Intro to Dimensionality Reduction

Data Matrix

Recall

The columns commonly represent attributes/properties of the data

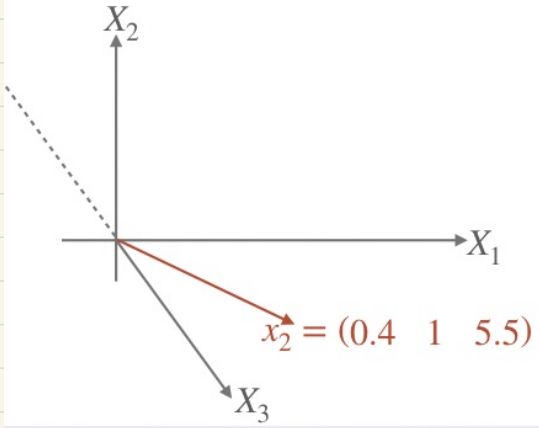
The rows commonly represent entities and their observed values for each attribute

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0 & C & 5.2 \\ x_4 & 5.6 & 50 & A & 5.1 \\ x_5 & -0.5 & 34 & A & 5.3 \\ x_6 & 0.4 & 19 & B & 5.4 \\ x_7 & 1.1 & 11 & A & 5.5 \end{array}$$

Geometric View of Data

Each row is a vector in m-dim space

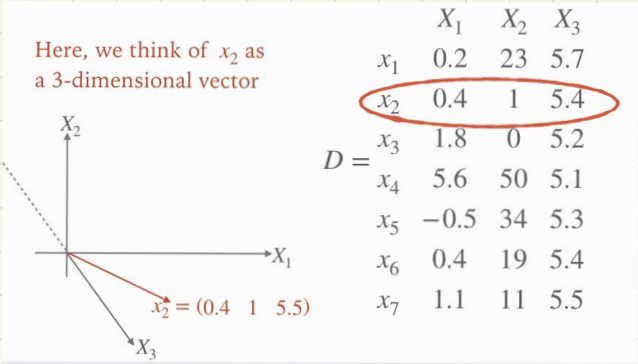
Here, we think of x_2 as a 3-dimensional vector



$$D = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

Geometric View of Data

What problems can occur in high dim space

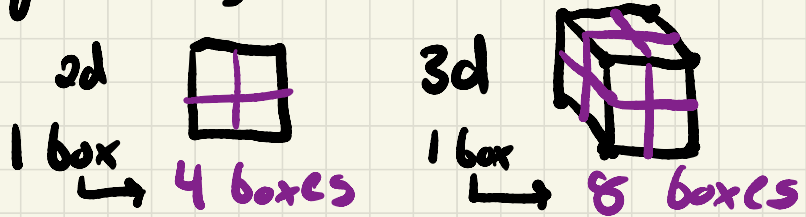


-
-
-

Geometric View in high dim

Some problems

- intuition from 2 & 3-dim doesn't always work
- volumes scale in non-intuitive ways
- near and far become harder to differentiate
- computing distances becomes more expensive
- algorithms become slower



Example: Volume of Ball w/ radius r

- 2D: πr^2
- 3D: $\frac{4}{3}\pi r^3$
- d-dim: $$\left(\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}\right) r^d$$

Euler's Gamma Function:

$$\Gamma\left(\frac{d}{2}+1\right) = \begin{cases} \left(\frac{d}{2}\right)! & \text{if } d \text{ is even} \\ \sqrt{\pi}\left(\frac{d!}{2^{(d+1)/2}}\right) & \text{if } d \text{ is odd} \end{cases}$$

Volume of d-dim ball w/ radius r

$$V_B(r,d) = \left(\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \right) r^d$$

$$\Gamma\left(\frac{d}{2}+1\right) = \begin{cases} \left(\frac{d}{2}\right)! & \text{if } d \text{ is even} \\ \sqrt{\pi}\left(\frac{d!}{2^{(d+1)/2}}\right) & \text{if } d \text{ is odd} \end{cases}$$

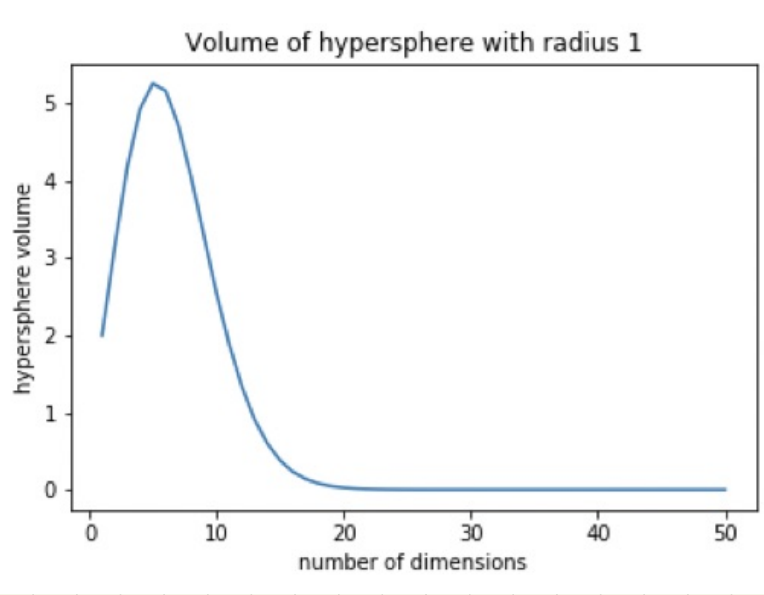
Let $r=1$, compute volume of a d-ball for:

- $d=4$ $V_B(1,4) = \frac{\pi^{\frac{4}{2}}}{\Gamma(\frac{4}{2}+1)} 1^4 = \frac{\pi^2}{2!} \approx 4.93$
- $d=12$ $V_B(1,12) = \frac{\pi^{\frac{12}{2}}}{\Gamma(\frac{12}{2}+1)} 1^{12} = \frac{\pi^6}{6!} \approx 1.34$

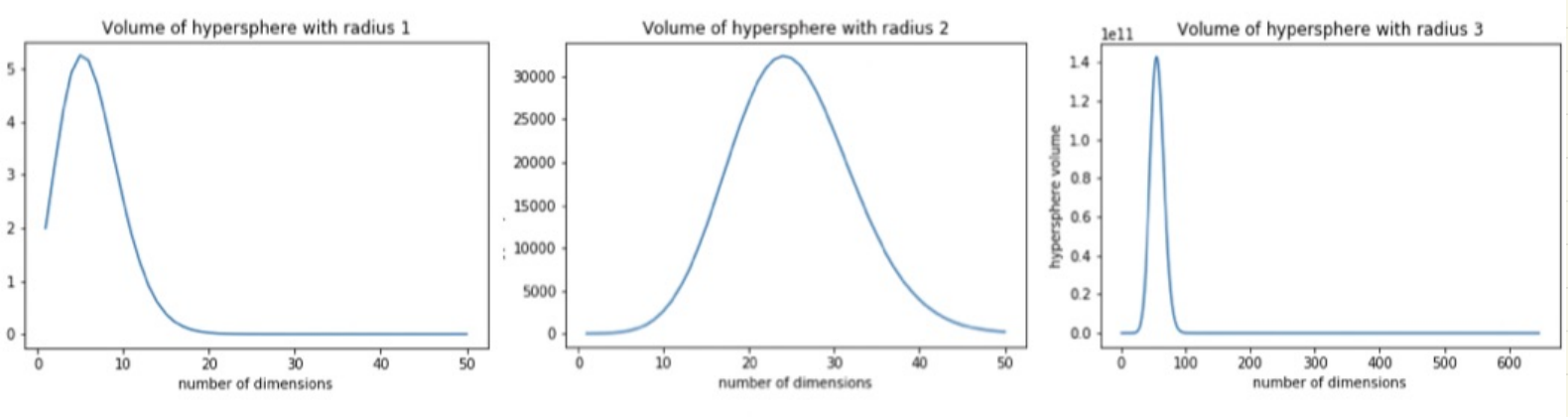
Volume of d-dim ball w/ radius r

Let r=1, compute volume of a d-ball for:

- d=2 $V_B(1,2) = \pi r^2 \approx 3.14$
- d=3 $V_B(1,3) = \frac{4}{3}\pi r^3 \approx 4.19$
- d=4 $V_B(1,4) \approx 4.93$
- d=12 $V_B(1,12) \approx 1.34$



Volume of d-dim ball w/ radius r



Other non-intuitive props:

- As dimensionality increases, the volume of a sphere with radius 1 increases up to a point, and vanishes as $d \to \infty$
- Consider a hypersphere inscribed inside a hypercube. As dimensionality increases, all the volume of the hyperspace is in the corners of the cube!

- The fraction of points with density at least α times the peak density of a multivariate normal distribution decreases rapidly (i.e., majority of the density is in the tail regions of the distribution!)
- The distance between the nearest neighbor of a point and the farthest neighbor of a point get closer together as dimensionality increases, making it difficult to find true "nearest neighbors"

How to solve? Work in lower dim!

Dim Reduction Brain Storm

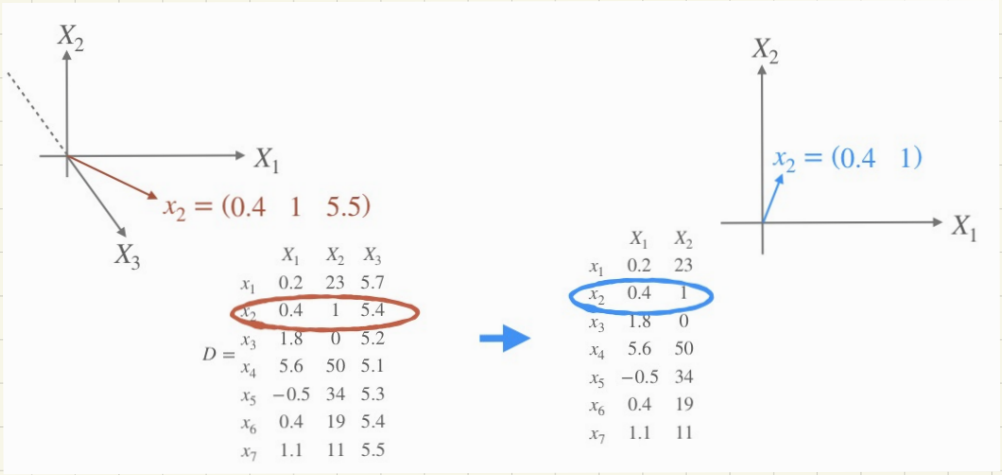
How can we reduce dim?

(+/- of that idea)

Feature Selection

select subset of attribs (feature selection)

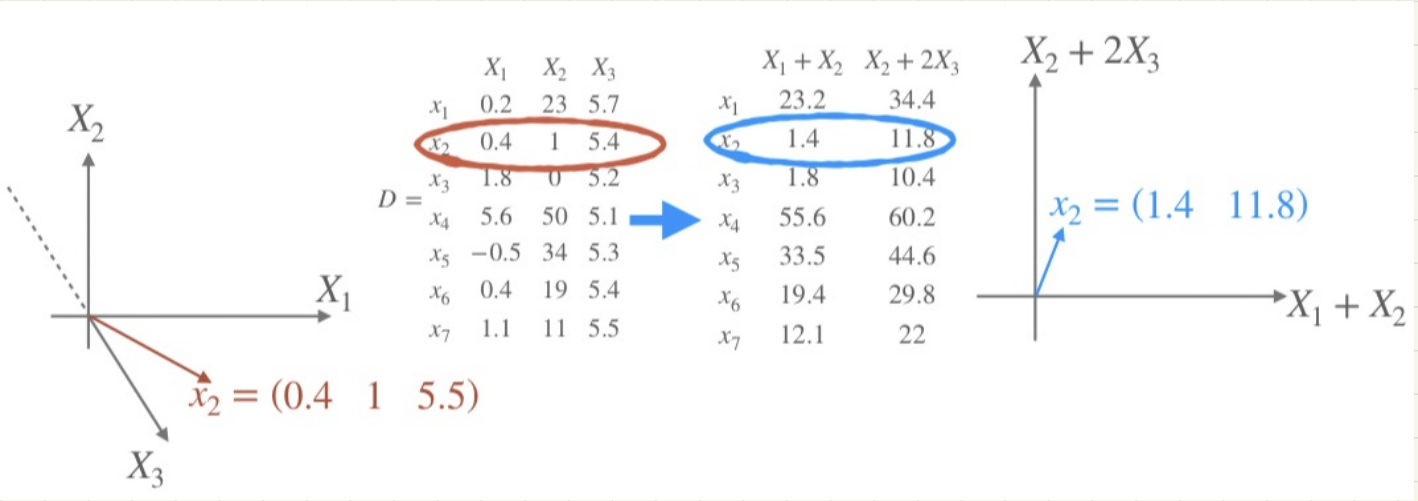
+ data still meaningful
- how to pick those attribs?
- lose some info



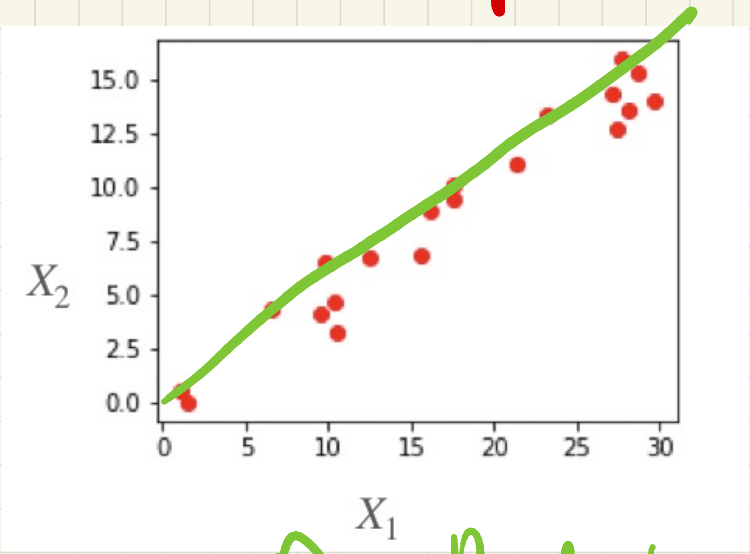
Different Representation

Use all attribs to create new attribs

- data may lose meaning
- how to pick the new attribs?
+ do not lose info



Low dim Rep - Often, data lies close to a low dim space

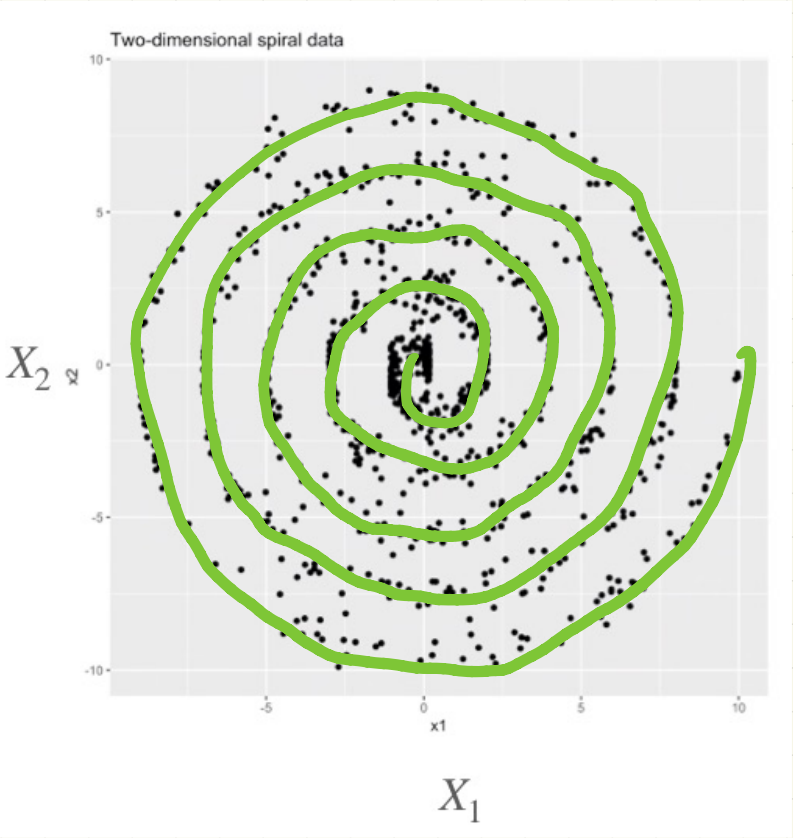


Idea find line that represents points pretty well

Linear Dim Reduction:

- Linear Discriminant Analysis (LDA)
- Principal Component Analysis (PCA)

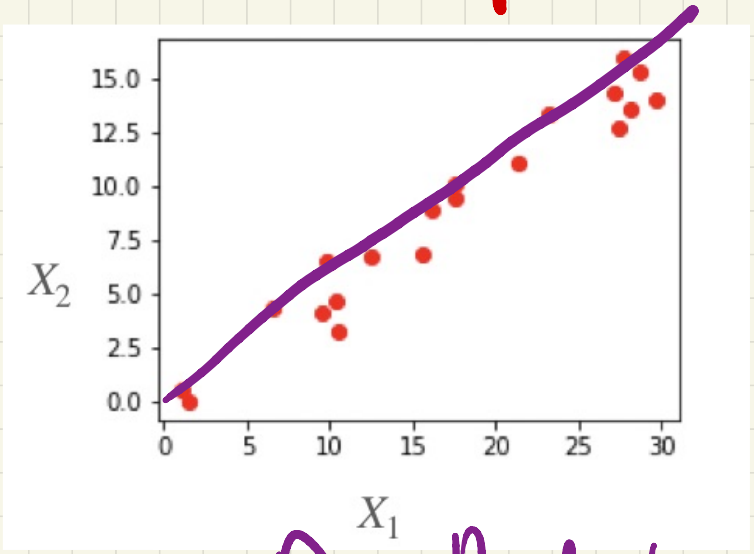
Low dim Rep - Often, data lies close to a low dim space



Idea find a curve that represents points pretty well

Nonlinear Dim Reduction:
- Multi Dim Scaling (MDS)
- Isomap

Low dim Rep - Often, data lies close to a low dim space



Idea find line that represents points pretty well

Linear Dim Reduction:

- Linear Discriminant Analysis (LDA)
- Principal Component Analysis (PCA)

Next time