtained by removing p from the original ranges.

Suppose $\exists\ \bar{r}$ such that $p \in r$ and also $r - \{p\} \in R$. Then both $r \cup \{p\}$ and $r - \{p\}$ project to same range in R_1. So to count $|R|$ we create a separate range space.

Let $R_2 = \{ r - \{p\} \mid r \cup \{p\} \in R$ and $r - \{p\} \in R \}$.

From this explanation we have

Claim: $|R| = |R_1| + |R_2|$.

Now we consider the two range spaces $(Q - \{p\}, R_1)$ and $(Q - \{p\}, R_2)$.

<u>Claim</u>: VC-dim $(Q - \{p\}, R_1) \leq d$

<u>Proof</u>: removing a point does not increase VC-dim

<u>Claim</u>: VC-dim of $(Q - \{p\}, R_2) \leq d-1$.

<u>Proof</u>: If $Q' \subseteq Q - \{p\}$ is shattered by R_2 then since every range

$r \in \mathcal{R}_2$ satisfies the property that $r \cup \{p\}$ and $r - \{p\} \in \mathcal{R}$ we would have $Q' \cup \{p\}$ is shattered by $\mathcal{R}$. Thus $|Q'| \leq d-1$.

$\square$.

Now by induction

$$|\mathcal{R}_1| \leq \sum_{i=0}^{d} \binom{n-1}{i}$$

and $|\mathcal{R}_2| \leq \sum_{i=0}^{d-1} \binom{n-1}{i}$

Thus

$$|\mathcal{R}| \leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i}$$

$$\leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d} \binom{n-1}{i-1}$$

$$\leq \sum_{i=0}^{d} \binom{n}{i}.$$

$\square$.

In many settings the only way VC-dim is used is via the bound given by Sauer's lemma. So it makes sense to define the following.

<u>Defn</u>: The <u>shattering dimension</u> of a range space (P, R) is d if $\forall\ Q \subseteq P$ with $|Q| \leq n$ the size of $R_{|Q} \leq n^d$.

VC-dim $(P, R) \leq d \Rightarrow$ shattering dim$(P, R) \leq d$

Converse is also true with weaker parameter

Shattering-dim$(P, R) \le d \Rightarrow$ VC-dim(P, R)
$\le O(d \log d)$.

One important aspect of VC-dim is kind of closure when combining.

<u>Theorem</u>: Suppose (P, R_1) and (P, R_2) are range spaces with VC-dim d_1 and d_2 respectively. Then VC-dim of (P, R) where $R = \{ (r_1 \cup r_2) \mid r_1 \in R_1, r_2 \in R_2 \}$ is $O(d_1 + d_2)$. Similarly for (P, R) when $R = \{ r_1 \cap r_2 \mid r_1 \in R_1, r_2 \in R_2 \}$.

ε-Sampling and ε-net theorems

We now discuss two theorems about how a random sample of a set from a set system (P, R) can approximate it.

For the following discussion it is useful to think of P as a finite set. Some of the concepts can be lifted to infinite sets with appropriate generalizations.

For a given system (P, R)

let $\mu(r) = \dfrac{|r \cap P|}{|P|}$ denote

the measure of r.

Suppose we take a "small" random sample Q from P. Does Q preserve the measure of all $r \in R$?

For this define

$$\mu_Q(r) = \frac{|r \cap Q|}{|Q|}.$$

Clearly a small sample cannot touch all ranges so we need to allow some "additive" error ε.

Defn: A subset $Q \subseteq P$ is an ε-sample for (P, R) if

$$|\mu(x) - \mu_Q(x)| \leq \varepsilon \quad \forall x \in R.$$

A related notion is the following.

Defn: Q is an ε-net for (P, R) if $|Q \cap x| \neq \phi$ $\forall x \in R$ where $\mu(x) > \varepsilon$.

Note that an ε-sample is automatically an ε-net.

<u>Theorem</u> Let (P, R) be a range space with VC-dimension $\leq d$.

Let $l \geq \frac{c}{\varepsilon^2}\left(d \log \frac{d}{\varepsilon} + \log \frac{1}{\delta}\right)$.

Then a random sample of l points with repetition from P is an ε-sample with probability $\geq 1-\delta$.

<u>Note</u>: The sample size does not depend on $|P|$. Could be infinite!

<u>Note</u>: The theorem relies only on the growth rate of the number of distinct ranges of a given size that follows from Sauer's lemma.

Hence the proof is not so tied to VC-dimension itself.

A stronger bound is known for ε-nets.

<u>Theorem:</u> Let (P, R) be a range space with VC-dim $\leq d$.
Let $l \geq \frac{C}{\varepsilon}\left(d \log \frac{d}{\varepsilon} + \log \frac{1}{\delta}\right)$

Then a random sample of l points with repetition from P is an ε-net with probability $\geq (1-\delta)$.

Proof of ε-sample theorem

It is a clever argument using a "double sample" argument.

First we recall the additive Chernoff bound.

Theorem: Let $X_1, X_2, \ldots, X_n \in [0,1]$ and independent. Let $Y = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $\mu = E[Y]$. Then

(i) $\Pr[Y > \mu + \varepsilon] \leq e^{-2\varepsilon^2 n}$.

(ii) $\Pr[Y < \mu - \varepsilon] \leq e^{-2\varepsilon^2 n}$.

To see how to use the above theorems in our setting. Fix a range r. $\mu(r) = \frac{|r \cap P|}{|P|}$. Suppose we take an l-sample Q with repetition.

What is $E[\mu_Q(r)]$?

Let X_i be indicator random variable for sample i being in r.

Let $Y = \frac{1}{l} \sum_{i=1}^{n} X_i$.

$E[X_i] = \frac{|r \cap P|}{|P|} = \mu(r)$.

Hence $E[Y] = \mu(r)$.

Therefore $|E[Y] - \mu(r)| > \varepsilon \leq 2e^{-2\varepsilon^2 l}$.

Proof of ε-sample theorem

It is a clever argument using a "double sample" argument.

Let Q_1 be a sample of size l.
Let Q_2 be an independent sample, also of size l.

Let A be the event that
$\exists$ some range R s.t.
$|\mu_1(R) - \mu(R)| > \varepsilon$. Here $\mu_1(R)$
$= \mu_{Q_1}(R)$ for short.

Let B be the event that $\exists R$ s.t.
$|\mu_2(R) - \mu_1(R)| > \varepsilon/2$

Claim: $\Pr[A] \leq 2 \Pr[B]$.

Let D be the event that $\exists x$ s.t. $|\mu_1(x) - \mu(x)| > \varepsilon$ and $|\mu_2(x) - \mu_1(x)| > \frac{\varepsilon}{2}$

We have
$$\Pr[B] \geq \Pr[D] = \Pr[D \text{ and } A]$$
$$= \Pr[D|A] \Pr[A].$$

We claim that $\Pr[D|A] \geq \frac{1}{2}$ which would imply that $\Pr[A] \leq 2\Pr[B]$.

To see this, suppose event A happens.
$\Rightarrow \exists x$ such that $|\mu_1(x) - \mu(x)| > \varepsilon$.
For this x, via the additive Chernoff bound

$\Pr\left[|\mu_2(x) - \mu(x)| \right] < \frac{\varepsilon}{2}$. This is because it is a fixed x and sample is big enough. and Q_2 is indep of Q_1.

If $|\mu_1(x) - \mu(x)| > \varepsilon$ and $|\mu_2(x) - \mu(x)| < \frac{\varepsilon}{2}$

then by triangle inequality

$$|\mu_2(x) - \mu_1(x)| \geq |\mu(x) - \mu_1(x)| - |\mu(x) - \mu_2(x)|$$

$$\geq \varepsilon - \frac{\varepsilon}{2} > \frac{\varepsilon}{2}.$$

$\Rightarrow$ $|\mu_2(x) - \mu_1(x)| > \frac{\varepsilon}{2} \Rightarrow D$ happens.

Thus $\Pr[D \mid A] \geq \frac{1}{2}$.

$\square$.

Thus we can focus on event B which is $P_x\left[|\mu_2(x)-\mu_1(x)|\right] > \frac{\varepsilon}{2}$ for some range x. Within factor of 2 we will get $P_x[A]$.

To analyzes B we think of the process differently. Instead of picking Q_1 and Q_2 independently as two separate steps we think of picking $2l$ elements Q_0 and then splitting Q_0 into two halves Q_1 and Q_2.

$$\Pr[B] = \sum_{Q_0} \Pr[Q_0] \Pr[B \mid Q_0]$$

$$\leq \max_{Q_0} \Pr[B \mid Q_0].$$

What is $\max_{Q_0} \Pr[B \mid Q_0]$?

We think of Q_0 as an arbitrary

multiset of $2l$ points and Q_1 and Q_2 are obtained by even splitting of Q_0 into l points each.

What is the advantage of this?

If we fix Q_0 then $R|_{Q_0}$ has $\leq (2l)^d$ ranges.

Thus we need to only worry about a "small" number of ranges.

For any fixed range r in $R|_{Q_0}$, via the additive Chernoff bound,

(i) $\Pr\left[\,|\mu_1(r) - \mu_0(r)| \geq \frac{\varepsilon}{4}\right] \leq \frac{\delta}{4(2l)^d}$

(ii) $\Pr\left[\,|\mu_2(r) - \mu_0(r)| \geq \frac{\varepsilon}{4}\right] \leq \frac{\delta}{4(2l)^d}$

This is because $l \geq \frac{C}{\varepsilon^2}\left(d \log \frac{d}{\varepsilon} + \ln \frac{1}{\delta}\right)$.

Since $|\mu_1(x) - \mu_2(x)| \leq |\mu_1(x) - \mu_0(x)| + |\mu_2(x) - \mu_0(x)|$

$$\Pr\left[\,|\mu_2(x) - \mu_1(x)| \geq \frac{\varepsilon}{2}\right] \leq \frac{\delta}{2(2l)^d}.$$

By the union bound we

$$\Rightarrow \Pr[B] \leq \frac{\delta}{2}.$$

$$\Rightarrow \Pr[A] \leq \delta.$$

$$\Rightarrow \Pr[\bar{A}] \geq 1-\delta.$$

$\square$.