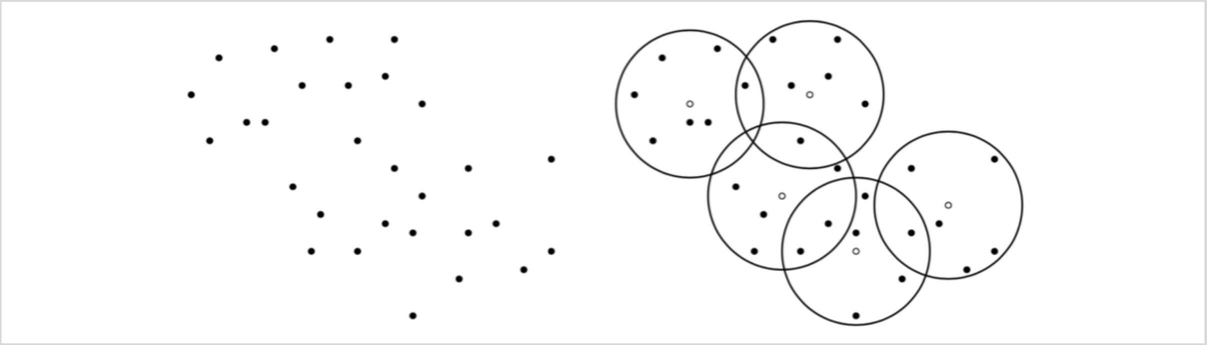
CS 6301.002.20S Lecture 25—April 21, 2020

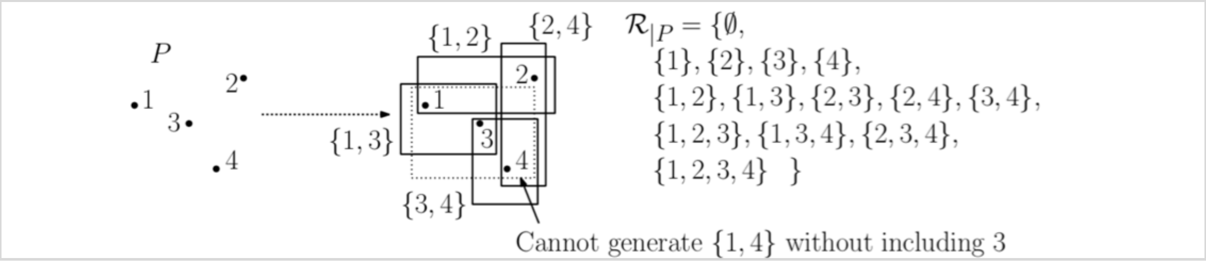
Main topics are #range_spaces and #VC_dimension.

Range Spaces

- Today, we're going to discuss grouping subsets of points by geometric objects and how to get a handle on these tasks.
- Here's one situation where you might group a subset based on geometric objects. Let's say you're installing wireless routers around campus. Each router effectively serves a disk of radius 1.
- You're given a list of locations that need service. How can you place as few routers as possible to service all those locations?

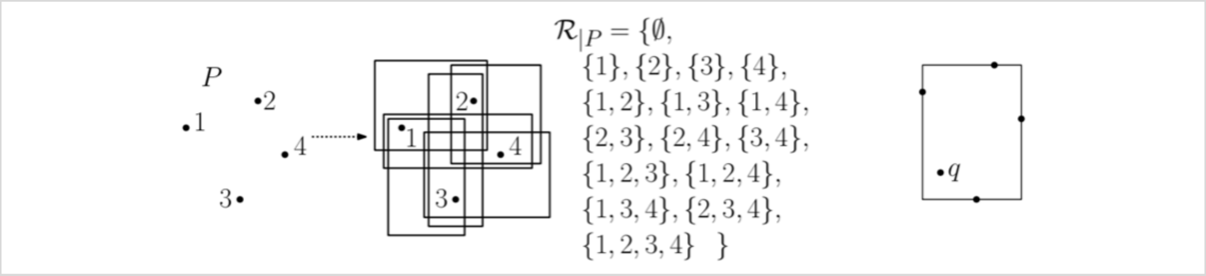


- In other words, you're given a collection of elements and a collection of subsets (who is served by each possible router location), and you want to cover all the elements with as few subsets as possible.
- This is a well-known NP-hard problem called set cover. It turns out defining the sets using these disks makes the problem a bit easier. We'll get to why after a good while.
- First, let's talk about a more general setting. We're given a set P of n points in R^d as usual.
- We let 2^P denote the *power set* of P, the set of all subsets of P.
- The power set has 2^n elements, but if we constrain ourselves to only considering subsets of P lying within simple geometric objects, the number of subsets decreases *a lot*.
- We describe these types of subsets using a range space. A *range space* is a pair (X, R) where X is an arbitrary set (finite or infinite) and R is a subset of 2^X.
- We usually care about how range spaces interact with finite sets. Given P subset X, the *restriction* of R to P is R_{|P} = {P intersect Q : Q in R}.
- For example, if X = R^d, P is a set of n points, and R consists of all interiors of some axis-parallel rectangle, then R_{|P} is the subsets of P that lie within some axis-parallel rectangle.
- This means there may be some subsets of P that *don't* lie in the restriction. For example, {1, 4} from the figure below.



VC-Dimension

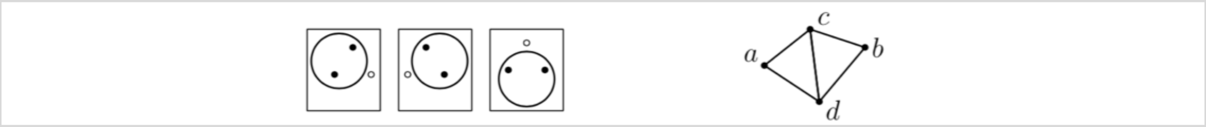
- The restrictions of range spaces based on simple geometric objects have a couple nice properties.
- One is that the restrictions don't have very many sets.
- For example, given a point set P, we can take any axis-parallel rectangle containing a subset of P and shrink it down to touch four points of P without changing the set of points in that rectangle.
- So in general, a rectangle for each of these subsets can be determined by four points of P. The total number of subsets is O(n^4).
- This idea that we can't create all 2^n subsets motivates the next few definitions.
- Given range space (X, R) and a point set P subset X, we say R *shatters* P if R_{|P} is equal to 2^P.
- For example, those four points I drew earlier are *not* shattered by the range space of axis-parallel rectangles.
- However, the four points below to the left *are* shattered by this range space (ignore that box on the right for now).



- The *Vapnik-Chervonenkis dimension* (VC-dimension) of (X, R) is the size of the *largest* subset P of X that is shattered by R.
- I know this definition is bazaar, but it does sort of capture what we're looking for. If the VC dimension is very high (or even infinite) then you could have a very large number of subsets in a restriction. But preventing shattering reduces the number of restrictions as we already saw with axis-parallel rectangles (kind of).
- It's also useful when you want to talk about sampling points to reduce the size of a data set, but we'll get to that in a bit.
- First, some examples:
    - Axis-parallel rectangles have VC-dimension 4. We can shatter that example above, so

it must be at least 4. But suppose we have five points P (and, for simplicity, they are in general position). One point q in P does not lie on the smallest axis-parallel rectangle holding all five points. But then there is no rectangle containing P \ {q}.

- Euclidean disks in the plane have VC-dimension 3. Below is a set of 3 points shattered by disks. But say we have four points P (in general position). If any point is inside the convex hull of the others, then any disk containing the others also contains the forth point, similar to before. Otherwise, all the points are on the convex hull. Consider their Delaunay triangulation, and let a and b be the pair *not* connected by an edge. By the empty-circle property of Delaunay triangulations, any circle containing a and b must contain at least one of c or d as well. So no disk contains exactly {a, b}.



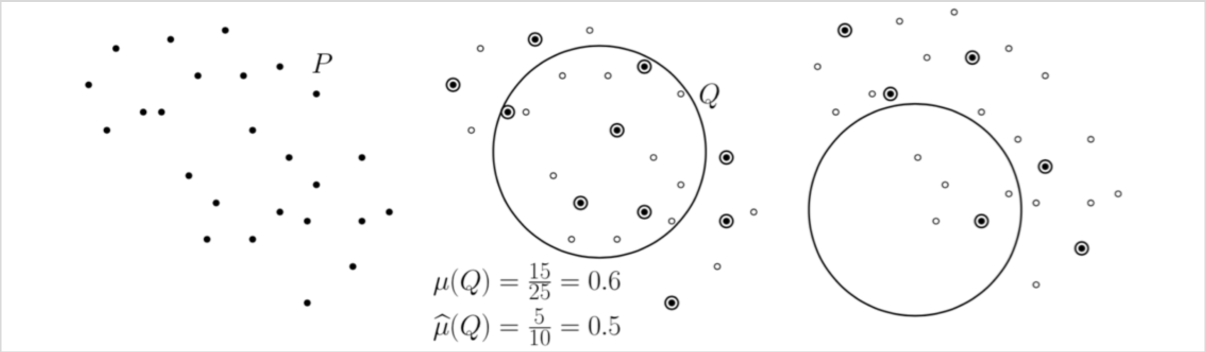
Sauer's Lemma

- I keep arguing that these nice range spaces have a small number of sets in their restriction. Let's finally make that precise.
- We need to define a useful function. Given 0 ≤ d ≤ n, define Phi_d(n) to be the number of subsets of size at most d over a ground set of size n.
- So Phi_d(n) = (n choose 0) + (n choose 1) + … + (n choose d) = sum_{i=0}^d (n choose i) = Theta(n^d)
- It turns out this function satisfies the following recurrence:
  - Phi_d(n) = Phi_d(n - 1) + Phi_{d - 1}(n - 1)
  - To see why, fix one element x_0 of the n-element set.
  - There are Phi_d(n - 1) subsets of size at most d that *do not* contain x_0.
  - The number of subsets that *do* contain x_0 is Phi_{d-1}(n - 1) because we have up to d - 1 additional elements to choose from for each of these subsets.
- Sauer's Lemma: If (X, R) has VC-dimension d and |X| = n, then |R| ≤ Phi_d(n) = Theta(n^d).
- Proof:
  - We'll do induction on d and n. The lemma is trivially true if d = n = 0.
  - Fix some element x in X. We define two new range sets:
    - R_x = {Q \ {x} : Q union {x} in R and Q \ {x} in R}
    - R \ {x} = {Q \ {x} : Q in R}
  - So R_x has pairs of ranges that are identical except one contains x and the other does not. Maybe x is on the side of some axis-parallel rectangle. So you take the points in the rectangle and the points in the slightly smaller rectangle not containing x.
  - R \ {x} is just what remains of the ranges after throwing out x.
  - I claim |R| = |R_x| + |R \ {x}|. To see why, charge each range of R to its corresponding

  - range in R \ {x}. Every range in R \ {x} receives at least one charge. It receives two if it is one of two identical ranges in R except one contains x and the other does not. But R_x contains a second copy of that range so we're counting correctly.
  - I claim (X \ {x} , R_x) has VC-dimension at most d-1. Otherwise, there is some set P' subseteq (X \ {x}) of size d that can be shattered by R_x. But for each member of 2^P', there are two subsets in R, one with x and one without. We can shatter the d + 1 element set P' union {x} which contradicts R having VC-dimension d.
  - Also, R \ {x} has VC-dimension at most d. Any subset P' of R \ {x} with d + 1 members is a subset of X as well, so we cannot shatter such subsets.
  - So, we have |R|
    - = |R_x| + |R \ {x}|
    - ≤ Phi_{d - 1}(n - 1) + Phi_d(n - 1)
    - = Phi_d(n)

Measures, Samples, and Nets

- So range spaces of low VC-dimension have small restrictions.
- Another property they have is that we can easily approximate the ways in which a large set of points intersect ranges. How do we make this precise?
- For simplicity, say we have a range space (P, R) where P is finite.
- Given Q in R, Q's *measure* is the fraction of P that it contains. We denote it as mu(Q) = |Q intersect P| / |P|.
- Given a *sample* S subset P, the *estimate* of Q is mubar(Q) = |Q intersect S| / |S|.
- S is a good sample if estimates are about equal to measures.
- Given eps > 0, we'll say S is an eps-sample if for *any* range Q in R, we have |mu(Q) - mubar(Q)| ≤ eps.
- So if eps = 0.1 and Q encloses 60% of P, then Q encloses 50-70% of S if S is an eps-sample.
- Samples provide good estimates on how much of P lies in any given range.
- But maybe we just want a sense of what are the important ranges. In our router example, maybe we only care about router locations that cover at least 10% of sites.
- Given eps > 0, we'll say S is an eps-net if for any range Q in R with mu(Q) ≥ eps, Q contains at least one point of S. In other words, the net "catches" every large range Q.
- So if eps = 0.2 and |P| = 25, then S will contain at least one point from any range with at least 5 points of P.
- The figure below shows the example set P, an eps-sample, and an eps-net in left-to-right order.



P
Q
$\mu(Q) = \frac{15}{25} = 0.6$
$\widehat{\mu}(Q) = \frac{5}{10} = 0.5$

- Any eps-sample is an eps-net, but eps-nets are often much smaller than would be needed to get an eps-sample.
- So how do we find a good sample or net? It turns out we can just take elements at random, assuming the VC-dimension is small.
- Theorem: Let (X, R) be a range space of VC-dimension d, and let P be any finite subset of X. There exists a positive constant c (independent of R) such that with probability 1 - phi a random sample S of P of size at least

  c / eps^2 (d log d / eps + log 1 / phi)

  is an eps-sample for (P, R_{|P}).
- So if d and phi are constants, you just need O((1 / eps^2) log (1 / eps)) points in your random sample. That number has no dependency on the size of P!
- Theorem: Same as above, but now a random sample S of P of size at least

  c / eps (d log 1 / eps + log 1 / phi)

  is an eps-net for (P, R_{|P}) with probability 1 - phi.
- So for constant d and phi, you need only O((1 / eps) log (1 / eps)) points in the sample.
- The proofs are fairly technical, so I will not be going over them.
- Next time, we'll discuss some algorithmic applications of the two theorems.