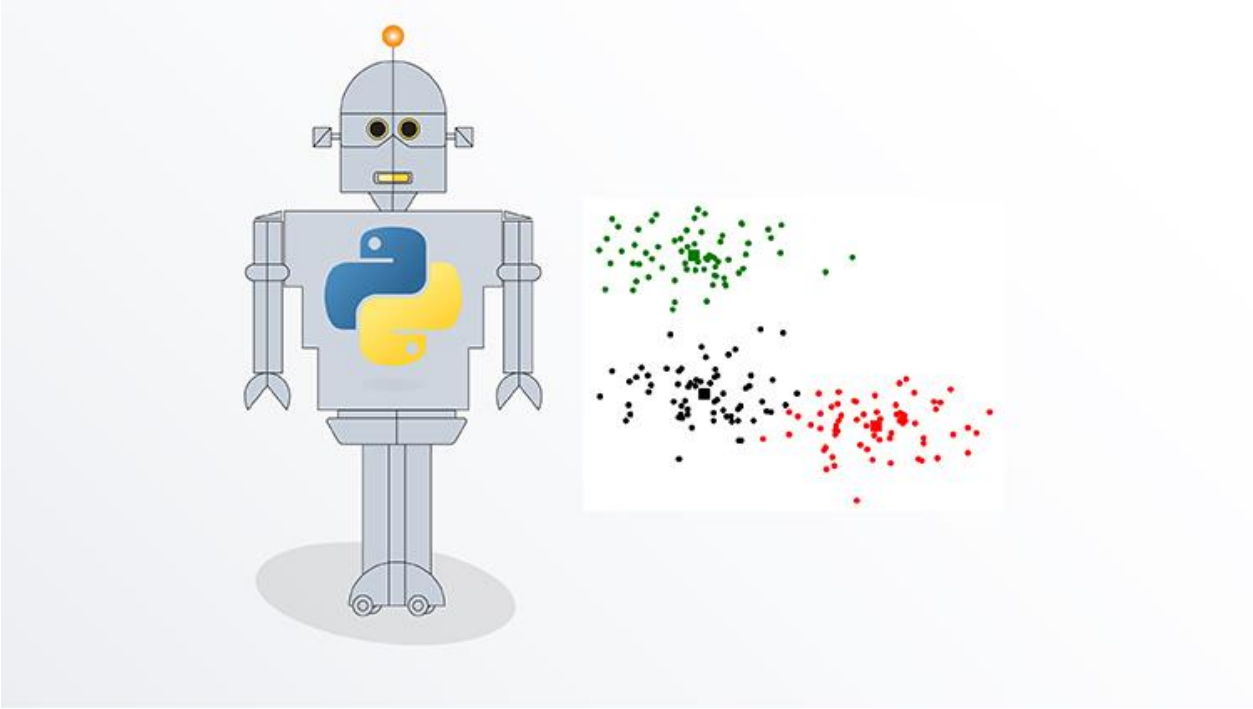
What is Machine Learning?

Copyright Satish Reddy 2020

Chapter 3: Unsupervised Machine Learning



Copyright Satish Reddy 2020

Unsupervised Machine Learning

Definition (see Wikipedia page on Unsupervised Learning)

- Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision.

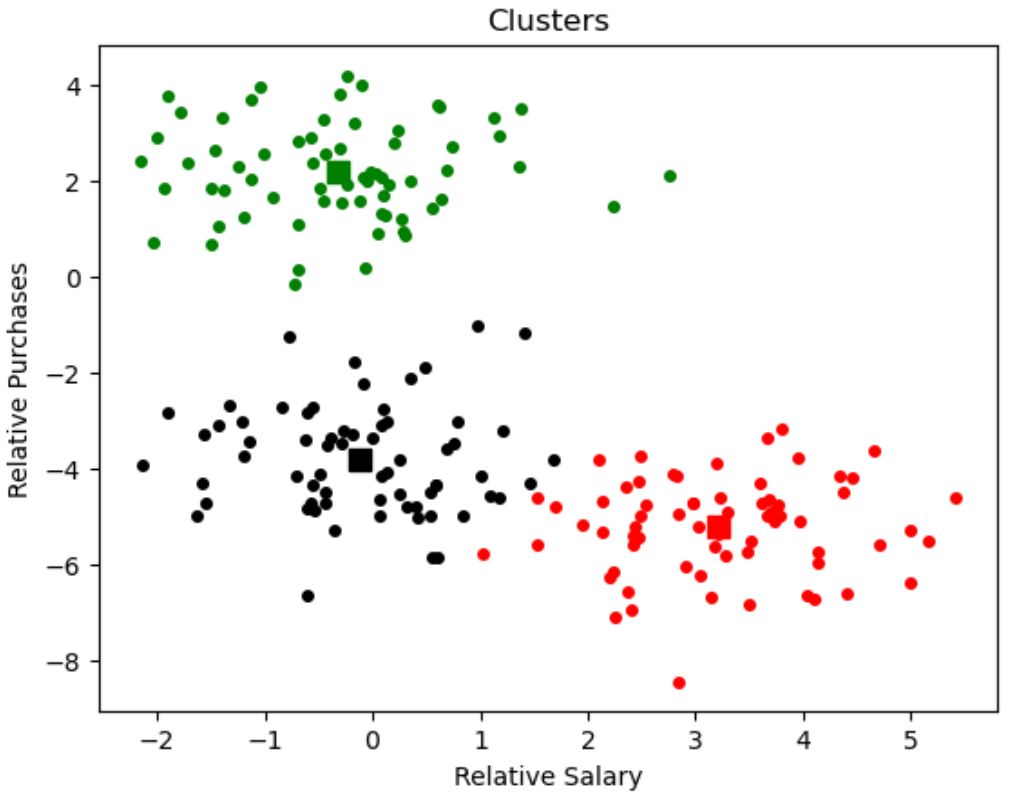
Informal Definition:

- Learn patterns in data

Copyright Satish Reddy 2020

Clustering Problem

- Given data points, find clusters
- Example: customer data



Copyright Satish Reddy 2020

What is the Data?

Typically, a data point is a vector X (d features)

- Customer Segmentation: X represents customer information
  - Each entry is feature of customer (age, sex, salary, number of purchases, etc)
- Image Classification: X represents an image
  - For black and white, entry is greyscale intensity for a pixel
  - Colour image has 3 intensities (red/blue/green) per pixel
  - May have a large number of pixels (400x300 image = 120,000 pixels)
- Natural Language Processing: X may represent a document
  - Example:
    - Have a large dictionary Word1, Word2, Word3, ... (d words)
    - Entry j of X is number of times Word j appears in the document

Note that the output Y used in Supervised Learning is not used in Unsupervised Learning

Copyright Satish Reddy 2020

Types of Clustering

See Wikipedia page on Cluster Analysis for more details

Connectivity-Based

- Combine nearby points to create clusters or nearby clusters to create larger clusters

Centroid-Based

- Estimate the centroid/mean of clusters

Distribution-Based

- Define clusters in terms of statistical distributions

Density-Based

- Define set of points to be in cluster if density of data points exceeds specified threshold

Copyright Satish Reddy 2020

Hierarchical Clustering

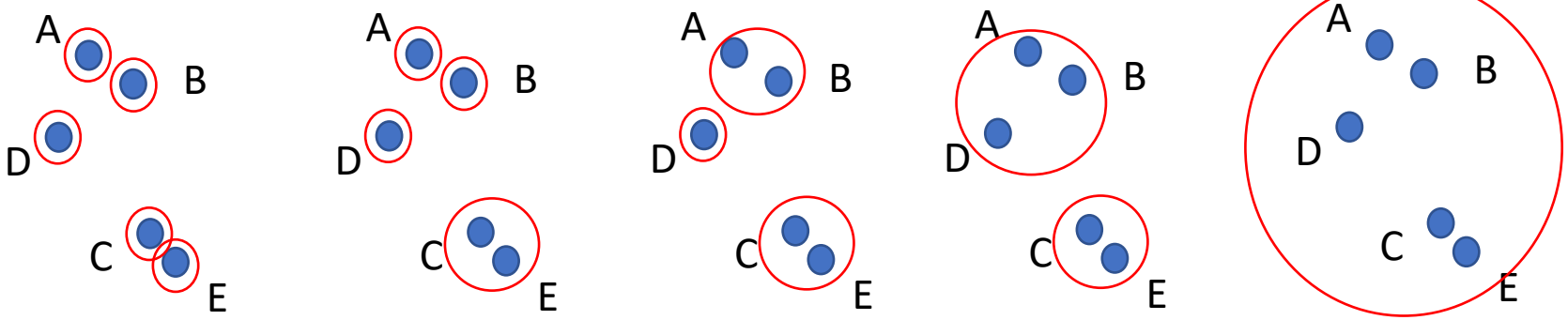
Hierarchical Clustering is a Connectivity-Based approach creating clusters at “all levels”

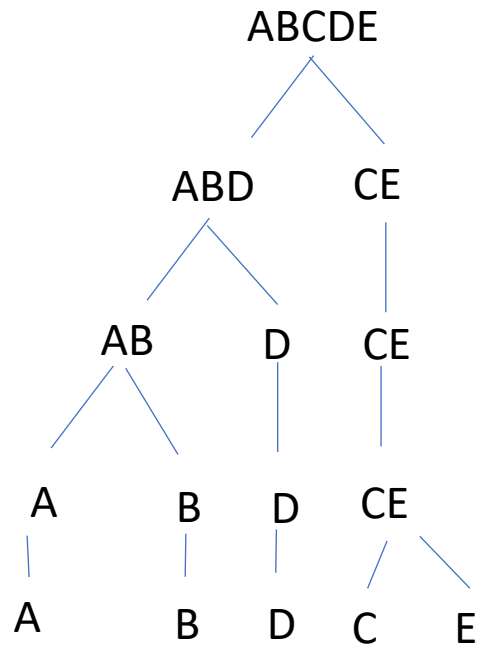
Algorithm:

(1) Assume N data points and define each as a cluster

(2) Compute distances between each of the clusters and combine the 2 clusters with the shortest distance between them into a single cluster

(3) Repeat (2) until all data points in a single cluster





Copyright Satish Reddy 2020

K Means Algorithm

K Means is a Centroid-Based approach:

Algorithm:

(1) Specify number of clusters K

(2) Make an initial guess for centre (mean) of each cluster

(3) For each cluster mean, find points that are closer to it than to other cluster means

(4) Based on points assigned to each cluster, re-compute cluster means and go back to (3)

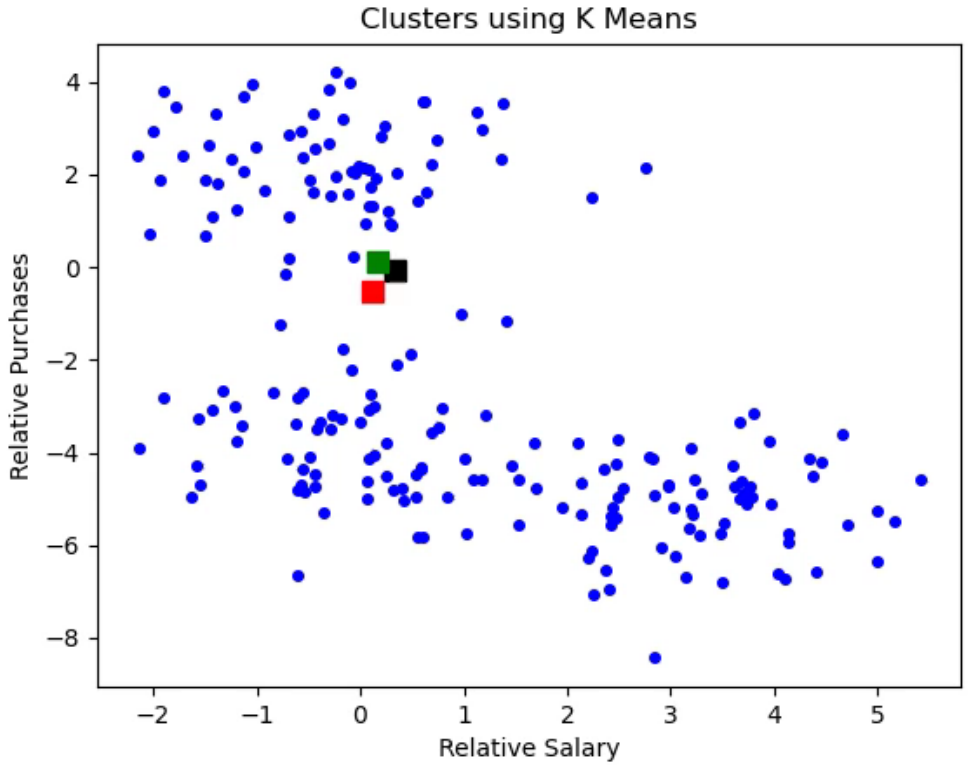
(5) Continue (3) and (4) until change in cluster means is sufficiently small

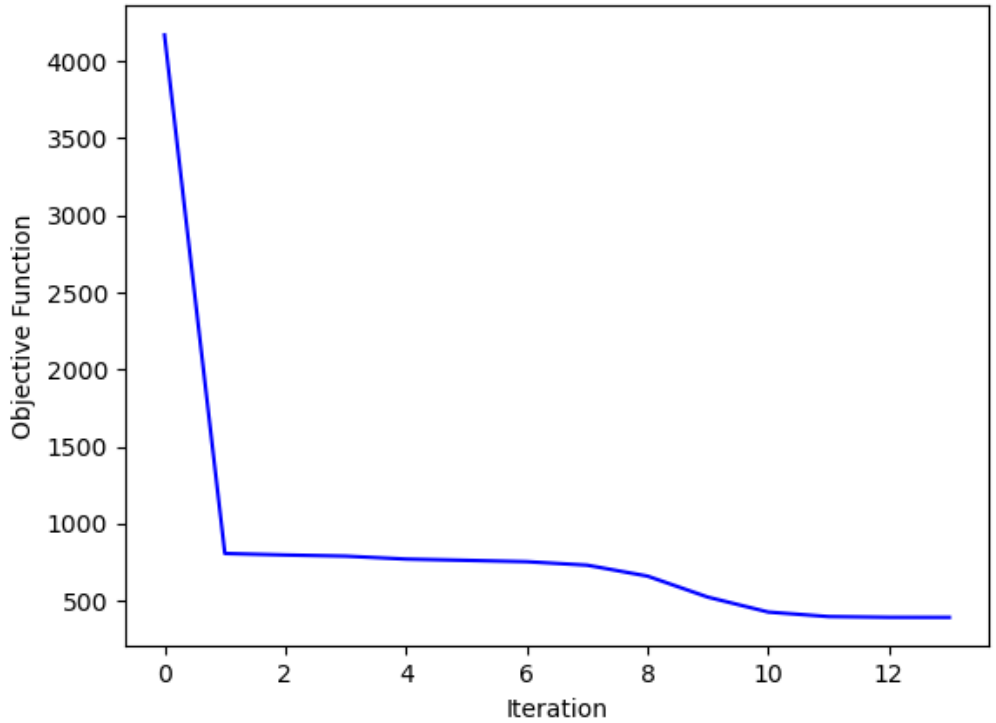
Can measure “goodness” of clustering by tracking objective function = sum of square distance to nearest mean (like a loss function)

Copyright Satish Reddy 2020

K Means Algorithm - Example

- Start with data and initial guess for cluster means
- Movie shows computation of clusters using K Means
- Plot shows objective function



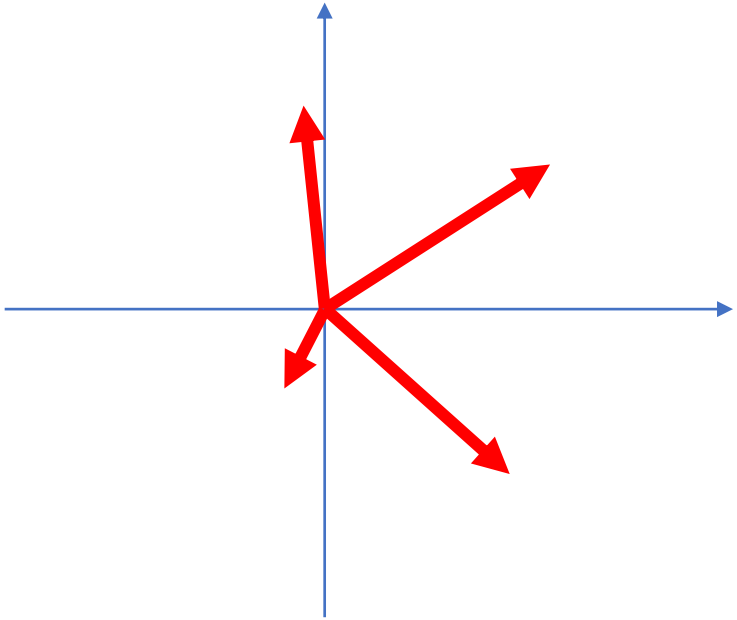


Copyright Satish Reddy 2020

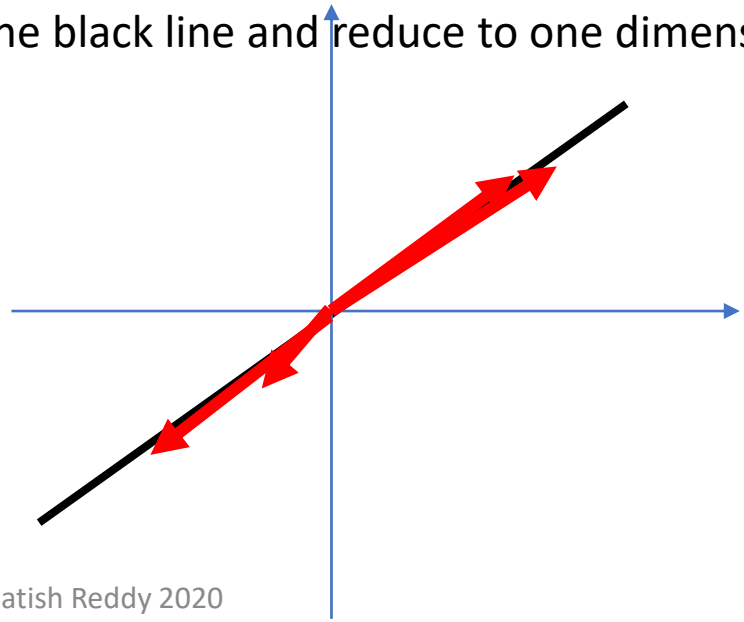
Principal Component Analysis (PCA)

- Machine learning problems may deal with data in 1000s of dimensions
- More dimensions generally means slower computation
- PCA reduce dimensions by retaining information in directions of most relevant principal components
- Typically choose # of principal components to capture specified amount of variance

Example 1: 4 data points in 2d. Since data points in all directions, both dimensions are relevant



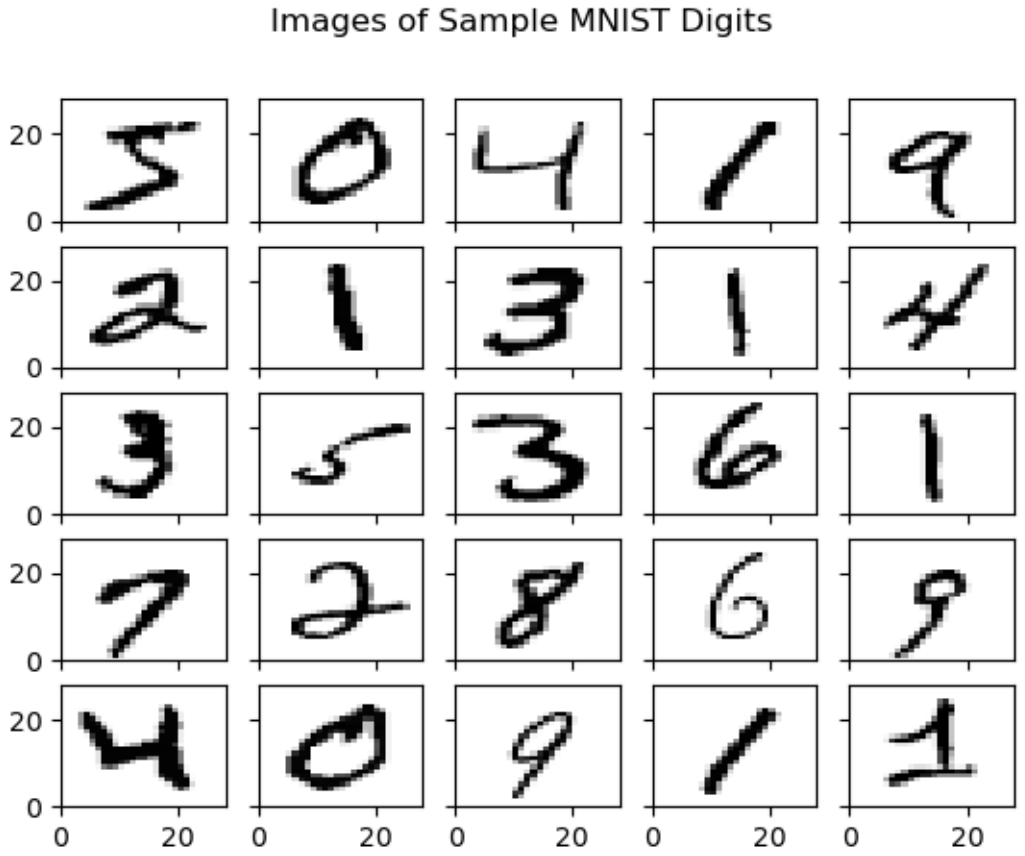
Example 2: 4 data points in 2d. Data points nearly align with black line. In this case one can project data onto the black line and reduce to one dimension



Copyright Satish Reddy 2020

PCA for MNIST Data Set

- Apply PCA to MNIST data set with 6000 images
- Data matrix X: 6000 samples with 784 entries/features per sample



Copyright Satish Reddy 2020

PCA for MNIST Data Set

PCA Algorithm:

- Subtract sample mean from X
- Compute Singular Value Decomposition of X - Xmean
- Variance is sum of squares of singular values
- Choose first n singular values so partial variance captures specified percentage

% Variance to be Captured	Number of Components
100.0	784
99.9	475
99.0	322
90.0	84

Copyright Satish Reddy 2020

Unsupervised Learning: Applications

Application	Data	Notes
Customer Segmentation	Customer features and behaviours	Group customers with similar features to be create customized marketing campaigns for each group
Image Segmentation	Images	Segment a set of unlabelled images into clusters based on how “close” images are to each other
Anomaly Detection	Features of a product	Perform cluster analysis and identify anomalies as outliers

Copyright Satish Reddy 2020

Unsupervised Learning: Notes

Component	Notes:
Definition of Distance	Need to define a distance measure • Hierarchical clustering: distance between clusters • K means: distance between data points Can use Euclidean distance, but there are other choices
Hierarchical Clustering	Not suitable for large amounts of data
K means	Issues: • Need to specify number of means K • May find unsuitable cluster means: estimated mean may get stuck between actual clusters • Final cluster means depend on initial guesses Despite issues, this is a good, straightforward starting point
Other approaches	Other approaches: • Based on probability distributions • Based on defining clusters in terms of density of data points

Copyright Satish Reddy 2020