Monday 4/6

VC–Dimensions

Before we get started:

We have a matrix C (for cities) where each row is the x and y coordinates of a city.

CC^T = Similarity matrix

The similarity matrix is something like the inverse distance between cities.

If you do a SVD on the similarity matrix, you should get 2 large singular values (b/c map is 2D) plus some noise (since Earth is round). The first 2 columns of matrix U should give you x, y values, and when you plot these values, you should get a map of the US.

VC–Dimension

We want to be able to answer questions about a data set without actually having all the data. If there n are pieces of data, there are 2^n possible subsets of this data. Unless we actually have all these 2^n subsets, we cannot answer arbitrary questions about the data. But what if we put restrictions on the questions we can ask?

Say we restrict our questions to rectangles. Two rectangles (or squares, circles, etc.) are equivalent if they contain the same points.

- There are n^4 rectangles
- There are n^3 circles

What do we know about our estimate?

$|$ probability mass – estimate $| \leq \epsilon$ with probability $1 - \delta$. δ depends on ϵ and the VC-Dimension of the shape.

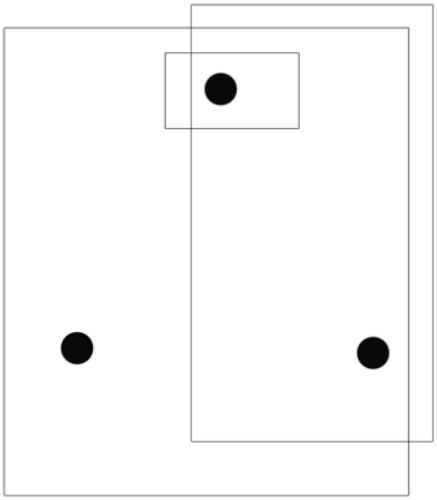
For example, (1-D case) suppose we have a line with points evenly distributed along it. Pick a rectangle that just barely fits 3 points. We can shift this rectangle to fit only 2 points, but the size of the rectangle limits us to enclosing just 2 or 3 points, so we can get a pretty good estimate of what is going on here.

On the other hand, (2-D case) suppose we have a plane with points distributed evenly like a grid. We can align our rectangle in the grid such that moving it by some ϵ can significantly increase the number of points contained in it. In this case, we can be off by +/- sqrt(n).

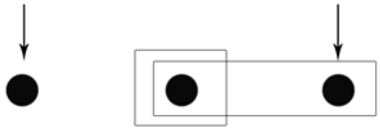
Set System (X, S) consists of a set X and a collection of subsets of X. (i.e. all points in plane and all points in a rectangle with horizontal and vertical lines).
Let (X, S) be a set system. A subset $A \epsilon X$ is **shattered** by S if each subset of A can be expressed as the intersection of A with an element of S.

The VC-Dimension of the set system is the size of the largest set A that can be shattered.

Example: 3 points and rectangle (can shatter)

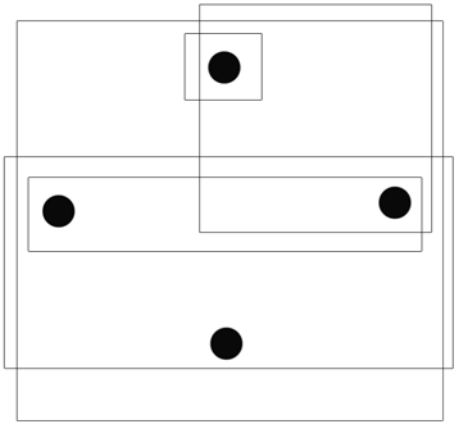


Example: 3 points in a line (cannot shatter)

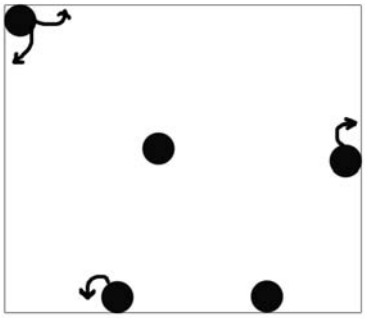


What is the largest set of points that can be shattered by a rectangle?

Example: 4 points and rectangle



But, we cannot shatter a set of 5 points.



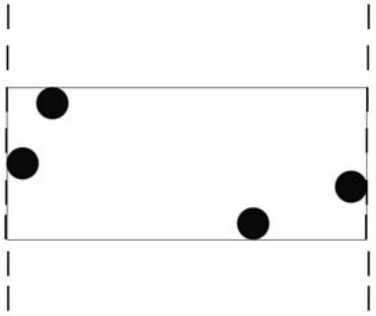
We can see that if we create a minimal size rectangle which contains all 5 points, that we cannot pick the points along the outside without picking the points on the inside.

What is the largest set of points that can be shattered by a circle?

We can shatter 3 points, but we cannot shatter 4 points and we want to show this. You can do this by proving that any circle that contains points A, B, and C also contains D.

What is the largest set of points that can be shattered by a square?

We can shatter 3 points, but not 4.
Find the minimum enclosing rectangle.



Any square that contains A and C (AC must be longer than DB) must contain 1 of B or D.

Justin Choi (jc626) & Cari Reiche (car56)