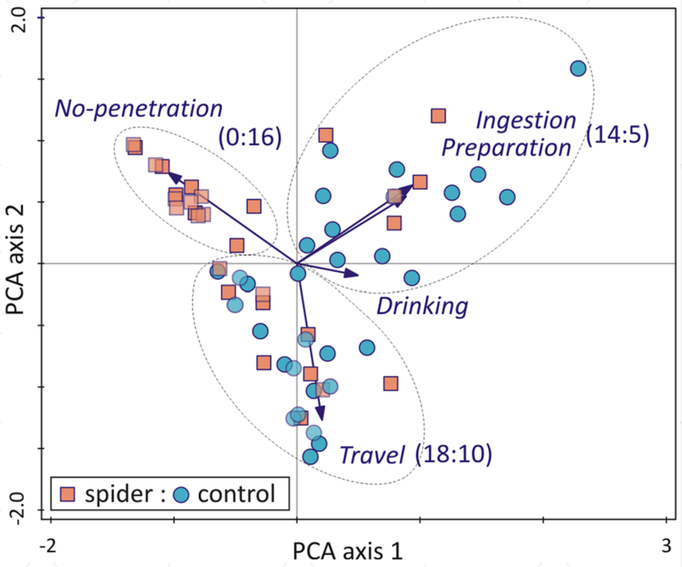


Principal Component Analysis

~Abhishek Kumar

PCA

- Unsupervised

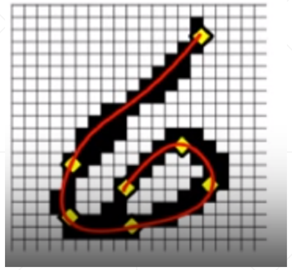
Curse of dimensionality

- Data points
  - # of skidding incidents
  - # of burst water pipes
  - Snow-plow expenditure
  - # of school closures
  - # of patient with heat stroke

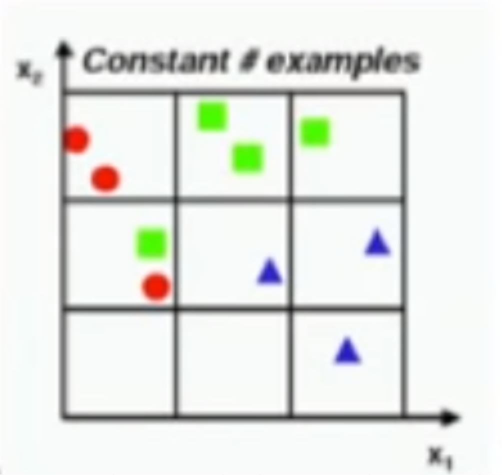
  - ~~ Temperature

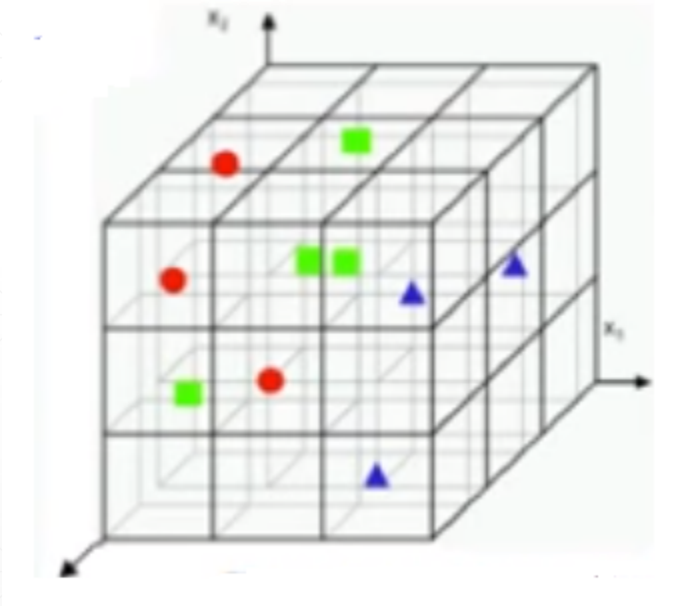
Curse of dimensionality

- Dataset which are high dimensional
- Examples?
- Machine learning are statistical methods
- Dimensionality grows, less observation



More dimensions ----- Sparse dataset





Dealing with High diemsionality

- Domain knowledge
- Make assumptions about dimension
  - Independence
  - Smoothness
  - Symmetry
- Reduce the dimensionality
  - Create new set of dimension

Goal

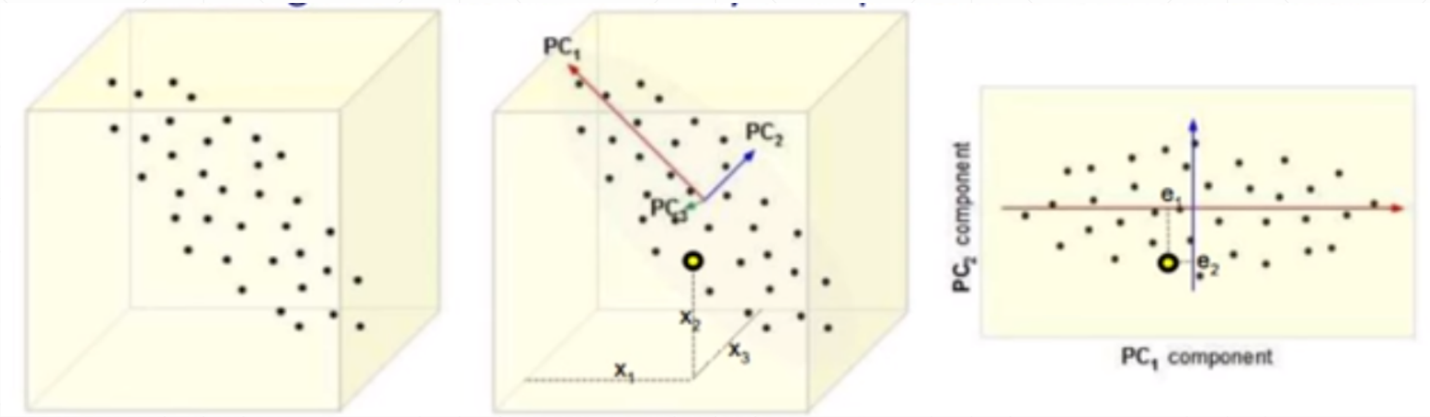
- Represent instances with fewer variables

Two ways

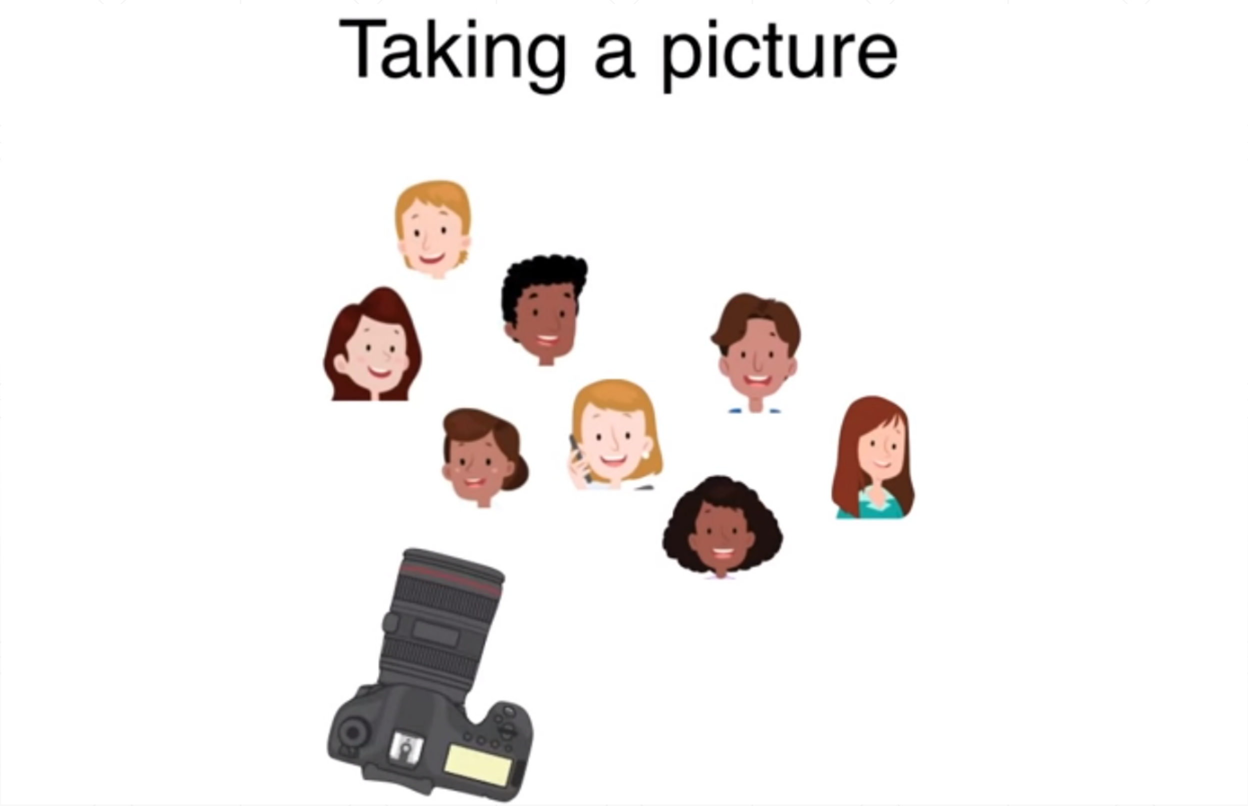
- Feature Selection
  - Pick a subset of original dimension
- Feature extraction
  - Construct new set of dimension

Principal Component Analysis

- Defines the set of principal components
- 1st : direction of the greatest variability
- 2nd: perpendicular to 1st
- 3rd :perpendicular to 2nd and so on
- $m << d$



Why greatest variablity



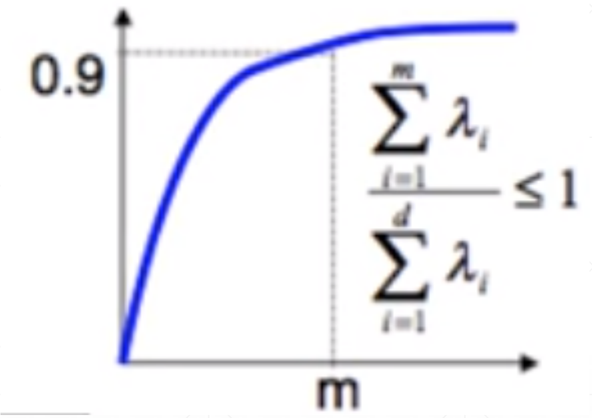
Covariance Eigenvector

- Center data to zero
- Compute co-variance metrics
- Multiply a vector by Σ:
- Want vector which are not turned

Finding Eigenvectors

How many dimensions?

- Eigenvector e_1 ------- e_d
- Eigenvalue = variance across e_i
- Pick e_i that explains most variance



- Pick first m eigenvector which covers 90% of the total variance
- Use scree-plot

PCA

- 1. Correlated High dimensional data
- 2. Center the points (want dimension of highest variance)
- 3. Compute Covariance metrics
- 4. Find Eigenvectors and Eigenvalues
- 5. Pick m<d eigenvectors with highest eigenvalues
- 6. Project data points to those eigenvectors
- 7. Uncorrelated low-d data

Issues

- Covariance extremely sensitive to large values
- Multiply one attribute by large number
- Dominates covariance
- Becomes principal component
- Normalize each dimension to zero mean and unit variance
- PCA assume underlying space is linear

Discussion

Thank you!