

# Clustering: Grouping Related Docs



CS229: Machine Learning  
Carlos Guestrin  
Stanford University

Slides include content developed by and co-developed with  
Emily Fox

# Motivating clustering approaches

# Goal: Structure documents by topic

Discover groups (*clusters*) of related articles

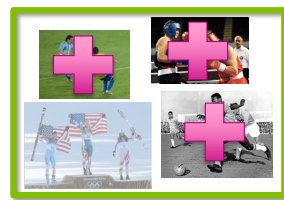


## Why might clustering be useful?



# Learn user preferences

Set of clustered documents read by user



Cluster 1



Cluster 2



Cluster 3



Cluster 4



Use feedback  
to learn user  
preferences  
over topics



## Clustering: An unsupervised learning task

# What if some of the labels are known?

Training set of labeled docs



SPORTS



WORLD NEWS



ENTERTAINMENT



SCIENCE

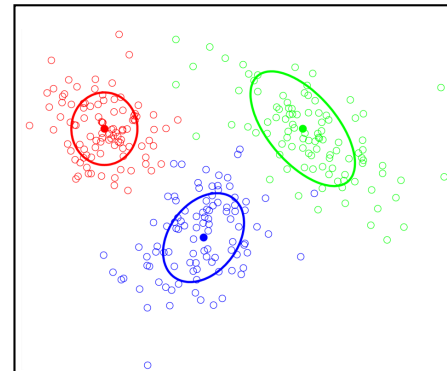
# Clustering

No labels provided  
...uncover cluster structure  
from input alone

Input: docs as vectors  $\mathbf{x}_i$

Output: cluster labels  $z_i$

An unsupervised  
learning task



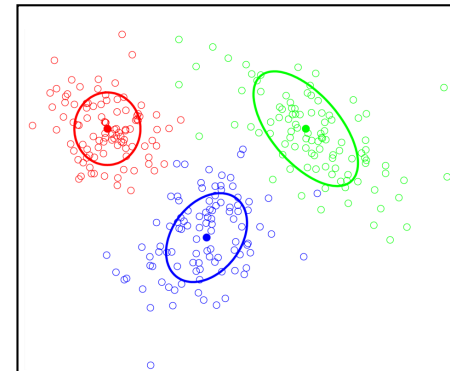


# What defines a cluster?

Cluster defined by **center** & **shape/spread**

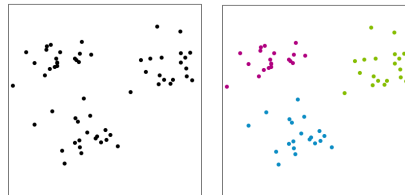
Assign observation  $x_i$  (**doc**)  
to cluster  $k$  (**topic label**) if

- Score under cluster  $k$  is higher than under others
- For simplicity, often define score as **distance to cluster center** (ignoring shape)

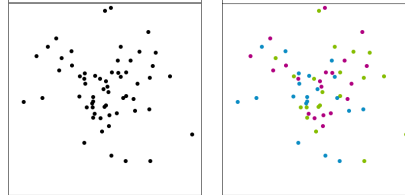


# Hope for unsupervised learning

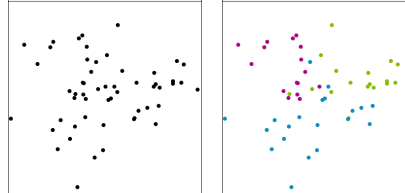
Easy



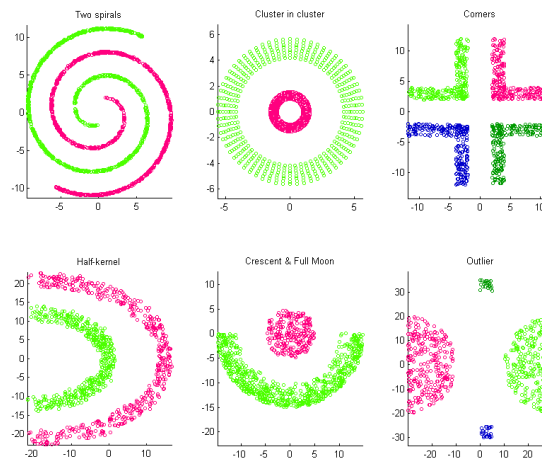
Impossible



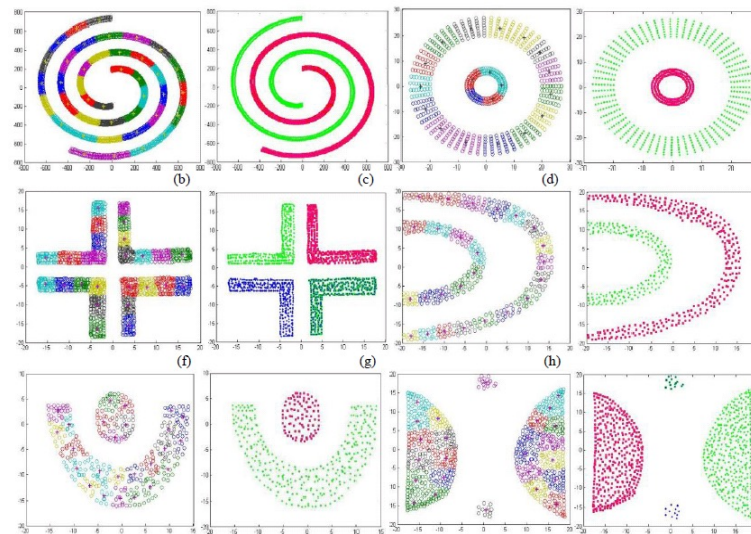
In between



## Other (challenging!) clusters to discover...



Other (challenging!) clusters to discover...



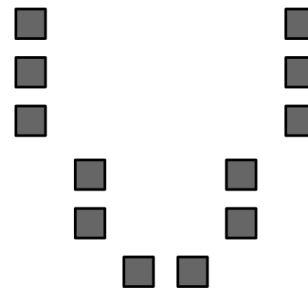


## k-means: A clustering algorithm

# k-means

Assume

–Score= distance to  
cluster center  
(smaller better)

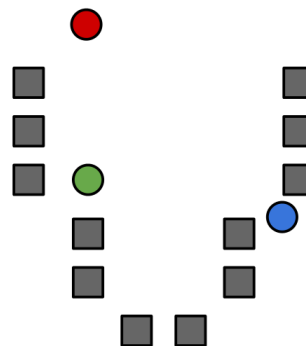


DATA  
to  
CLUSTER

# k-means algorithm

0. Initialize cluster centers

$$\mu_1, \mu_2, \dots, \mu_k$$

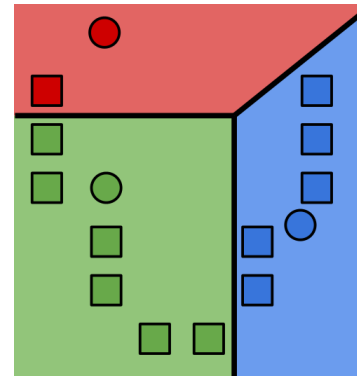


# k-means algorithm

0. Initialize cluster centers
1. Assign observations to closest cluster center

$$z_i \leftarrow \arg \min_j ||\mu_j - \mathbf{x}_i||_2^2$$

$z_i$  is the **Inferred label** for obs  $i$ , whereas supervised learning has **given label**  $y_i$

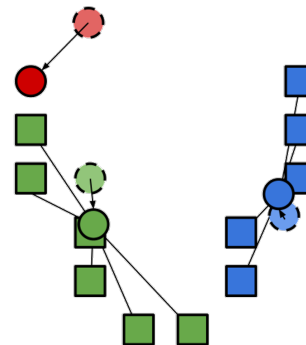




# k-means algorithm

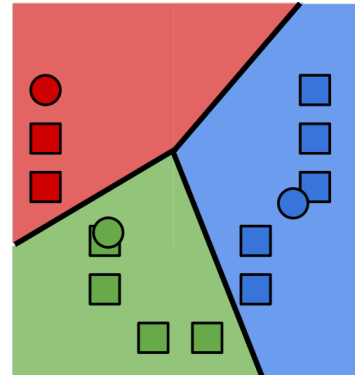
0. Initialize cluster centers
1. Assign observations to closest cluster center
2. Revise cluster centers as mean of assigned observations

$$\mu_j = \frac{1}{n_j} \sum_{i: z_i=j} \mathbf{x}_i$$



# k-means algorithm

0. Initialize cluster centers
1. Assign observations to closest cluster center
2. Revise cluster centers as mean of assigned observations
3. Repeat 1.+2. until convergence



## Why does K-means work???

- What's k-means optimizing?
- Does it always converge?

## What is k-means optimizing?

- Potential function  $F(\mu, \mathbf{z})$  of centers  $\mu$  and point allocations  $\mathbf{z}$ :
- Optimal k-means:

## Does K-means converge??? Part 1

- Optimize potential function:

$$\min_{\mu} \min_{\mathbf{z}} F(\mu, \mathbf{z}) = \min_{\mu} \min_{\mathbf{z}} \sum_{j=1}^N \|\mu_{z_j} - x_j\|_2^2$$

- Fix  $\mu$  and minimize  $\mathbf{z}$ :

## Does K-means converge??? Part 2

- Optimize potential function:

$$\min_{\mu} \min_{\mathbf{z}} F(\mu, \mathbf{z}) = \min_{\mu} \min_{\mathbf{z}} \sum_{j=1}^N \|\mu_{z_j} - x_j\|_2^2$$

- Fix  $\mathbf{z}$  and minimize  $\mu$ :

# Coordinate descent algorithms

$$\min_{\mu} \min_z F(\mu, z) = \min_{\mu} \min_z \sum_{i=1}^N \|\mu_{z_i} - x_i\|_2^2$$

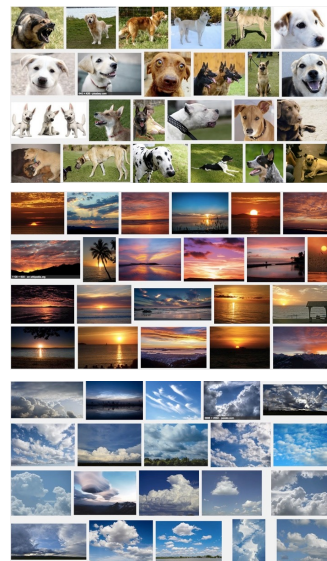
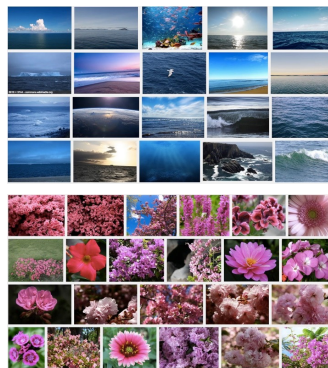
- Want:  $\min_a \min_b F(a,b)$
- Coordinate descent:
  - fix a, minimize b
  - fix b, minimize a
  - repeat
- Converges!!!
  - if F is bounded
  - to a (often good) local optimum
    - as we saw in applet (play with it!)
      - (For LASSO it converged to the global optimum, because of convexity)
- K-means is a coordinate descent algorithm!

## Summary for k-means



# Clustering images

- For search, group as:
  - Ocean
  - Pink flower
  - Dog
  - Sunset
  - Clouds
  - ...



# Limitations of k-means

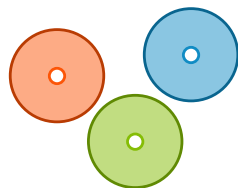
Assign observations to closest cluster center

$$z_i \leftarrow \arg \min_j \|\mu_j - \mathbf{x}_i\|_2^2$$

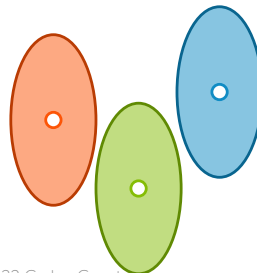
Can use weighted Euclidean,  
but requires *known* weights

Only center matters

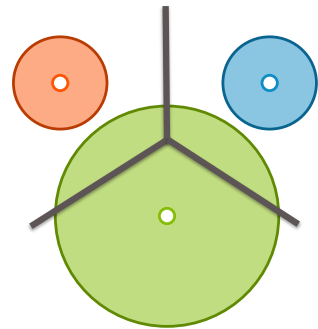
Equivalent to assuming  
*spherically symmetric* clusters



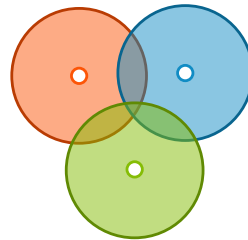
Still assumes all clusters have  
the same axis-aligned ellipses



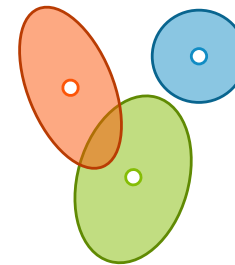
## Failure modes of k-means



disparate cluster sizes



overlapping clusters



different  
shaped/oriented  
clusters

## What you can do now...

- Describe the input (unlabeled observations) and output (labels) of a clustering algorithm
- Determine whether a task is supervised or unsupervised
- Cluster documents using k-means
- Describe potential applications of clustering