

DATA SCIENCE WITH MACHINE LEARNING: CLUSTERING

This lecture is
based on course by E. Fox and C. Guestrin, Univ of Washington

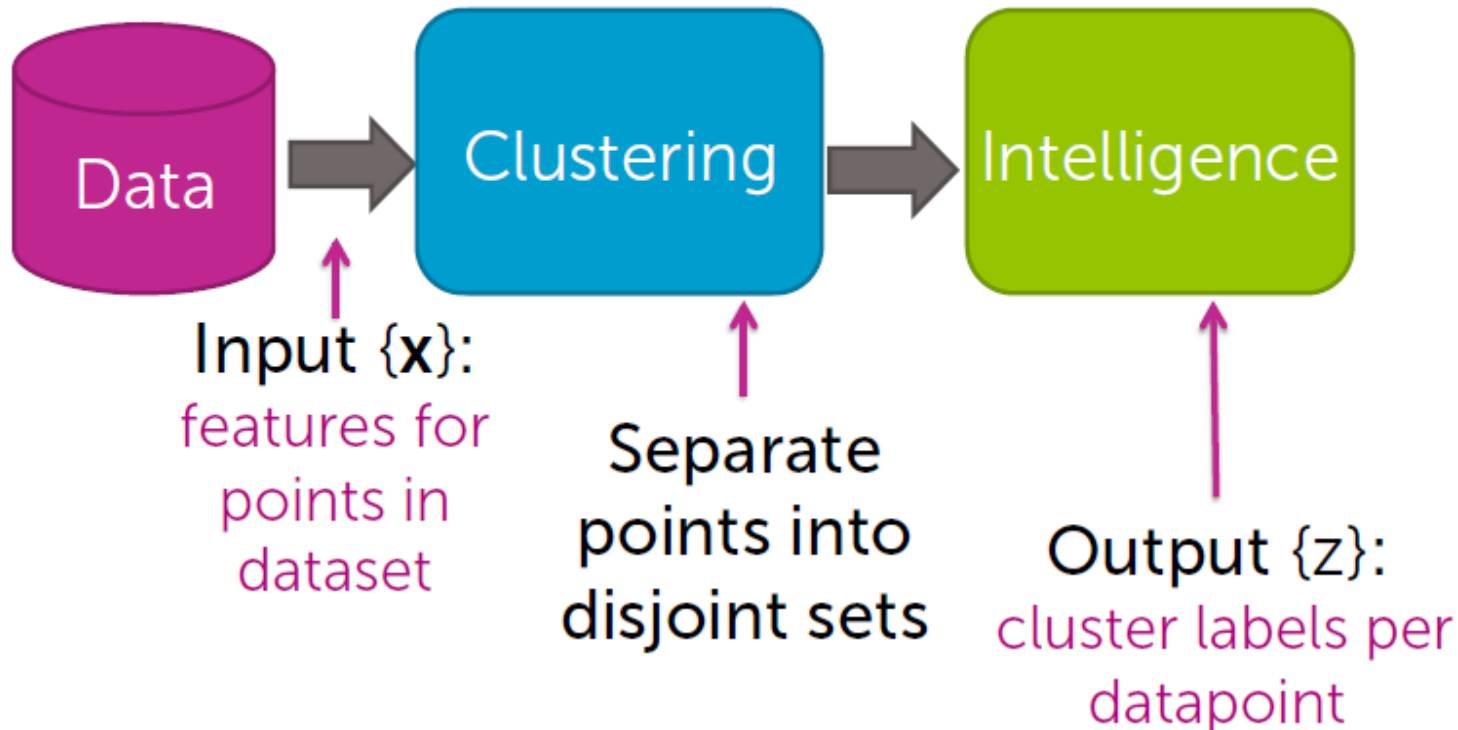
19/01/2021

WFAiS UJ, Informatyka Stosowana
I stopień studiów

What is clustering?

2

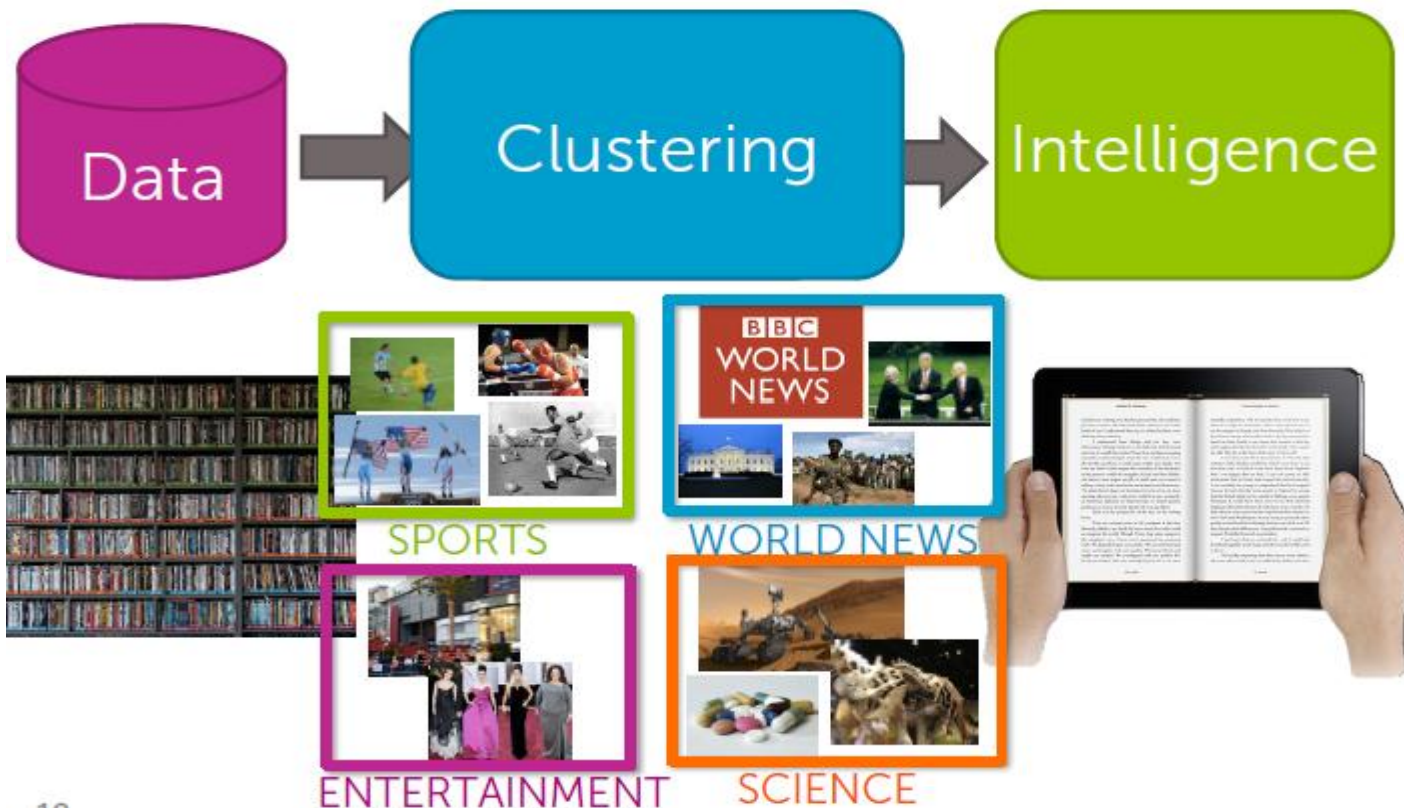
Discover groups of similar inputs



Clustering applications

3

Clustering documents by "topic"



Clustering applications

4

Clustering images

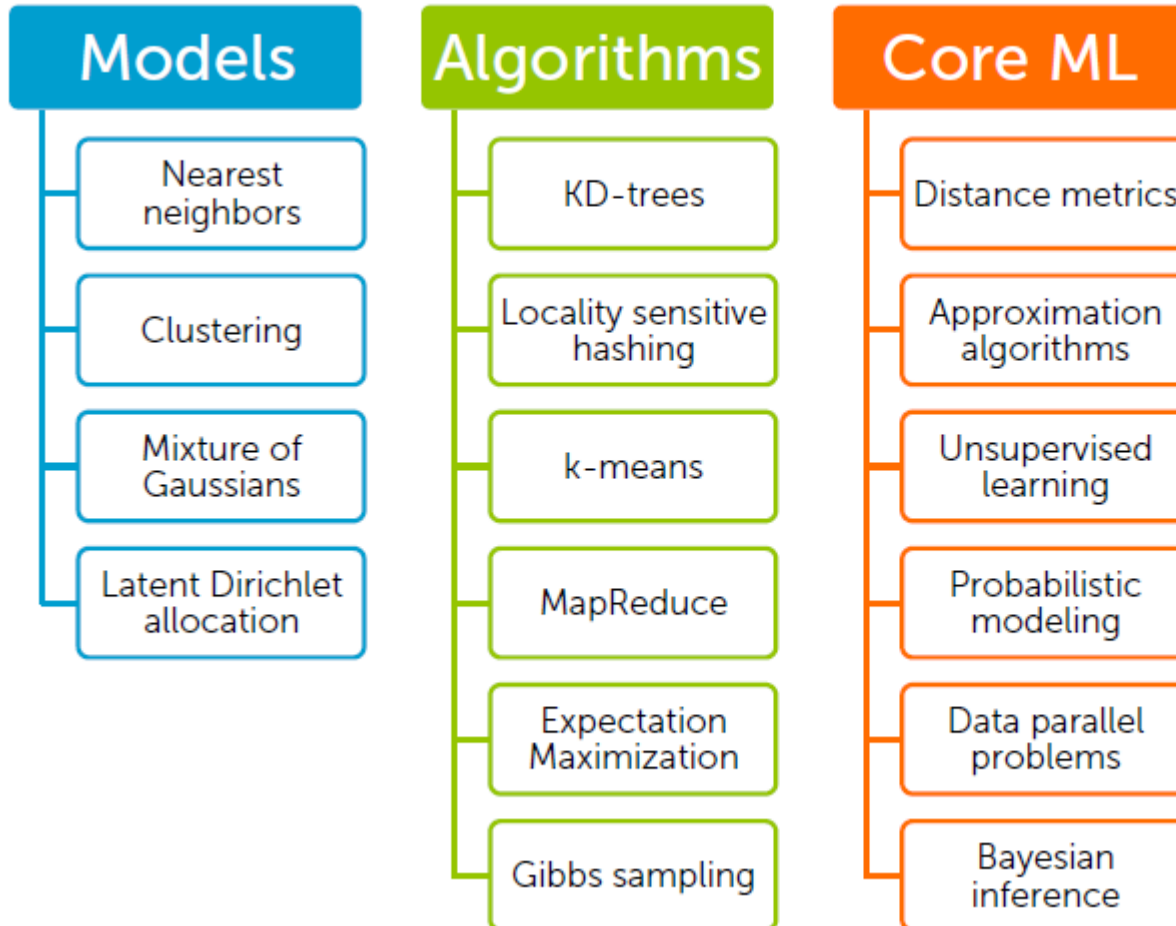
For search, group as:

- Ocean
- Pink flower
- Dog
- Sunset
- Clouds
- ...



Overview of content

5



Clustering:

An unsupervised learning task

Motivation

7

Goal: Structure documents by topic

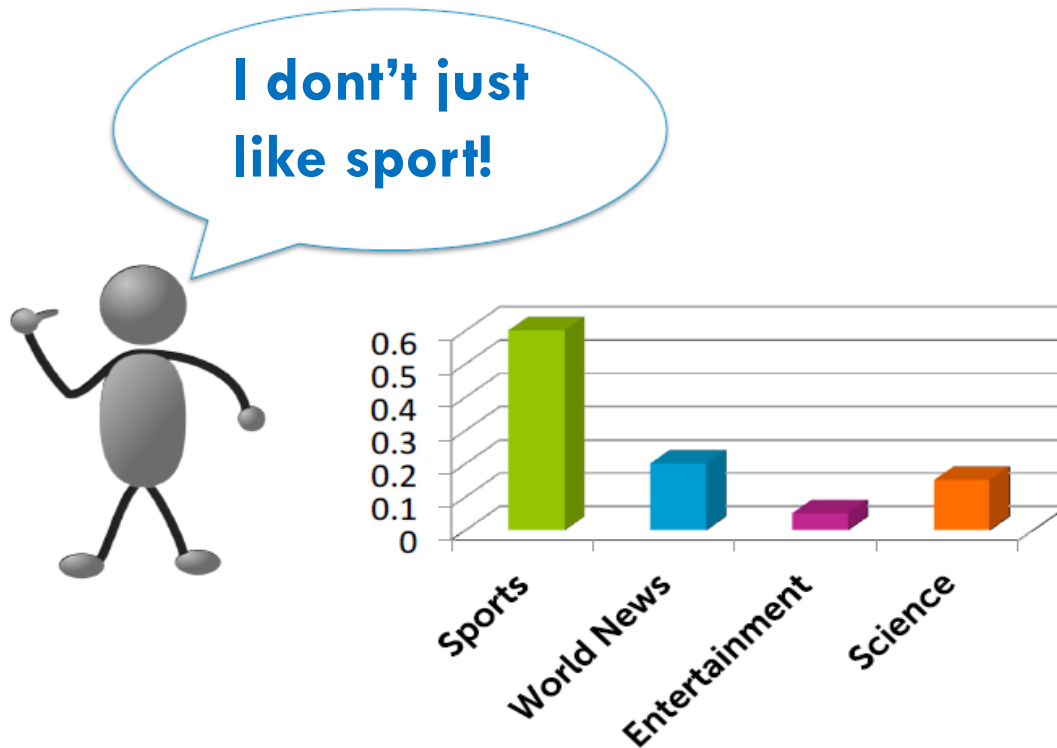
Discover groups (*clusters*) of related articles



Motivation

8

Why might clustering be useful?

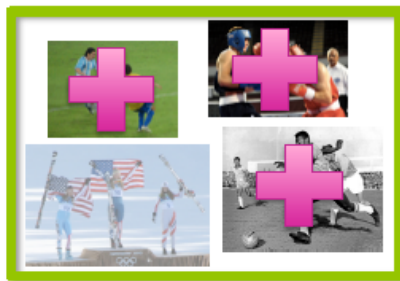


Motivation

9

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: a supervised learning

10

What if some of the labels are known?

Training set of labeled docs



SPORTS



WORLD NEWS



ENTERTAINMENT

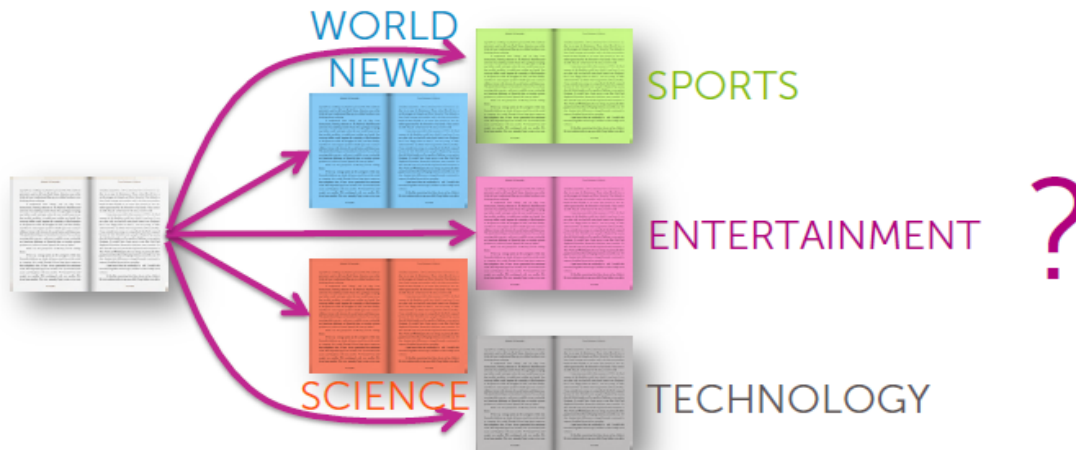


SCIENCE

Clustering: a supervised learning

11

Multiclass classification problem



Example of
supervised learning

Clustering: an unsupervised learning

12

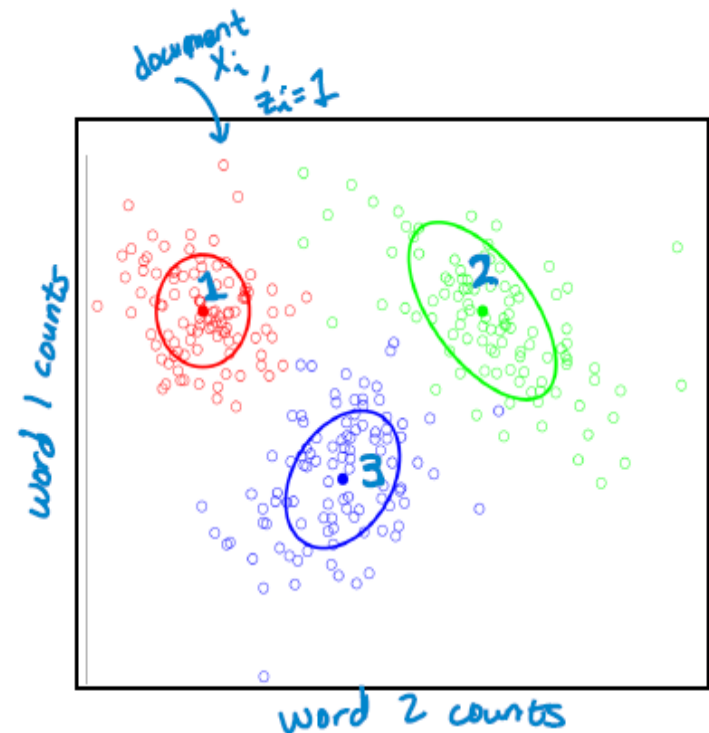
No labels provided

...uncover cluster structure
from input alone

Input: docs as vectors \mathbf{x}_i

Output: cluster labels z_i

An unsupervised
learning task



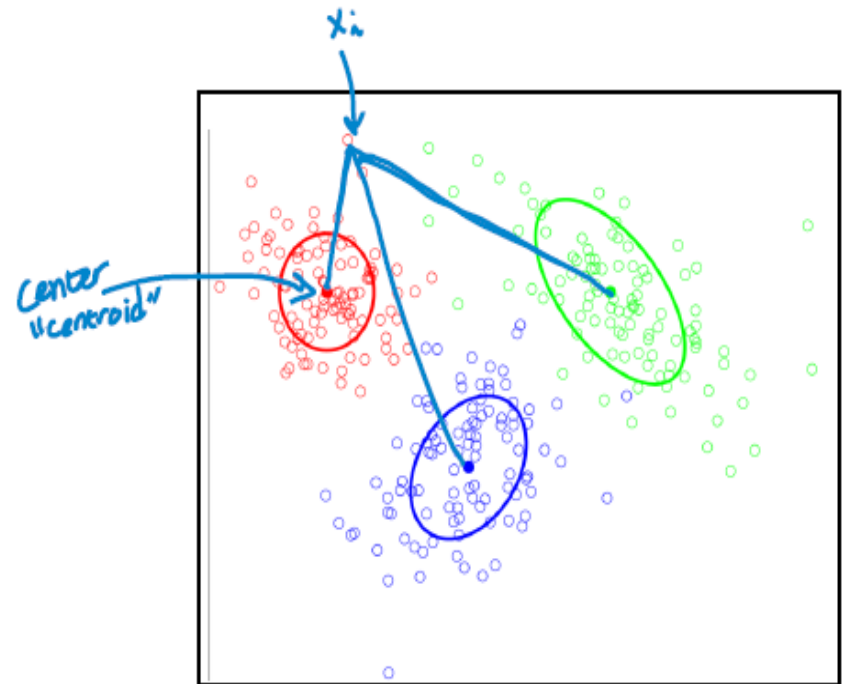
What defines a cluster ?

13

Cluster defined by
center & shape/spread

Assign observation \mathbf{x}_i (doc)
to cluster k (topic label) if

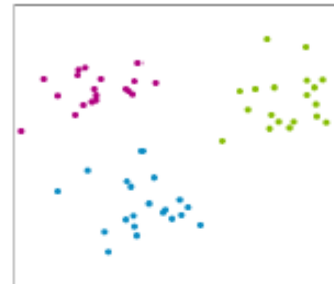
- Score under cluster k is higher than under others
- For simplicity, often define score as distance to cluster center (ignoring shape)



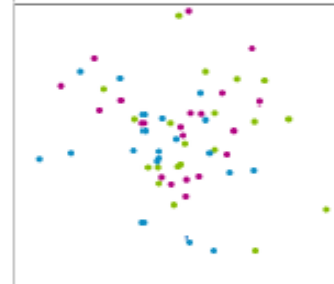
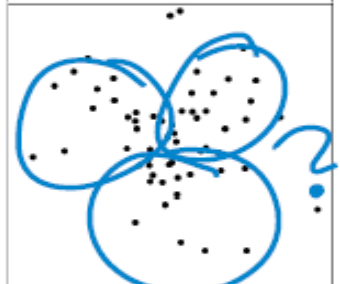
Hope for unsupervised learning

14

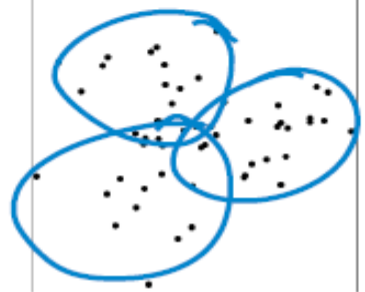
Easy



Impossible



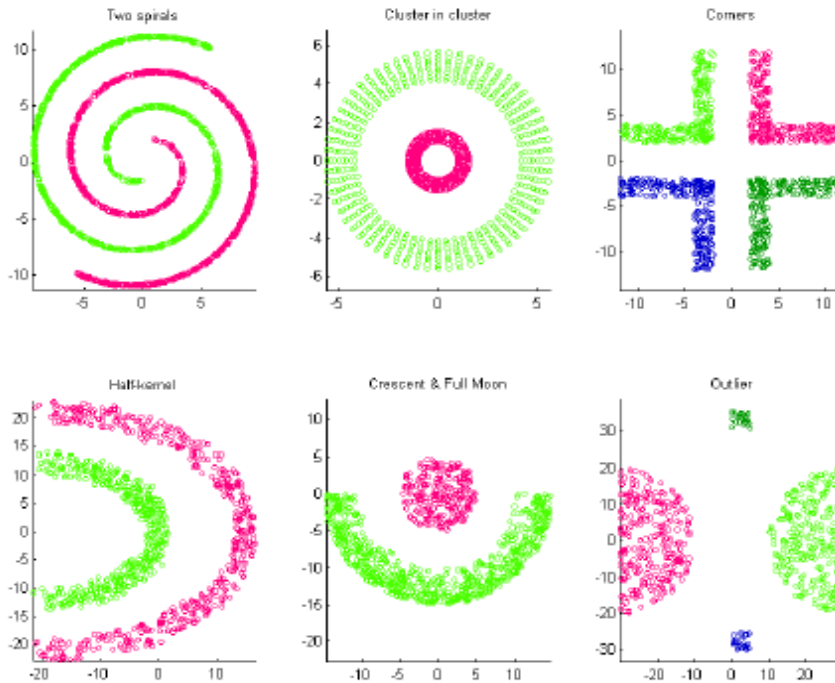
In between



Other (challenging!) clusters to discover

15

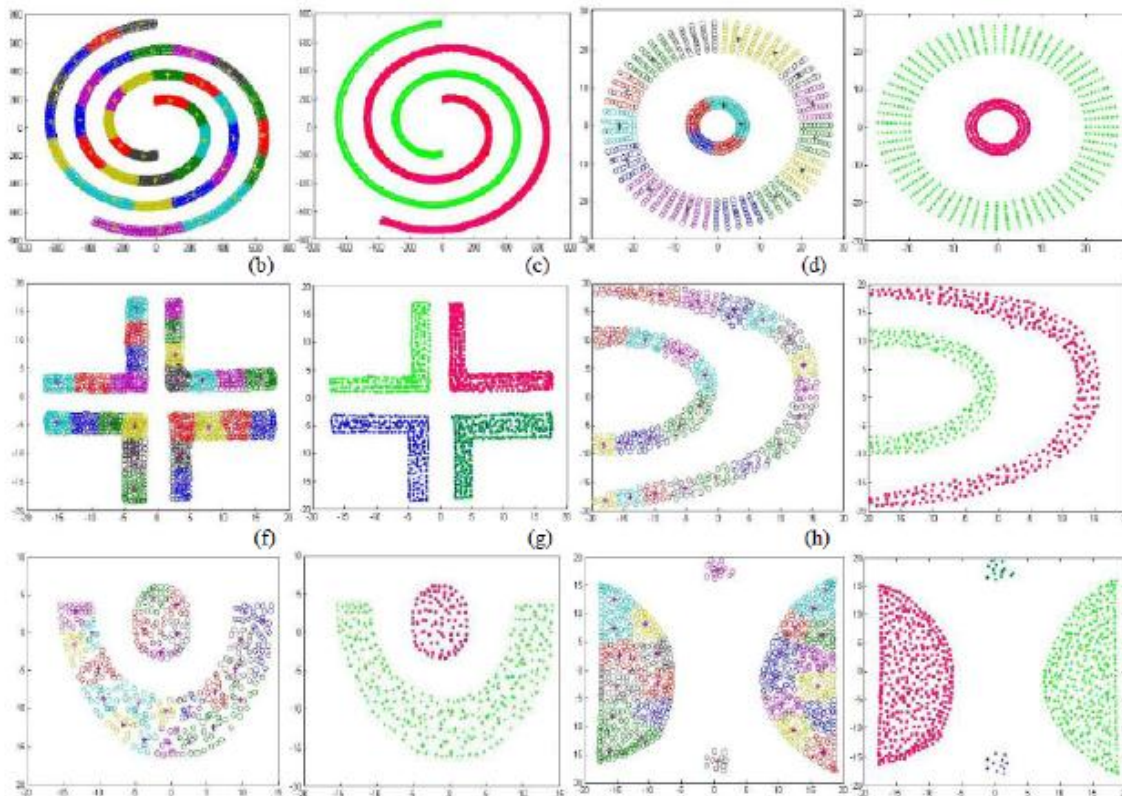
Analysed by your eyes



Other (challenging!) clusters to discover

16

Analysed by clustering algorithms



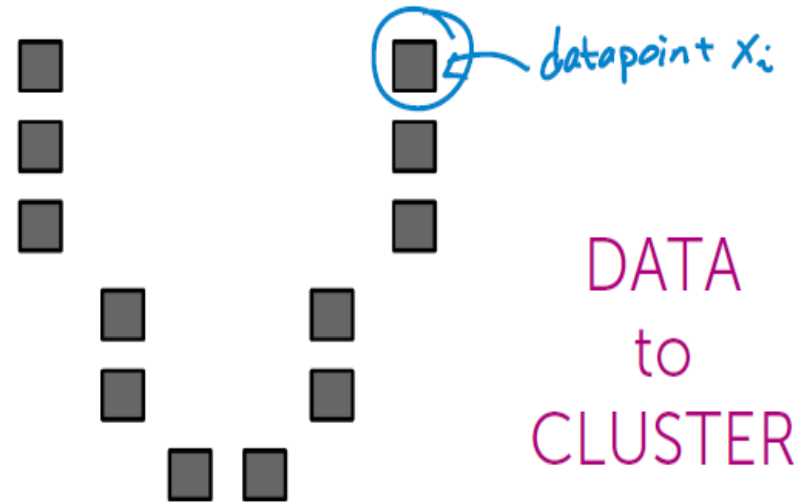
k-means clustering algorithm

k-means clustering algorithm

18

Assume

- Score = distance to cluster center
(smaller better)

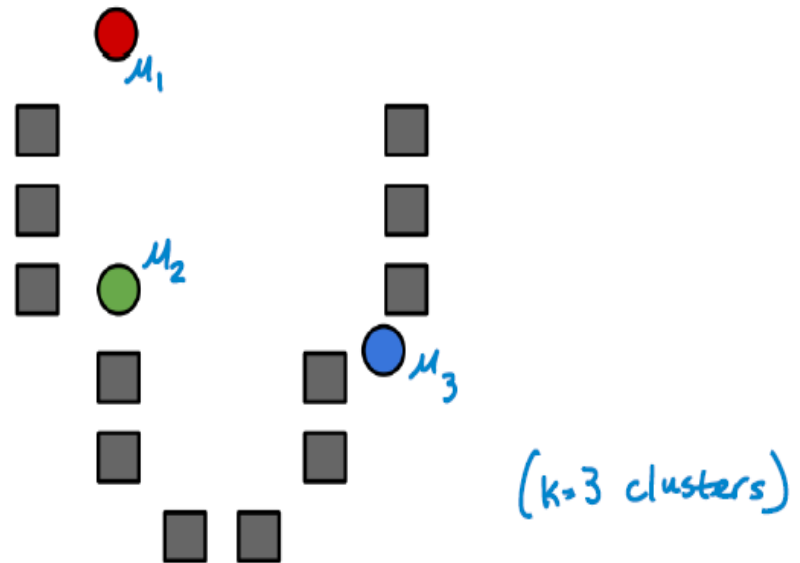


k-means clustering algorithm

19

0. Initialize cluster centers

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



k-means clustering algorithm

20

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

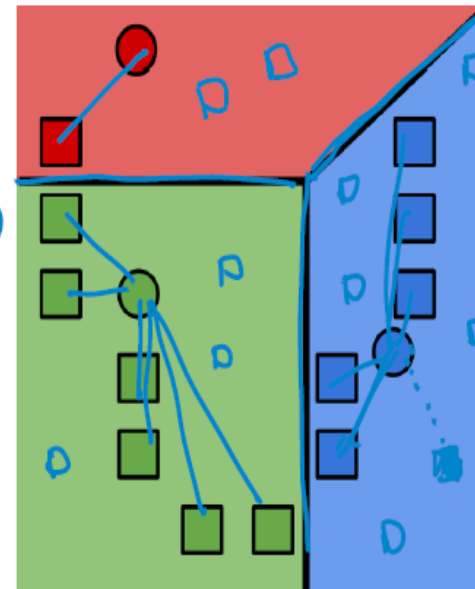
$$z_i \leftarrow \arg \min_j \|\mu_j - \mathbf{x}_i\|_2^2$$

z_i ← Inferred label for obs i , whereas supervised learning has given label y_i

j ← j th cluster center (varying)

i ← i th obs. (fixed)

return index j of the cluster whose center is closest to obs x_i (whereas min returning minimum value of $\|\cdot\|_2^2$)



Voronoi tessellation
(for visualization only... you don't need to compute this)

k-means clustering algorithm

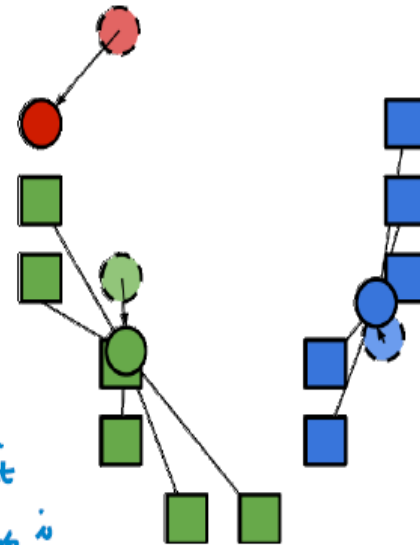
21

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

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

n_j ← # of obs. in cluster j

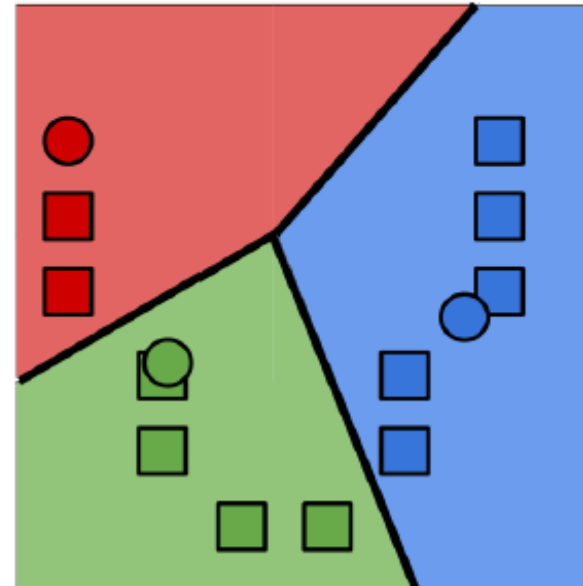
$i: z_i=j$ ← all obs. i such that $z_i=j$ (obs i is in cluster j)



k-means clustering algorithm

22

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



K-means as coordinate descent algorithm

23

1. Assign observations to closest cluster center

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

2. Revise cluster centers as mean of assigned observations

$$\mu_j \leftarrow \arg \min_{\mu} \sum_{i: z_i=j} \|\mu - \mathbf{x}_i\|_2^2$$

Alternating minimization
1. (z given μ) and 2. (μ given z)
= **coordinate descent**

Convergence of k-means

24

Converges to:

~~- Global optimum~~

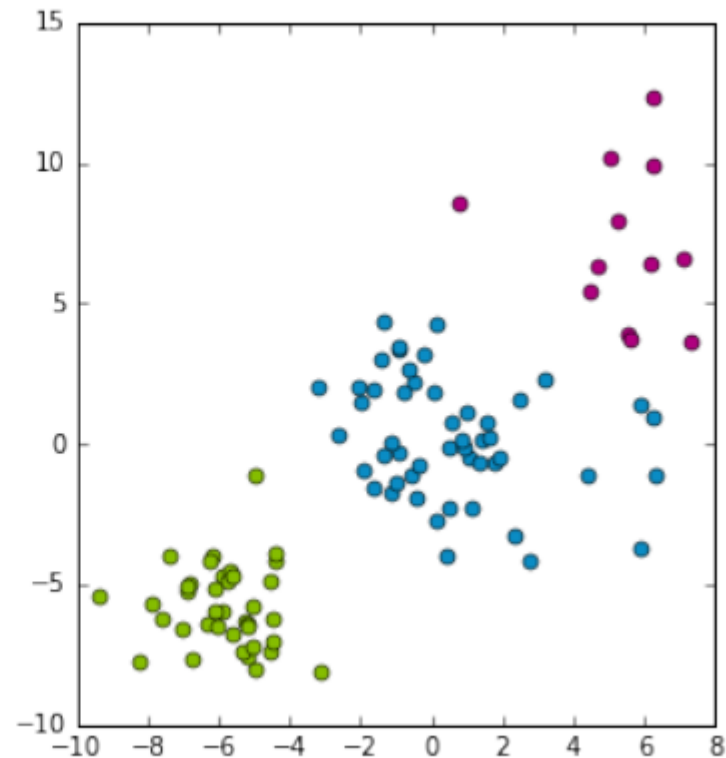
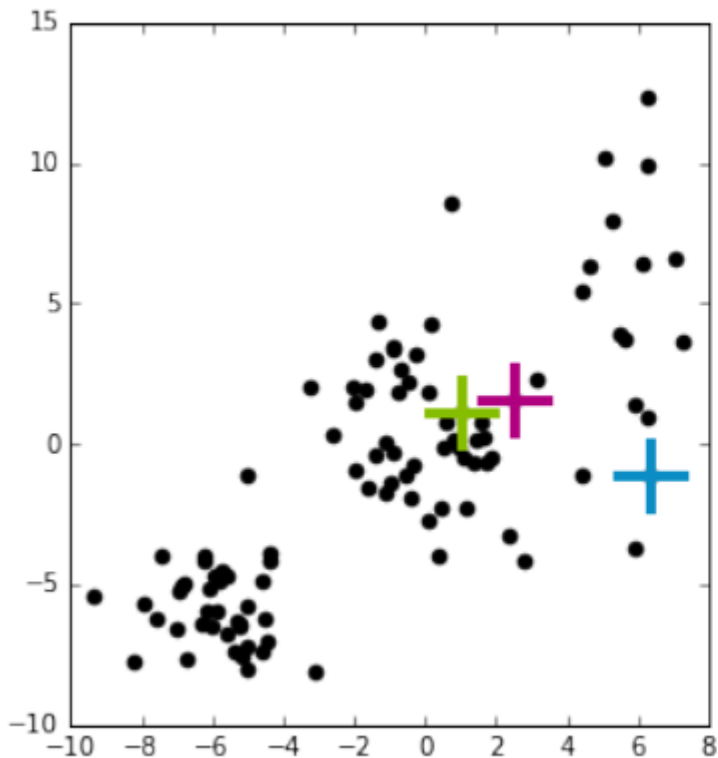
- Local optimum

~~- neither~~

Because we can cast k-means as coordinate descent algorithm we know that we are converging to local optimum

Convergence of k-mans to local mode

25



Crosses: initialised centers

Smart initialization: k-means++ overview

26

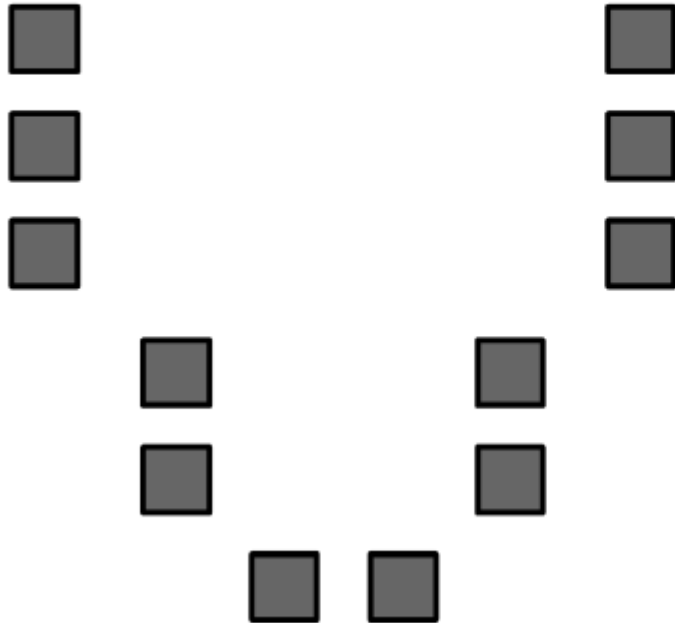
Initialization of k-means algorithm is critical to quality of local optima found

Smart initialization:

1. Choose first cluster center uniformly at random from data points
2. For each obs \mathbf{x} , compute distance $d(\mathbf{x})$ to nearest cluster center
3. Choose new cluster center from amongst data points, with probability of \mathbf{x} being chosen proportional to $d(\mathbf{x})^2$
4. Repeat Steps 2 and 3 until k centers have been chosen

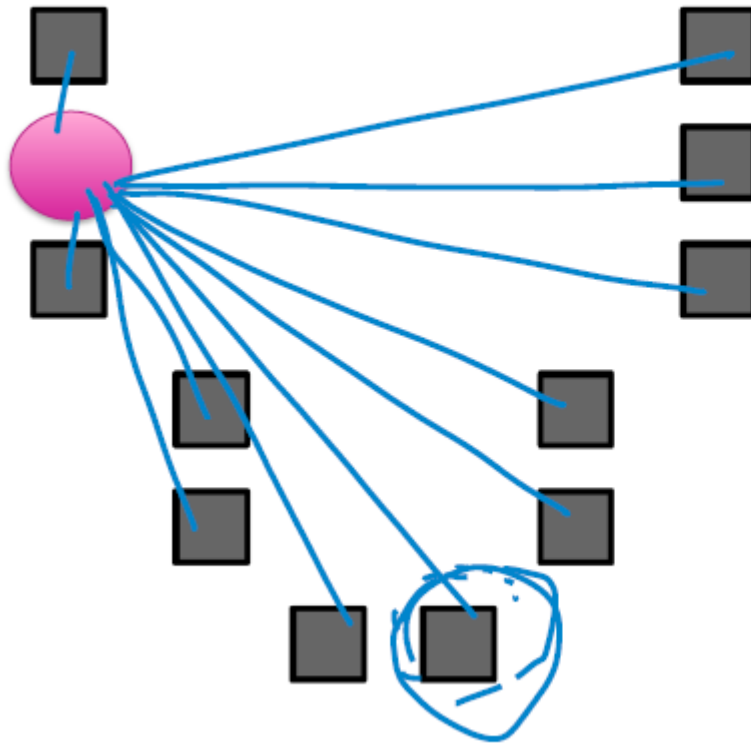
k-means++ visualised

27



k-means++ visualised

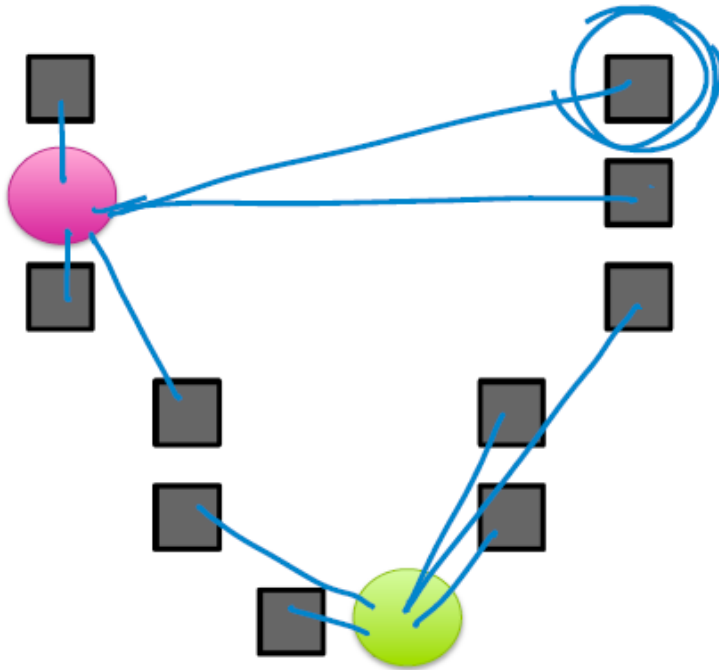
28



more likely to
select a datapoint
as a cluster center
if that datapoint is
far away
(dist^2 increases
this effect)

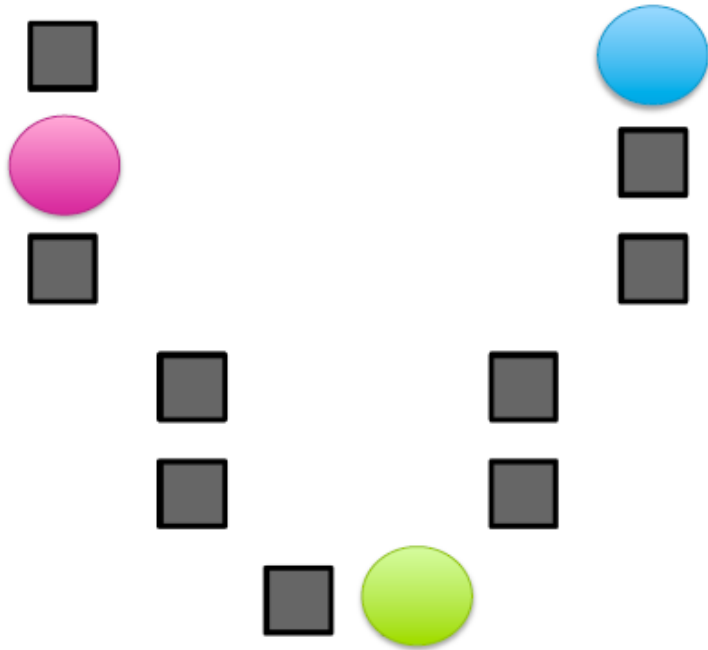
k-means++ visualised

29



k-means++ visualised

30



Smart initialisation: k-means++ overview

31

k-means++ pros/cons

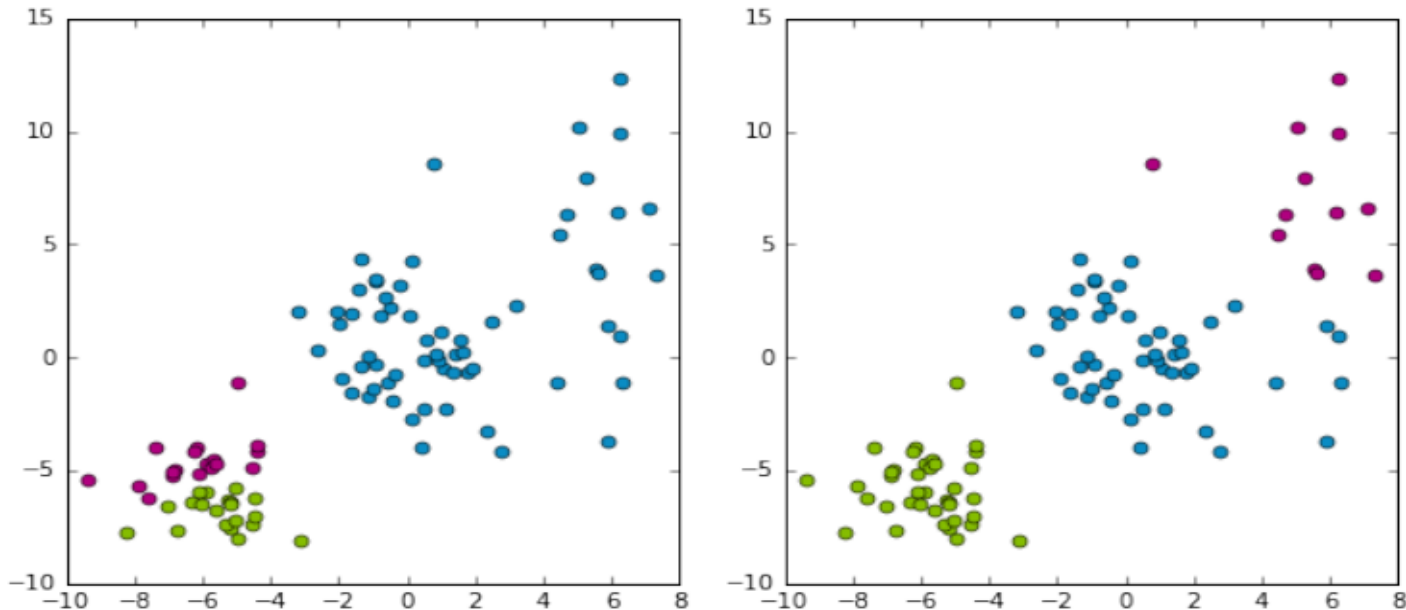
Computationally costly relative to random initialization, but the subsequent k-means often converges more rapidly

Tends to improve quality of local optimum and lower runtime

Assessing quality of the clustering

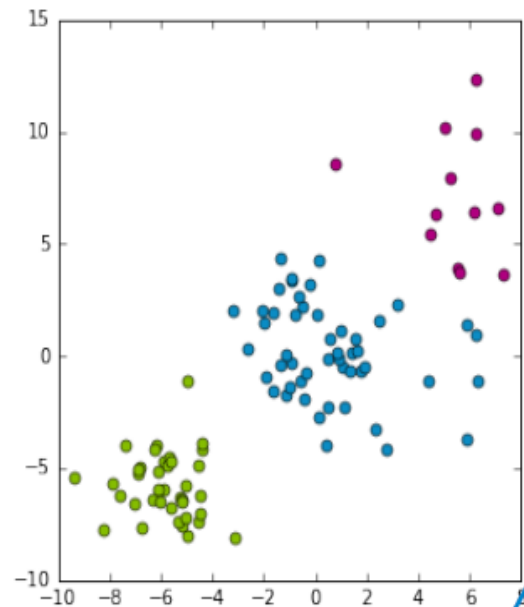
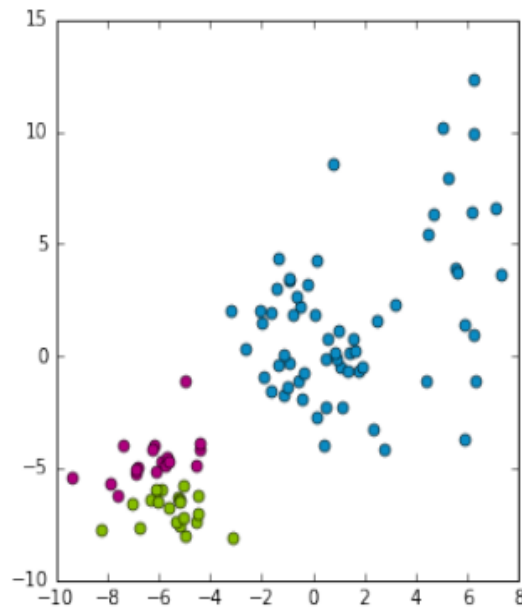
32

Which clustering do I prefer?



K-means objective

33



k-means is trying to minimize the **sum of squared distances**:

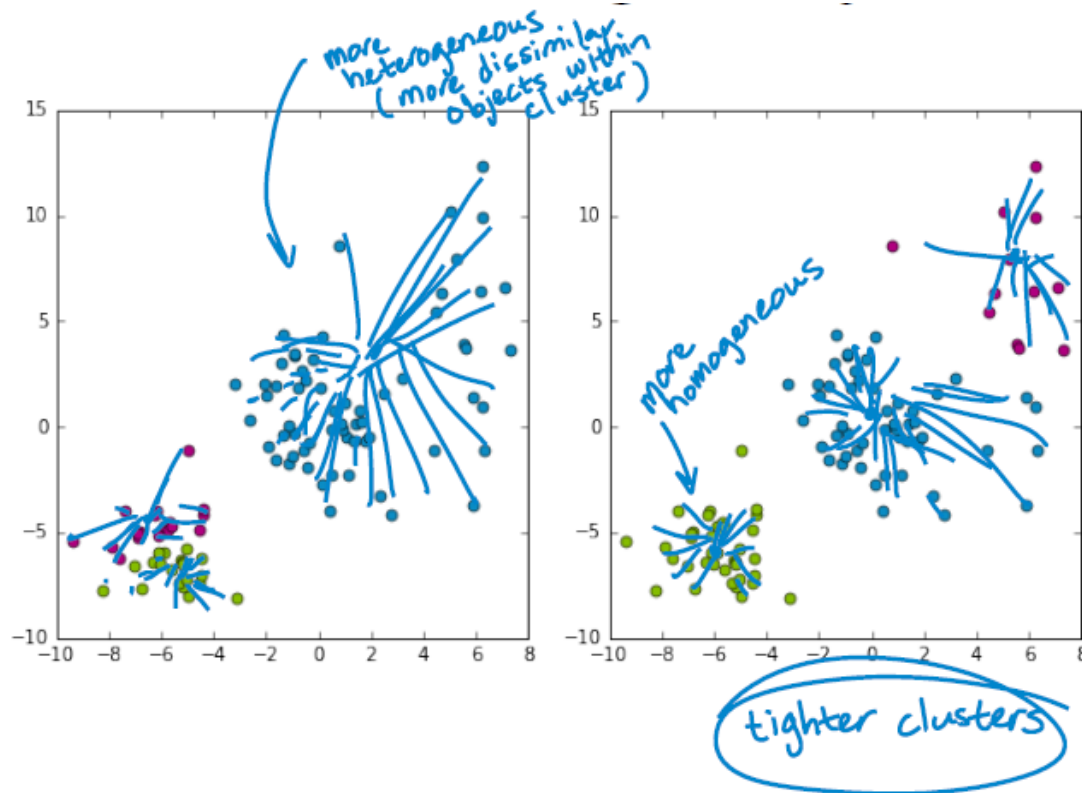
$$\sum_{j=1}^k \sum_{i:z_i=j} \|\mu_j - \mathbf{x}_i\|_2^2$$

sum over all clusters (pointing to k)
sum of squared distances in cluster j (pointing to the inner sum)

Min $\sum_{z_i} \sum_{\mu_j} \|\mu_j - \mathbf{x}_i\|_2^2$

Cluster heterogeneity

34



Measure of quality of given clustering:

$$\sum_{j=1}^k \sum_{i:z_i=j} \|\mu_j - \mathbf{x}_i\|_2^2$$

Lower is better!

What happens to heterogeneity as k increases?

35

Can refine clusters more and more to the data
→ overfitting!

Extreme case of $k=N$:

- can set each cluster center equal to datapoint
- heterogeneity = 0 !

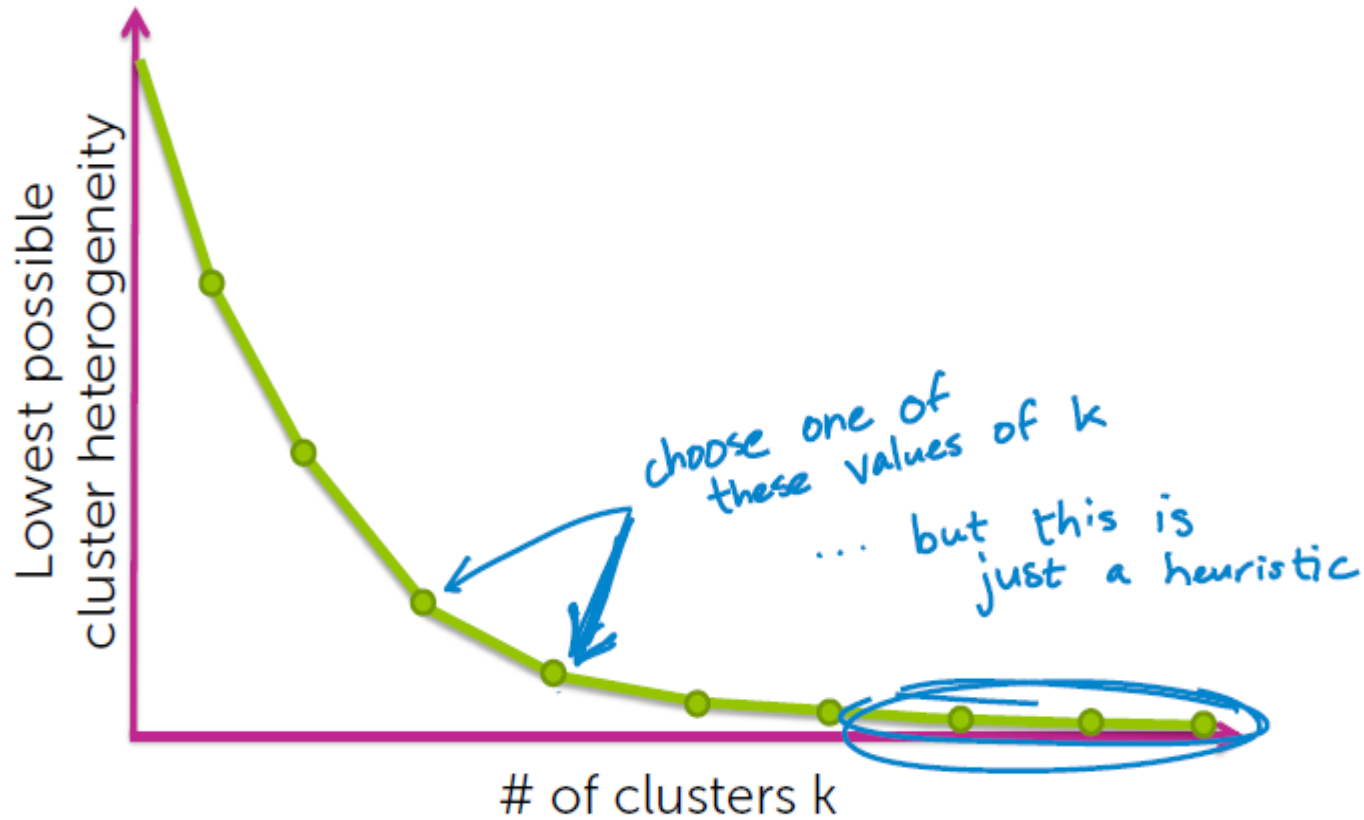
of observations

(all distances to cluster centers are 0)

Lowest possible cluster heterogeneity
decreases with increasing k

How to choose k?

36



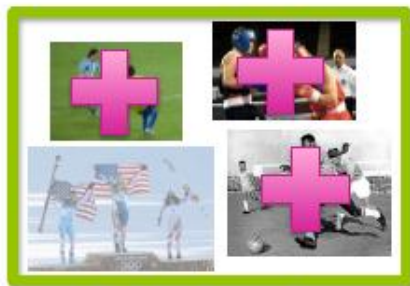
Probabilistic approach: mixture model

Why probabilistic approach?

38

Learn user preferences

Set of clustered documents read by user



Cluster 1



Cluster 2



Cluster 3



Cluster 4

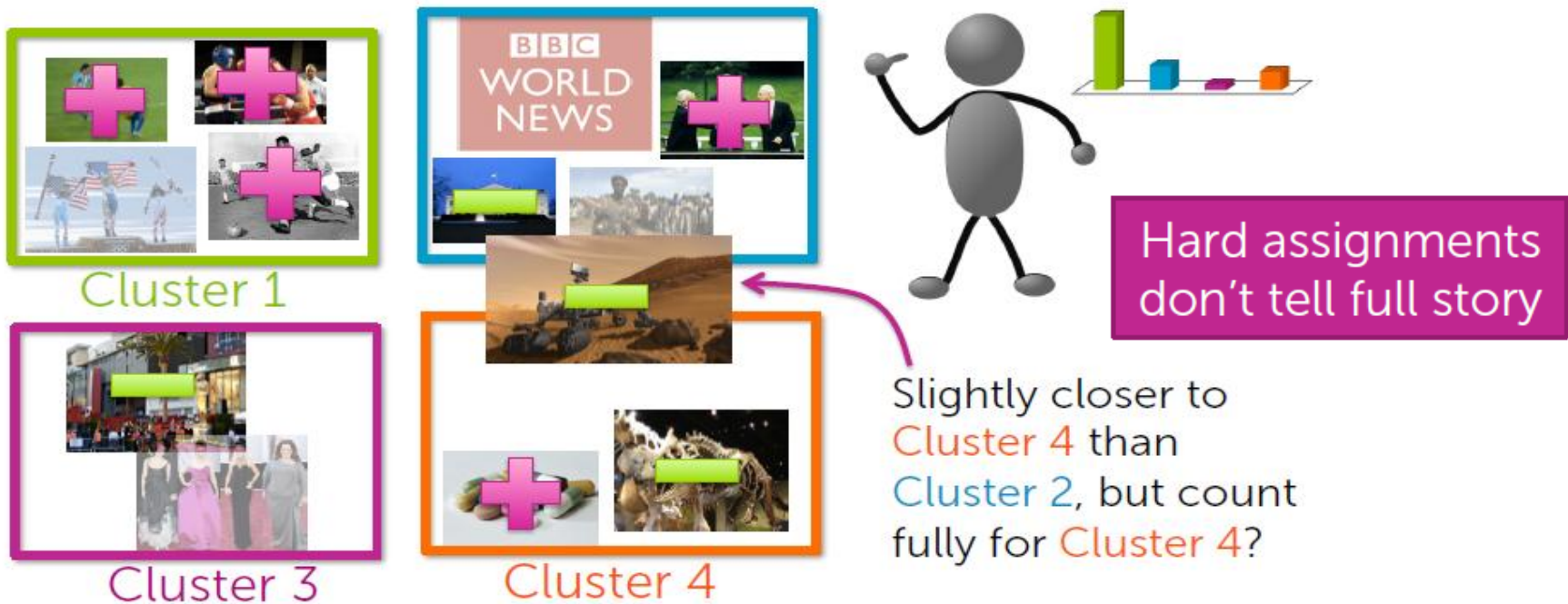


Use feedback
to learn user
preferences
over topics

Why probabilistic approach?

39

Uncertainty in cluster assignments



Why probabilistic approach?

40

Other limitations of k-means

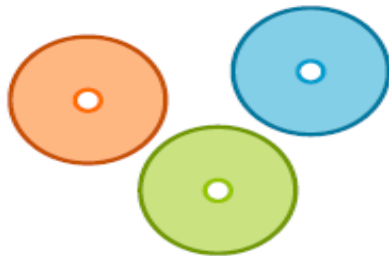
Assign observations to closest cluster center

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

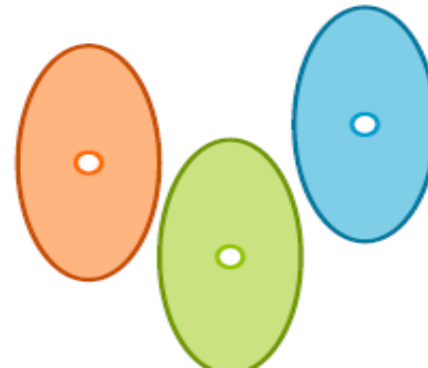
Can use weighted Euclidean, but requires *known* weights

Only center matters

Equivalent to assuming spherically symmetric clusters



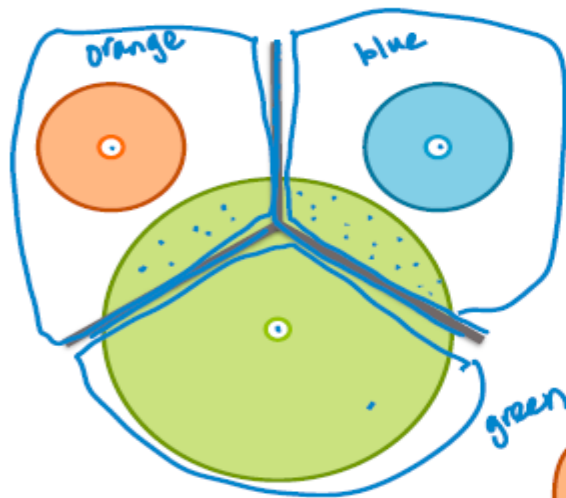
Still assumes all clusters have the same axis-aligned ellipses



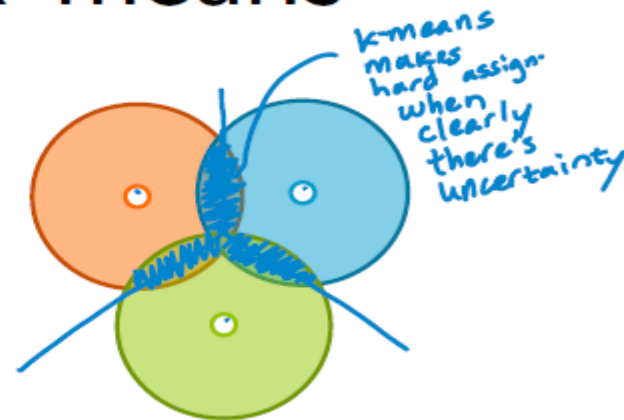
Why probabilistic approach?

41

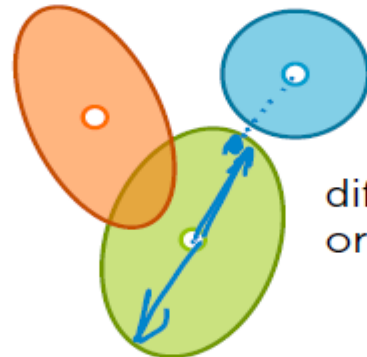
Failure modes of k-means



disparate cluster sizes



overlapping clusters



different shaped/
oriented clusters

Mixture models

42

- Provides **soft assignments** of observations to clusters (uncertainty in assignment)
 - e.g., 54% chance document is **world news**, 45% **science**, 1% **sports**, and 0% **entertainment**
- Accounts for cluster **shapes** not just **centers**
- Enables **learning weightings** of dimensions
 - e.g., how much to weight each word in the vocabulary when computing cluster assignment

Application: clustering images

43

Discover groups of similar images

- Ocean
- Pink flower
- Dog
- Sunset
- Clouds
- ...



Application: clustering images

44

Simple image representation

Consider average red, green, blue pixel intensities



[R = 0.05, G = 0.7, B = 0.9]



[R = 0.85, G = 0.05, B = 0.35]



[R = 0.02, G = 0.95, B = 0.4]

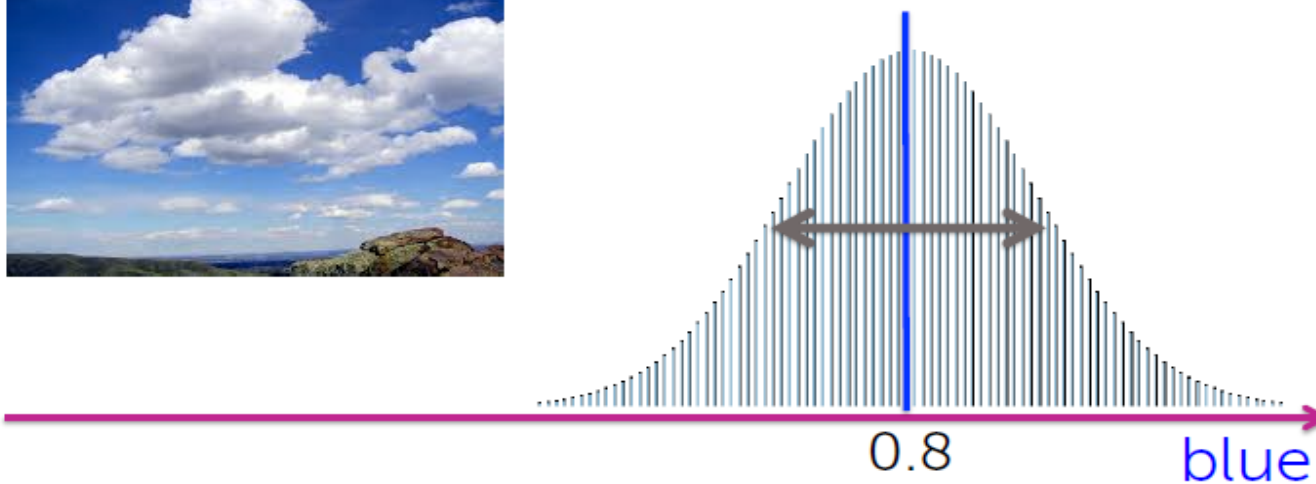
Single RGB vector per image

Application: clustering images

45

Distribution over all **cloud** images

Let's look at just the **blue** dimension

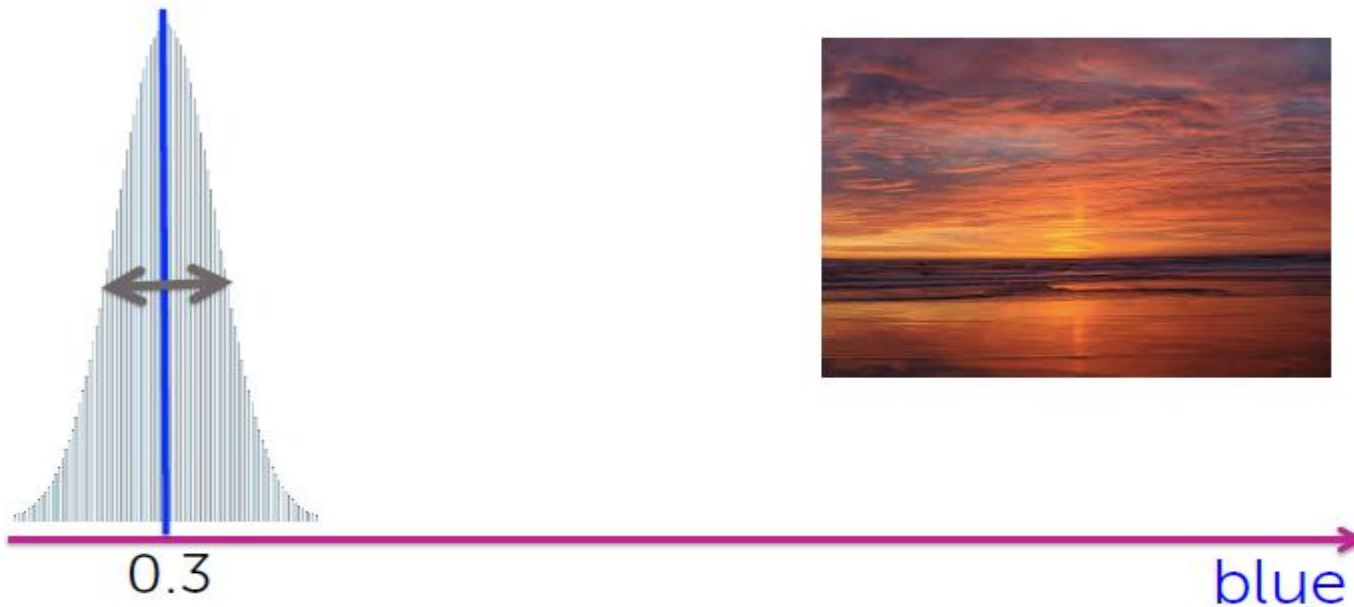


Application: clustering images

46

Distribution over all sunset images

Let's look at just the blue dimension



Application: clustering images

47

Distribution over all forest images

Let's look at just the blue dimension

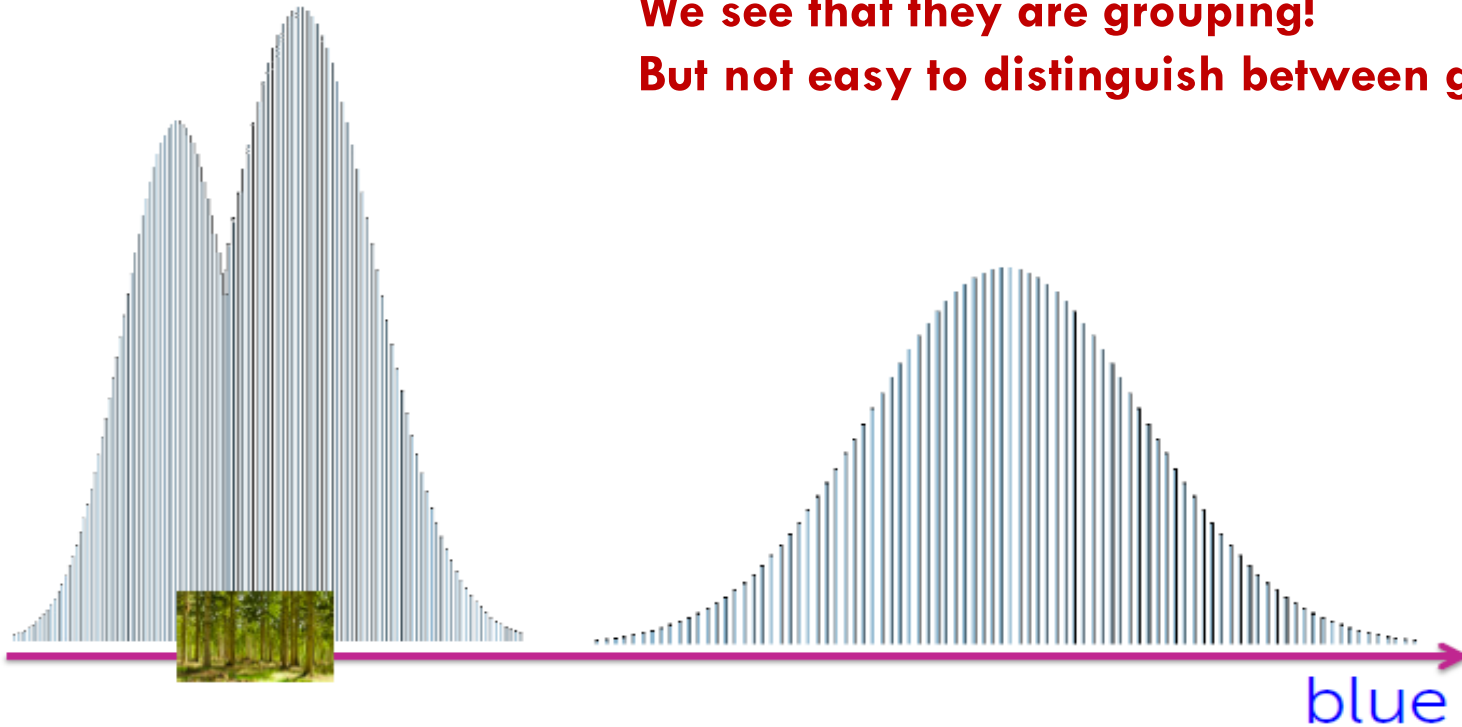


Application: clustering images

48

Distribution over **all** images

We see that they are grouping!
But not easy to distinguish between groups



Application: clustering images

49

Can be distinguished along other dim

Now look at the **red** dimension



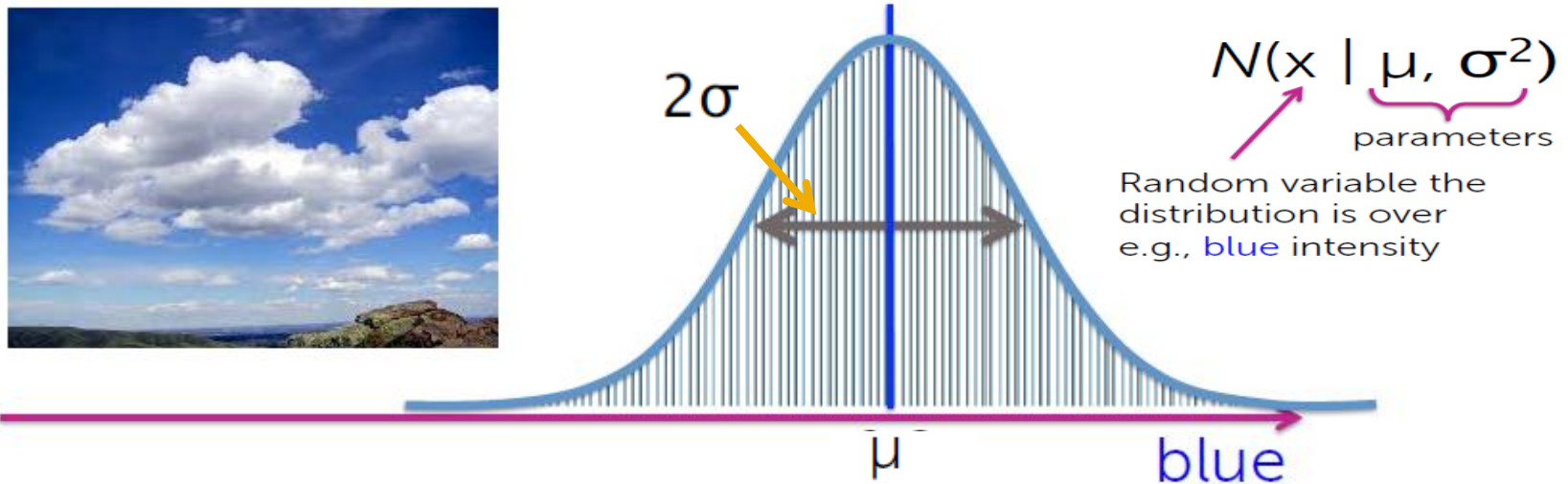
**In this dimension
separable groups!**



Model for a given image type

50

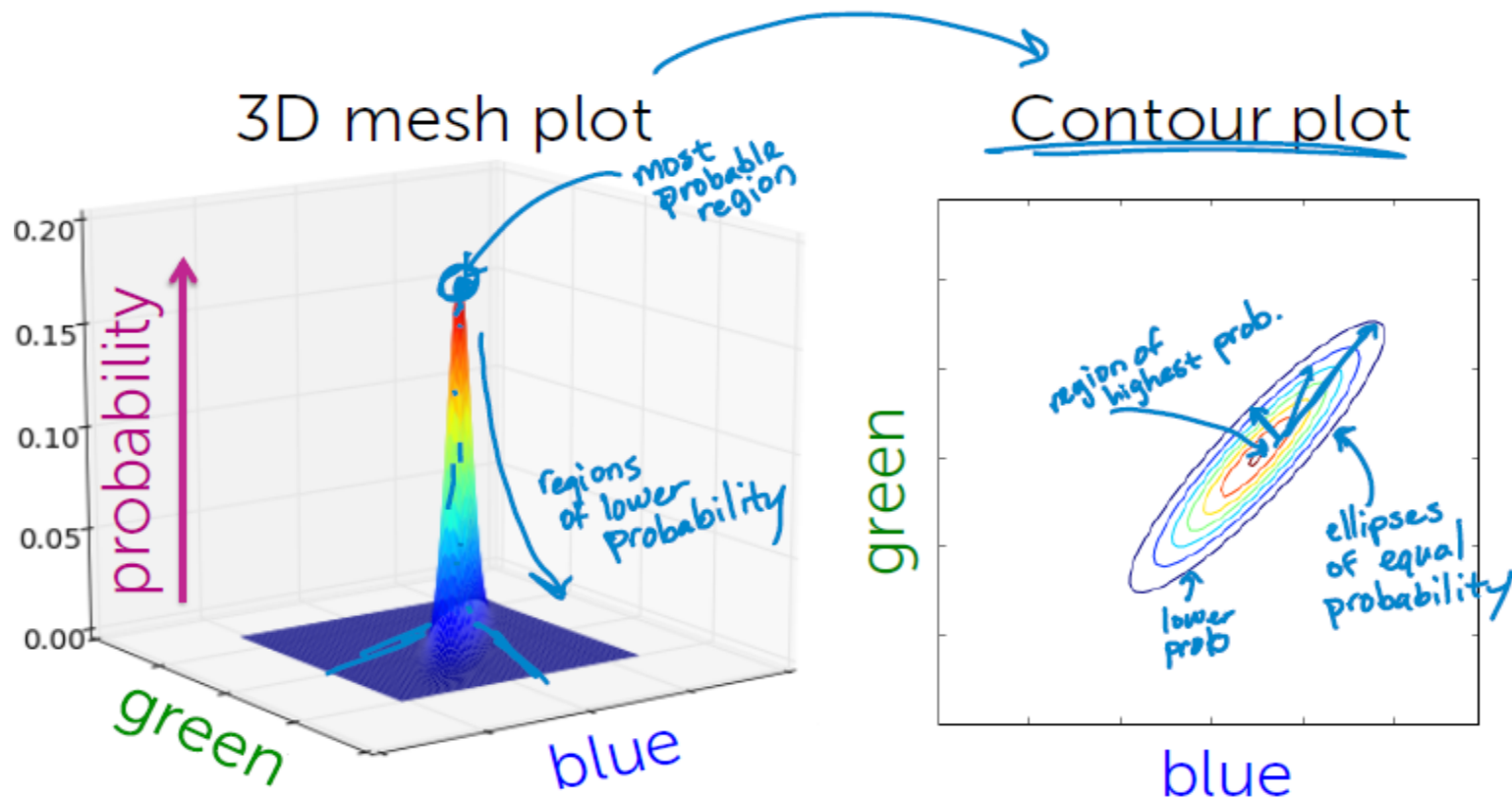
For **each dimension** of the [R, G, B] vector,
and **each image type**, assume a
Gaussian distribution over color intensity



Model for a given image type

51

2D Gaussians – Bird's eye view



Application: clustering images

52

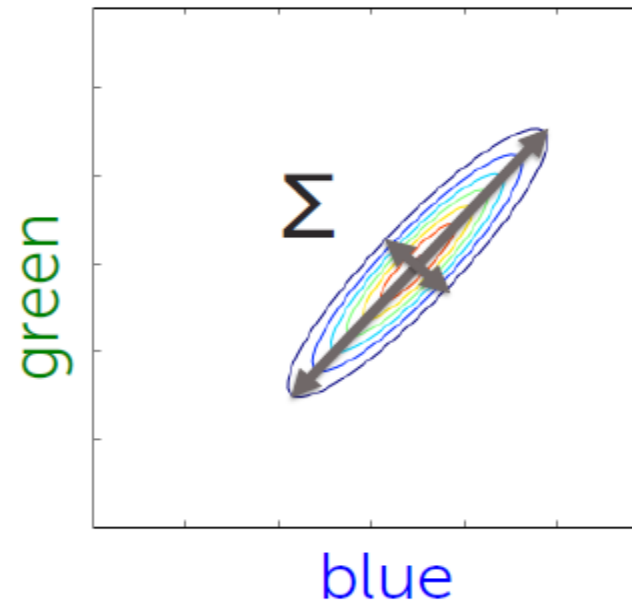
2D Gaussians – Parameters

Fully specified by **mean** μ and **covariance** Σ

$$\mu = [\mu_{\text{blue}}, \mu_{\text{green}}]$$

$$\Sigma = \begin{pmatrix} \sigma_{\text{blue}}^2 & \sigma_{\text{blue,green}} \\ \sigma_{\text{green,blue}} & \sigma_{\text{green}}^2 \end{pmatrix}$$

covariance determines
orientation + spread

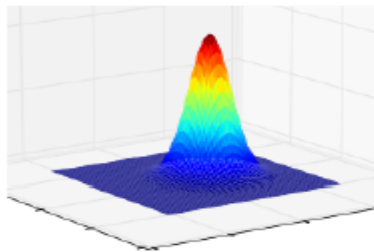
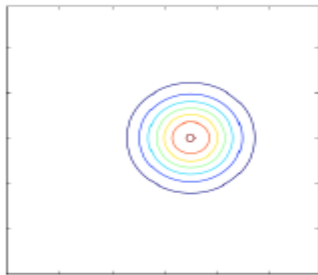


Application: clustering images

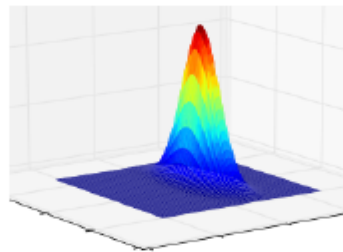
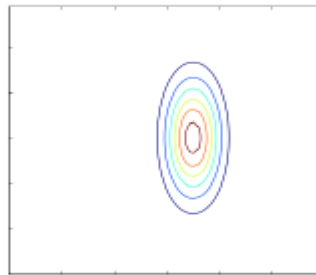
53

Covariance structures

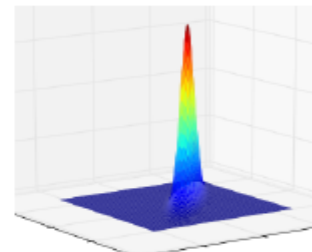
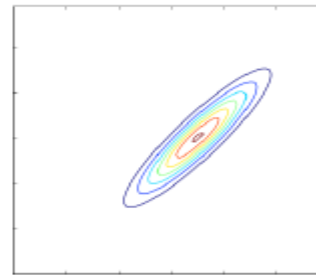
$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$



$$\Sigma = \begin{pmatrix} \sigma_B^2 & 0 \\ 0 & \sigma_G^2 \end{pmatrix}$$



$$\Sigma = \begin{pmatrix} \sigma_B^2 & \sigma_{B,G} \\ \sigma_{G,B} & \sigma_G^2 \end{pmatrix}$$



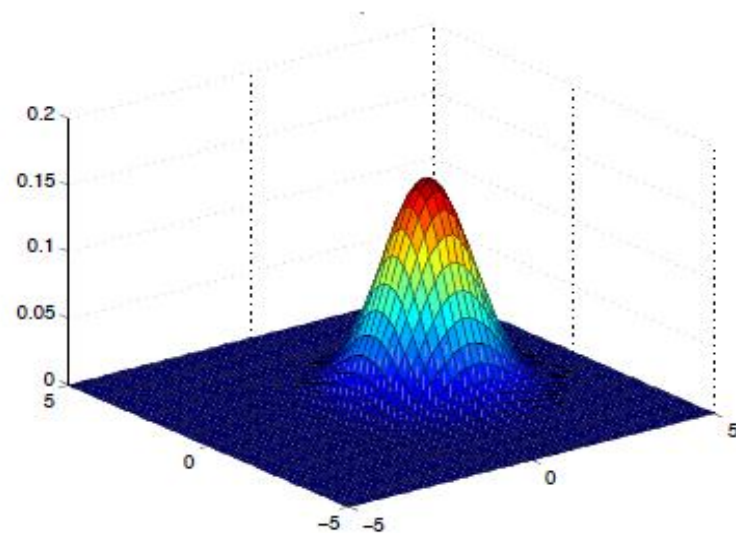
Application: clustering images

54

Notating a multivariate Gaussian

$$N(\mathbf{x} \mid \underbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}}_{\text{parameters}})$$

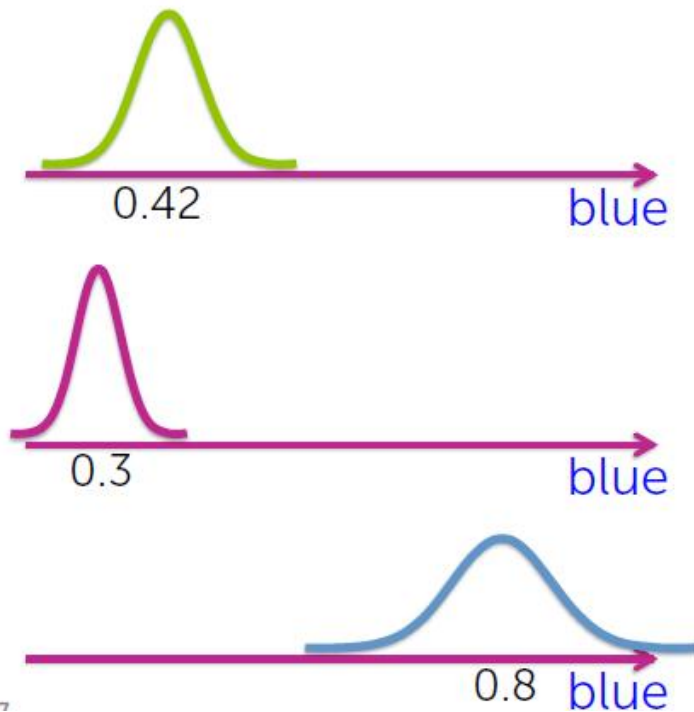
Random vector
e.g., [R, G, B] intensities



Mixture of Gaussians

55

Model as Gaussian per category/cluster

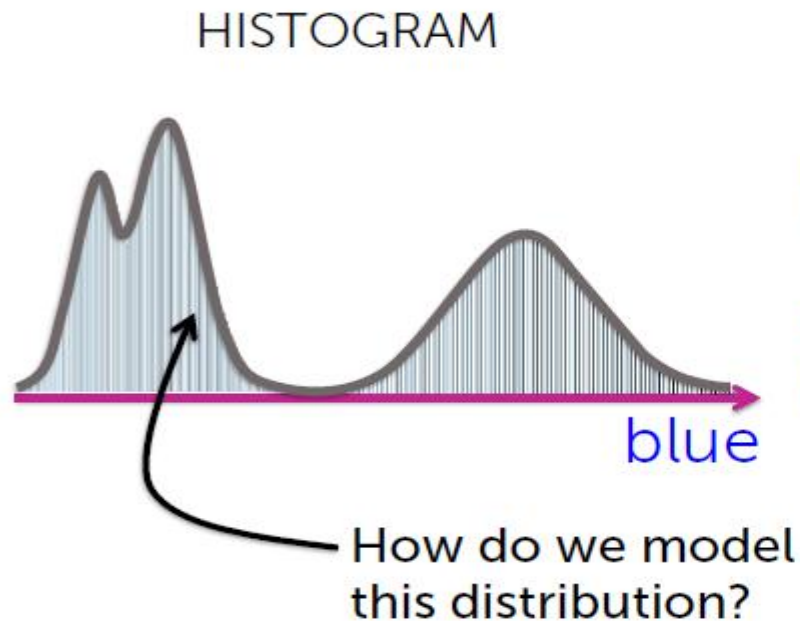


27

Mixture of Gaussians

56

Jumble of unlabeled images

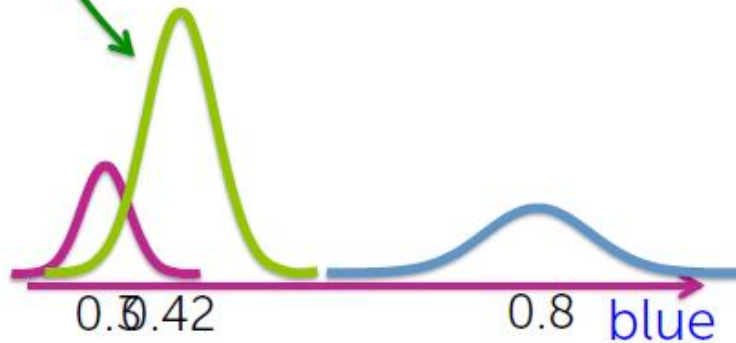


Mixture of Gaussians

57

What if image types not equally represented?

e.g., forest images are very likely in the collection

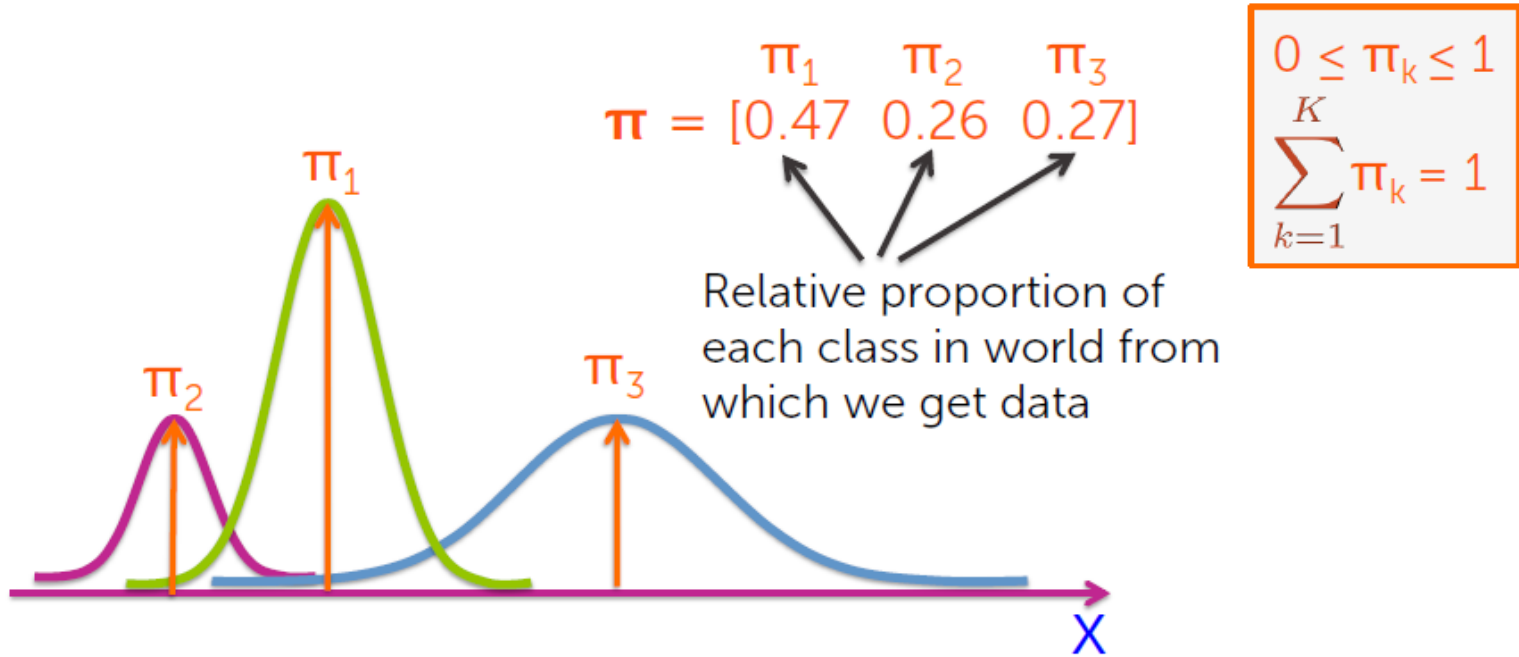


Mixture of Gaussians

58

Combination of weighted Gaussians

Associate a weight π_k with each Gaussian component

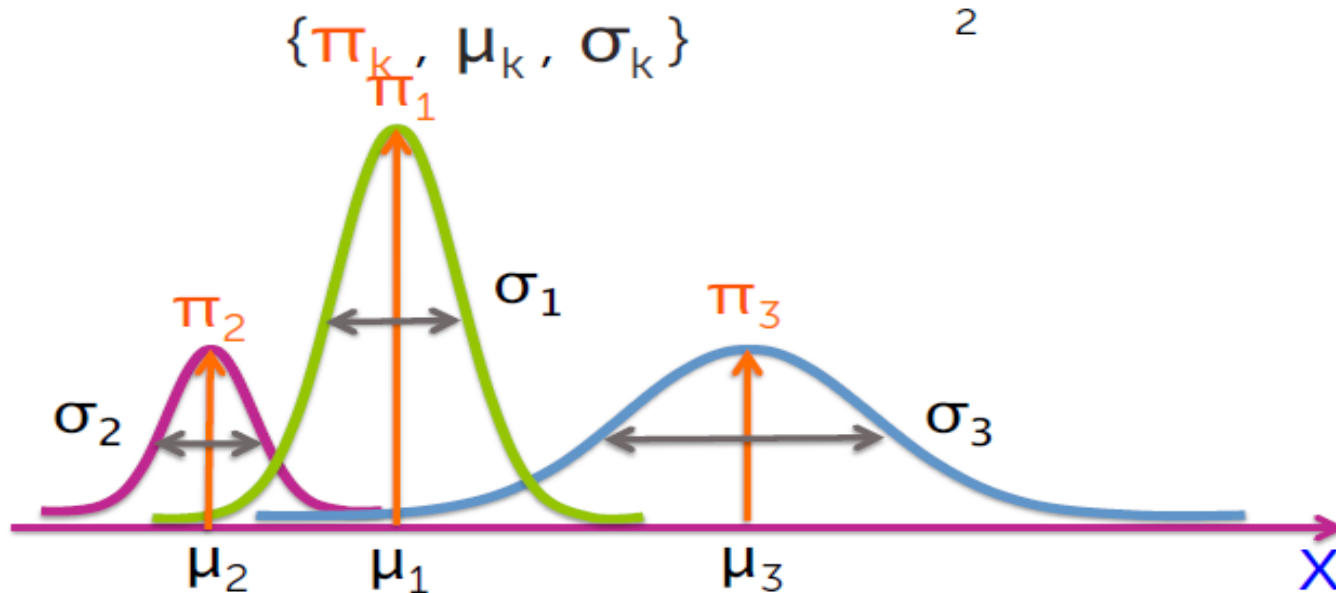


Mixture of Gaussians

59

Mixture of Gaussians (1D)

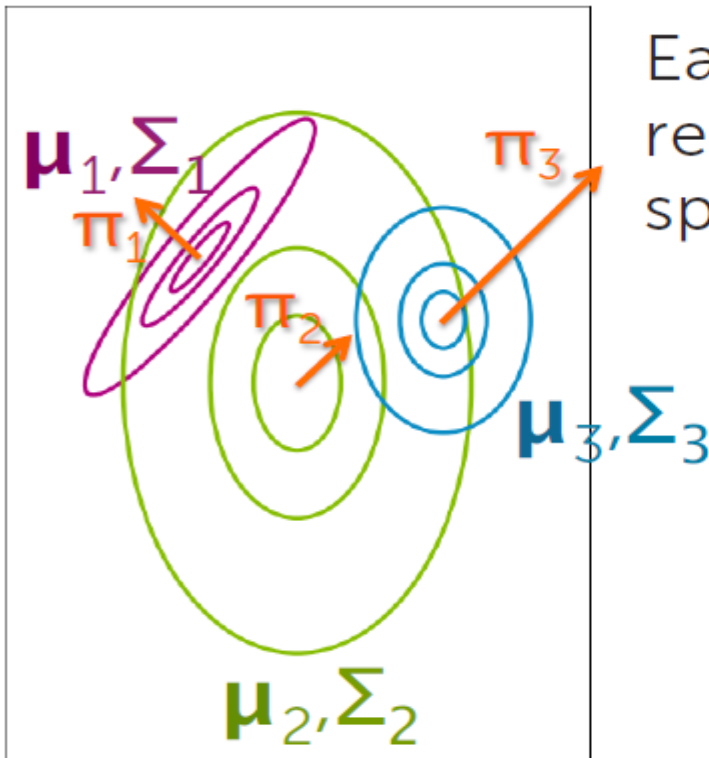
Each mixture component represents a unique cluster specified by:



Mixture of Gaussians

60

Mixture of Gaussians (general)



Each mixture component represents a unique cluster specified by:

$$\{\pi_k, \mu_k, \Sigma_k\}$$

Application: clustering documents

62

Discover groups of related documents



Application: clustering documents

63

Document representation



$x_i =$

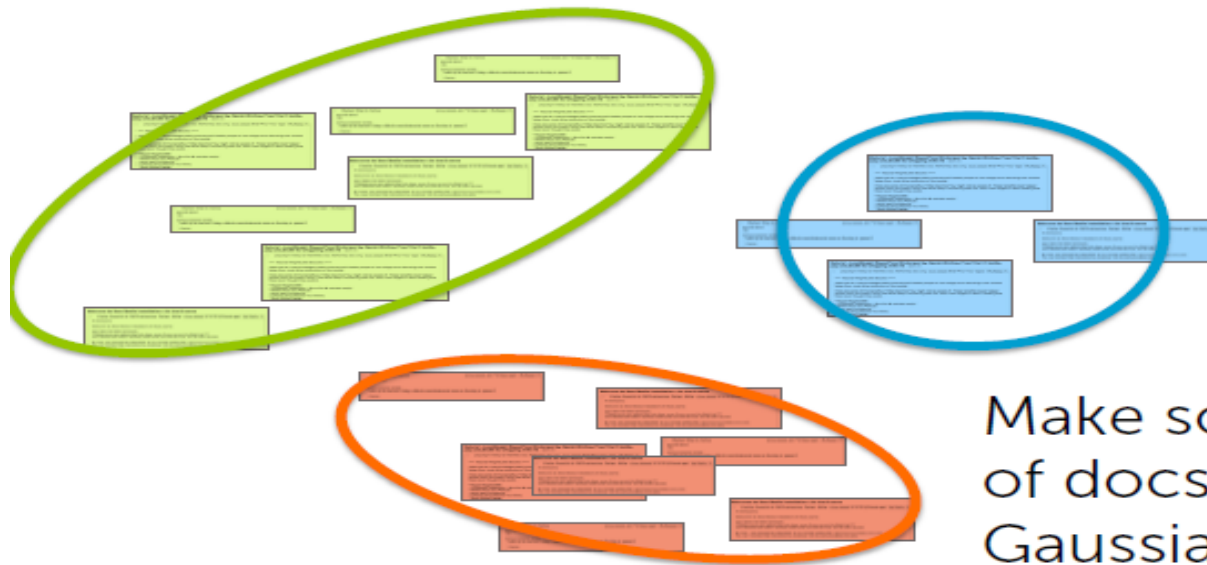


Application: clustering documents

64

Mixture of Gaussians for clustering documents

Space of all documents
(really lives in \mathbf{R}^V for vocab size V)



Application: clustering documents

65

Counting parameters

Each cluster has $\{\pi_k, \mu_k, \Sigma_k\}$



In 2D:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{pmatrix}$$

Application: clustering documents

66

Counting parameters

Each cluster has $\{\boldsymbol{\pi}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$



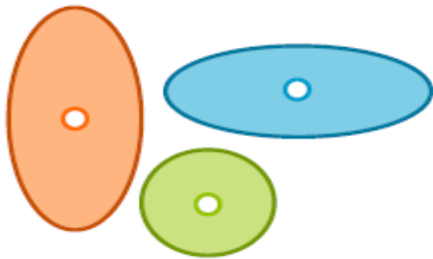
In V (vocab size) dims:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \frac{V(V+1)}{2} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

Application: clustering documents

68

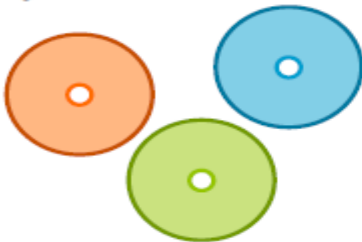
Restrictive assumption, but...



- Can **learn** weights on dimensions (e.g., weights on words in vocab)
- Can learn **cluster-specific** weights on dimensions

Still more flexible than k-means

Spherically symmetric clusters



Specify weights...

All clusters have same axis-aligned ellipses

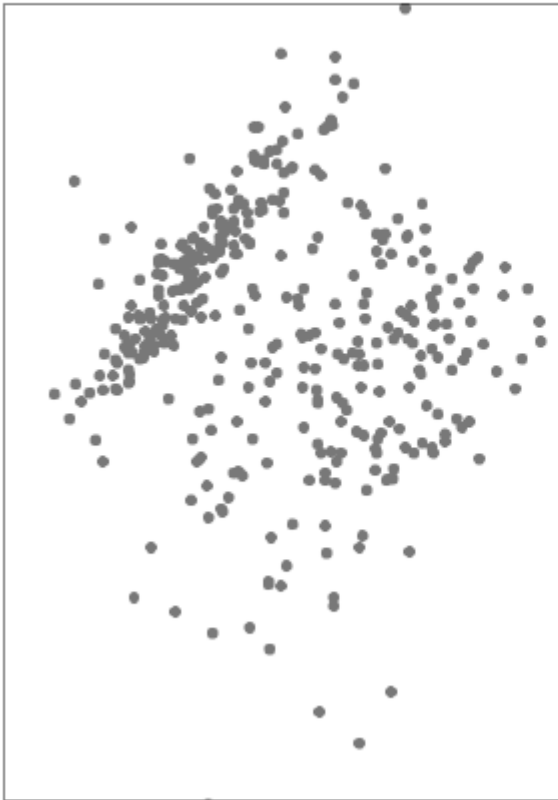
44

Inferring soft assignments with expectation maximization (EM)

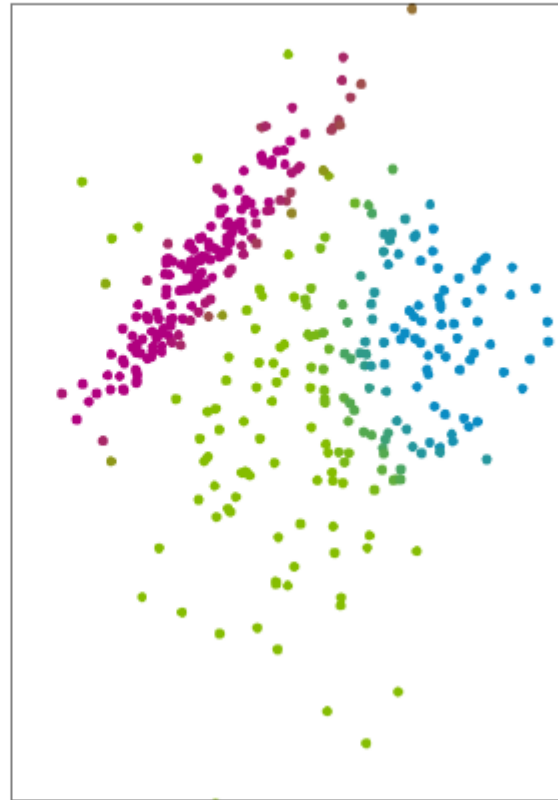
Inferring cluster labels

70

Data



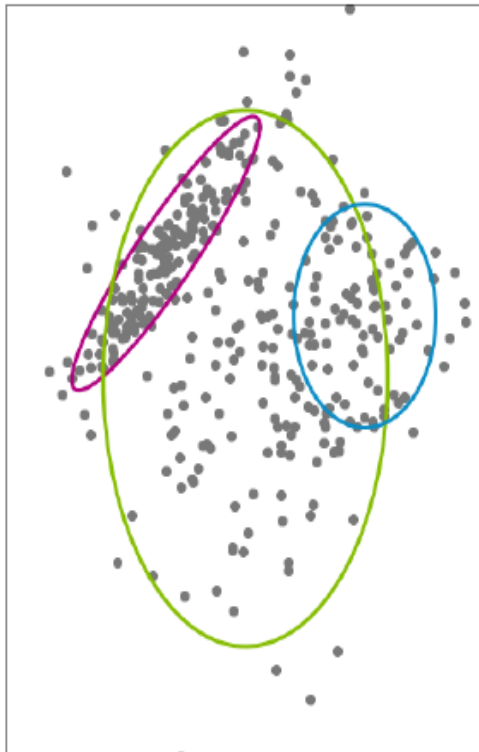
Desired soft assignments



What if we knew the cluster parameters $\{\pi_k, \mu_k, \Sigma_k\}$?

71

Compute responsibilities



$r_i = [r_{i1} \ r_{i2} \ \dots \ r_{iK}]$ # clusters

Responsibility cluster k takes for observation i

$$\underline{r_{ik}} = p(\underline{z_i = k} \mid \underbrace{\{\pi_j, \mu_j, \Sigma_j\}_{j=1}^K}_{\text{fixed values defining the distribution}}, \underline{x_i})$$

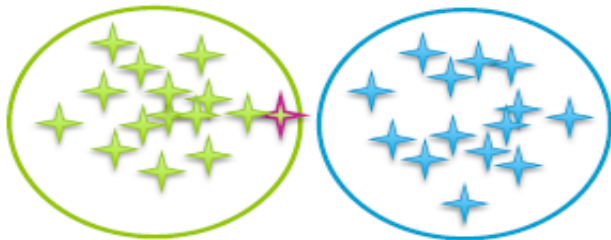
random variable probability of assignment to cluster k

given model parameters and observed value

What if we knew the cluster parameters $\{\pi_k, \mu_k, \Sigma_k\}$?

72

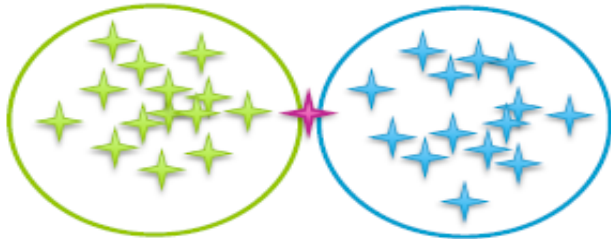
Responsibilities in pictures



Green cluster takes more responsibility



Blue cluster takes more responsibility



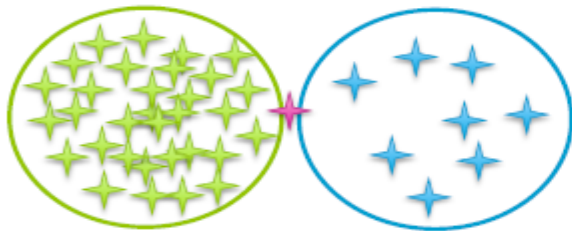
Uncertain... split responsibility

What if we knew the cluster parameters $\{\pi_k, \mu_k, \Sigma_k\}$?

73

Responsibilities in pictures

Need to weight by cluster probabilities, not just cluster shapes

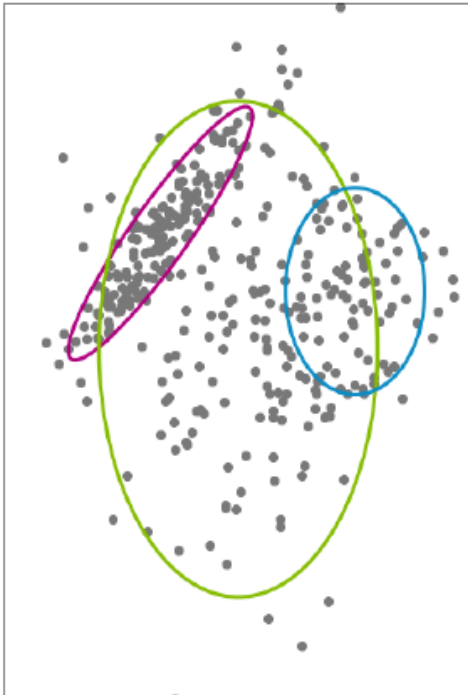


Still **uncertain**,
but **green** cluster seems
more probable...
takes more responsibility

What if we knew the cluster parameters $\{\pi_k, \mu_k, \Sigma_k\}$?

74

Responsibilities in equations



Responsibility cluster k takes for observation i

$$r_{ik} = \pi_k N(x_i | \mu_k, \Sigma_k)$$

Initial probability of being from cluster k

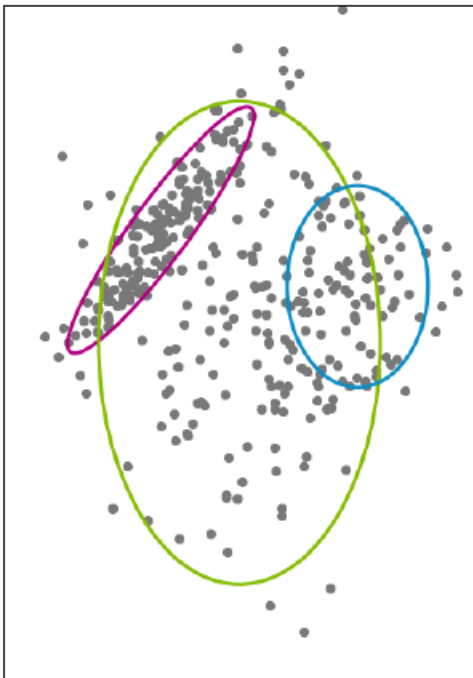
How likely is the observed value x_i under this cluster assignment?

very unlikely under the green cluster, even though the prior on green is higher

What if we knew the cluster parameters $\{\pi_k, \mu_k, \Sigma_k\}$?

75

Responsibilities in equations



Responsibility cluster k takes for observation i

$$r_{ik} = \frac{\pi_k N(x_i | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_i | \mu_j, \Sigma_j)}$$

Normalized over all possible cluster assignments

What if we knew the cluster parameters $\{\pi_k, \mu_k, \Sigma_k\}$?

76

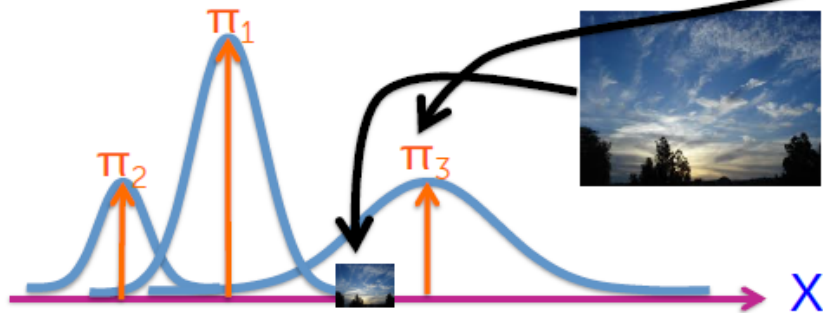
Recall: According to the model...

Without observing the image content, what's the probability it's from cluster k ? (e.g., prob. of seeing "clouds" image)

$$p(z_i = k) = \pi_k$$

Given observation \mathbf{x}_i is from cluster k , what's the likelihood of seeing \mathbf{x}_i ? (e.g., just look at distribution for "clouds")

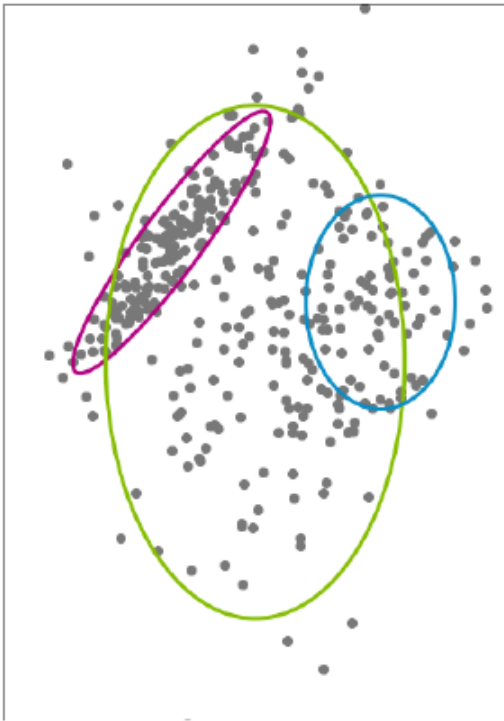
$$p(x_i | z_i = k, \mu_k, \Sigma_k) = N(x_i | \mu_k, \Sigma_k)$$



What if we knew the cluster parameters $\{\pi_k, \mu_k, \Sigma_k\}$?

77

Part 1: Summary



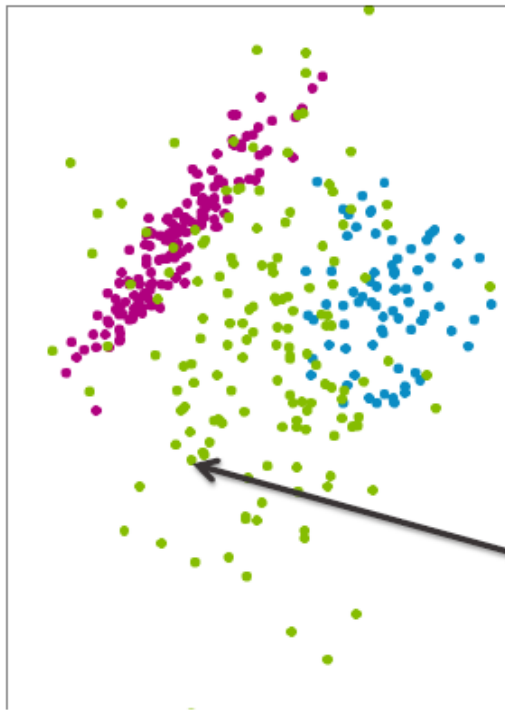
Desired soft assignments (responsibilities) are **easy** to compute when cluster parameters $\{\pi_k, \mu_k, \Sigma_k\}$ are known

But, we don't know these!

Imagine we knew the cluster
(hard) assignments z_i

78

Estimating cluster parameters



Imagine we know the
cluster assignments

Estimation problem
decouples across
clusters

Is **green** point informative of
fuchsia cluster parameters?

NO!

Imagine we knew the cluster
(hard) assignments z_i

79

Data table decoupling over clusters

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3
$x_4[1]$	$x_4[2]$	$x_4[3]$	1
$x_5[1]$	$x_5[2]$	$x_5[3]$	2
$x_6[1]$	$x_6[2]$	$x_6[3]$	2

Then split into separate tables and consider them independently.

Imagine we knew the cluster
(hard) assignments z_i

80

Maximum likelihood estimation

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3

Estimate $\{\pi_k, \mu_k, \Sigma_k\}$
given data assigned
to cluster k

maximum likelihood estimation
(MLE)

Find parameters that maximize the
score, or *likelihood*, of data

Imagine we knew the cluster
(hard) assignments z_i

81

Mean/covariance MLE

Sum these vectors

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3

divide by 3 (the total # of obs.)

denotes "estimate"

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i \text{ in } k} x_i$$

← average data points in cluster k
of obs. in cluster

$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_{i \text{ in } k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

Scalar case:

$$\hat{\sigma}_k^2 = \frac{1}{N_k} \sum_{i \text{ in } k} (x_i - \hat{\mu}_k)^2$$

Imagine we knew the cluster (hard) assignments z_i

82

Cluster proportion MLE

R	G	B	Cluster
$x_4[1]$	$x_4[2]$	$x_4[3]$	1

R	G	B	Cluster
$x_5[1]$	$x_5[2]$	$x_5[3]$	2
$x_6[1]$	$x_6[2]$	$x_6[3]$	2

R	G	B	Cluster
$x_1[1]$	$x_1[2]$	$x_1[3]$	3
$x_2[1]$	$x_2[2]$	$x_2[3]$	3
$x_3[1]$	$x_3[2]$	$x_3[3]$	3

obs in cluster k

$$\hat{\pi}_k = \frac{N_k}{N}$$

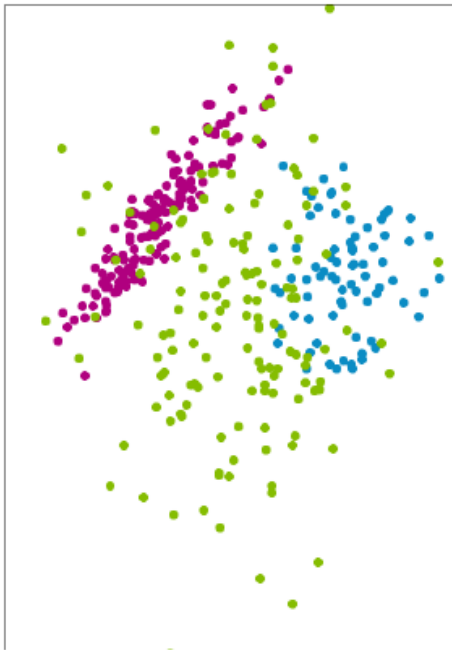
total # of obs

True for general mixtures of i.i.d. data,
not just Gaussian clusters

Imagine we knew the cluster
(hard) assignments z_i

83

Part 2a : Summary



needed to compute soft assignments



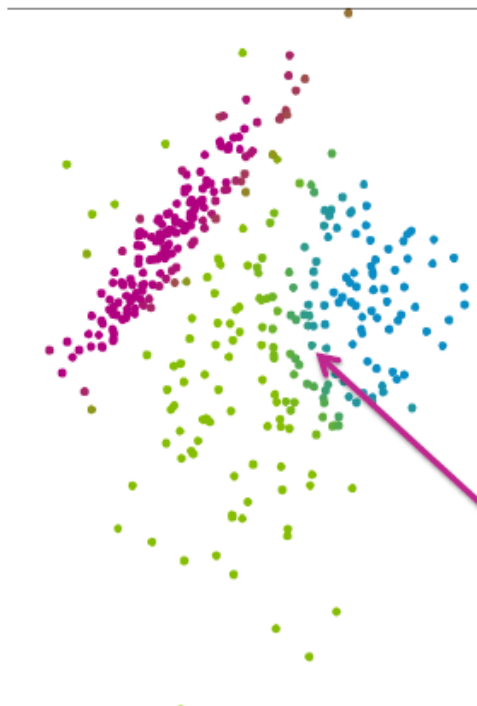
Cluster parameters are simple
to compute if we know the
cluster assignments

But, we don't know these!

What can we do with just soft assignments r_{ij} ?

84

Estimating cluster parameters from soft assignments



Instead of having a full observation \mathbf{x}_i in cluster k , just allocate a portion r_{ik}

\mathbf{x}_i divided across all clusters, as determined by r_{ik}

What can we do with just soft assignments r_{ij} ?

85

Maximum likelihood estimation from soft assignments

Just like in boosting with weighted observations...

R	G	B	r_{i1}	r_{i2}	r_{i3}
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30	0.18	0.52
$x_2[1]$	$x_2[2]$	$x_2[3]$	0.01	0.26	0.73
$x_3[1]$	$x_3[2]$	$x_3[3]$	0.002	0.008	0.99
$x_4[1]$	$x_4[2]$	$x_4[3]$	0.75	0.10	0.15
$x_5[1]$	$x_5[2]$	$x_5[3]$	0.05	0.93	0.02
$x_6[1]$	$x_6[2]$	$x_6[3]$	0.13	0.86	0.01

52% chance this obs is in cluster 3

Total weight in cluster:
(effective # of obs)

1.242

2.8

2.42

What can we do with just soft assignments r_{ij} ?

86

Maximum likelihood estimation from soft assignments

R	G	B	Cluster 1 weights		
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30		
$x_2[1]$	R	G	B	Cluster 2 weights	
$x_3[1]$					
$x_4[1]$	$x_1[1]$	$x_1[2]$	$x_1[3]$	0.18	
$x_5[1]$	$x_2[1]$	R	G	B	Cluster 3 weights
$x_6[1]$	$x_3[1]$				
	$x_4[1]$	$x_1[1]$	$x_1[2]$	$x_1[3]$	0.52
	$x_5[1]$	$x_2[1]$	$x_2[2]$	$x_2[3]$	0.73
	$x_6[1]$	$x_3[1]$	$x_3[2]$	$x_3[3]$	0.99
		$x_4[1]$	$x_4[2]$	$x_4[3]$	0.15
		$x_5[1]$	$x_5[2]$	$x_5[3]$	0.02
		$x_6[1]$	$x_6[2]$	$x_6[3]$	0.01

What can we do with just soft assignments r_{ij} ?

87

Cluster-specific location/shape MLE

R	G	B	Cluster 1 weights
$x_1[1]$	$x_1[2]$	$x_1[3]$	0.30
$x_2[1]$	$x_2[2]$	$x_2[3]$	0.01
$x_3[1]$	$x_3[2]$	$x_3[3]$	0.002
$x_4[1]$	$x_4[2]$	$x_4[3]$	0.75
$x_5[1]$	$x_5[2]$	$x_5[3]$	0.05
$x_6[1]$	$x_6[2]$	$x_6[3]$	0.13

$$\hat{\mu}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^N r_{ik} x_i$$

$$\hat{\Sigma}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^N r_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

$$N_k^{\text{soft}} = \sum_{i=1}^N r_{ik}$$

Total weight in cluster k
= effective # obs

Compute cluster parameter estimates with weights on each row operation

What can we do with just soft assignments r_{ij} ?

88

MLE of cluster proportions $\hat{\pi}_k$

r_{i1}	r_{i2}	r_{i3}
0.30	0.18	0.52
0.01	0.26	0.73
0.002	0.008	0.99
0.75	0.10	0.15
0.05	0.93	0.02
0.13	0.86	0.01

$$\hat{\pi}_k = \frac{N_k^{\text{soft}}}{N}$$

$$N_k^{\text{soft}} = \sum_{i=1}^N r_{ik}$$

Total weight in cluster k
= effective # obs

Total weight
in cluster:

1.242	2.8	2.42
-------	-----	------

Total weight
in dataset:

6

datapoints N

Estimate cluster proportions from relative weights

What can we do with just soft assignments r_{ij} ?

89

Defaults to hard assignment case when r_{ij} in $\{0,1\}$

Hard assignments have:

$$r_{ik} = \begin{cases} 1 & i \text{ in } k \\ 0 & \text{otherwise} \end{cases}$$

R	G	B	r_{i1}	r_{i2}	r_{i3}
$x_1[1]$	$x_1[2]$	$x_1[3]$	0	0	1
$x_2[1]$	$x_2[2]$	$x_2[3]$	0	0	1
$x_3[1]$	$x_3[2]$	$x_3[3]$	0	0	1
$x_4[1]$	$x_4[2]$	$x_4[3]$	1	0	0
$x_5[1]$	$x_5[2]$	$x_5[3]$	0	1	0
$x_6[1]$	$x_6[2]$	$x_6[3]$	0	1	0

One-hot encoding of cluster assignment

Total weight in cluster:

1	2	3
---	---	---

What can we do with just soft assignments r_{ij} ?

90

Equating the estimates...

$$\hat{\pi}_k = \frac{N_k^{\text{soft}}}{N}$$

$N_k^{\text{soft}} = \sum_{i=1}^N r_{ik}$ if $\{0,1\}$ just count obs i in cluster k if $r_{ik}=1 = N_k$ ✓

$$\hat{\mu}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^N r_{ik} x_i$$

only add x_i if i in k ($r_{ik}=1$) $\rightarrow \frac{1}{N_k} \sum_{i \text{ in } k} x_i$ ✓

$$\hat{\Sigma}_k = \frac{1}{N_k^{\text{soft}}} \sum_{i=1}^N r_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

same as above $\rightarrow \frac{1}{N_k} \sum_{i \text{ in } k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$ ✓

What can we do with just soft assignments r_{ij} ?

91

Part 2b: Summary



Still straightforward to compute cluster parameter estimates from soft assignments

Expectation maximization (ME)

92

An iterative algorithm

Motivates an iterative algorithm:

1. **E-step:** estimate cluster responsibilities given current parameter estimates

$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i | \hat{\mu}_k, \hat{\Sigma}_k)}{\sum_{j=1}^K \hat{\pi}_j N(x_i | \hat{\mu}_j, \hat{\Sigma}_j)}$$

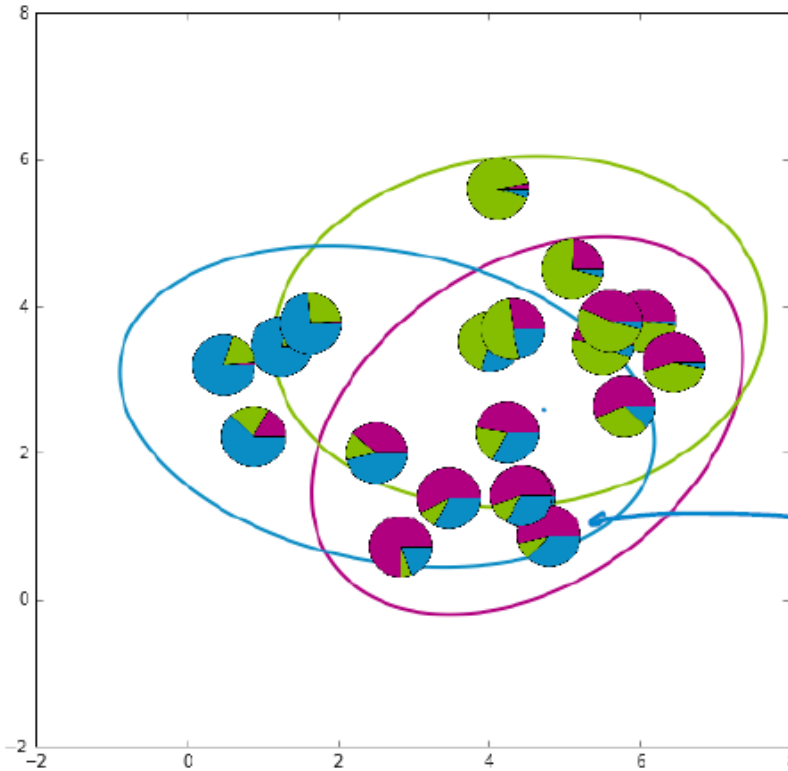
2. **M-step:** maximize likelihood over parameters given current responsibilities

$$\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}_k | \{\hat{r}_{ik}, x_i\}$$

Expectation maximization (EM)

93

EM for mixtures of Gaussians in pictures – initialization



Initialize
iter counter
 $\{\pi_k^{(0)}, \mu_k^{(0)}, \Sigma_k^{(0)}\}$

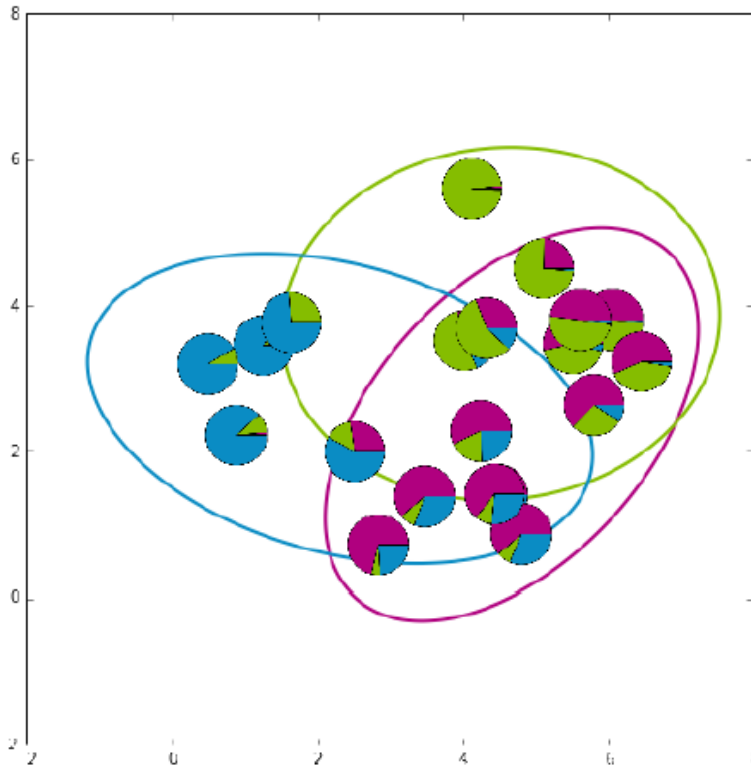
Then compute
 $\hat{r}_{ik}^{(1)}$

$$\hat{r}_i^{(1)} = \begin{matrix} \text{fuchsia} & \text{blue} & \text{green} \\ [0.52 & 0.4 & 0.08] \end{matrix}$$

Expectation maximization (EM)

94

EM for mixtures of Gaussians
in pictures – after 1st iteration



Maximize likelihood
given soft assign. $r_{ik}^{(1)}$

$$\rightarrow \{ \hat{\pi}_k^{(1)}, \hat{\mu}_k^{(1)}, \hat{\Sigma}_k^{(1)} \}$$

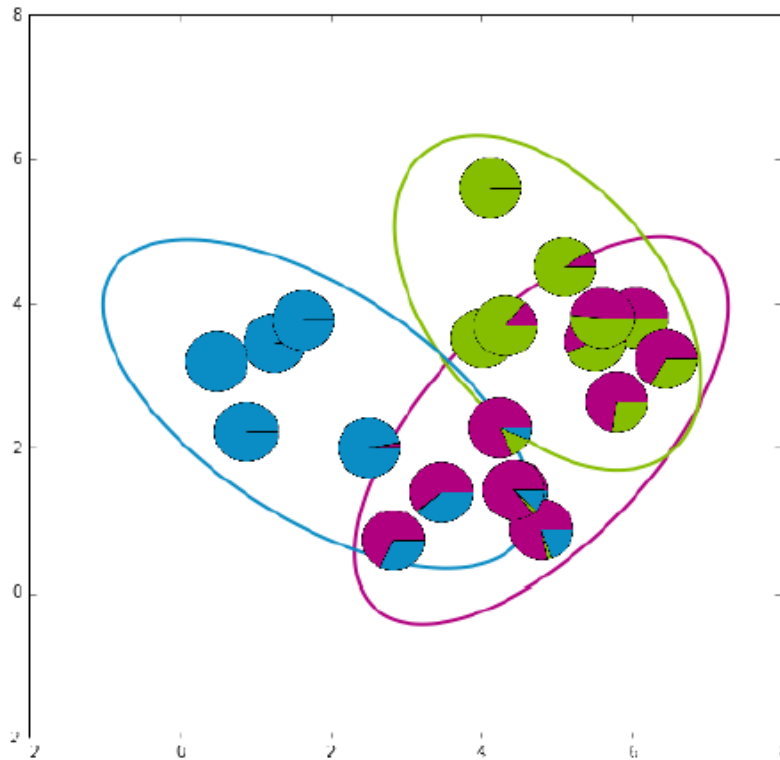
Then recompute responsibilities

$$\hat{r}_{ik}^{(2)}$$

Expectation maximization (EM)

95

EM for mixtures of Gaussians
in pictures – after 2nd iteration

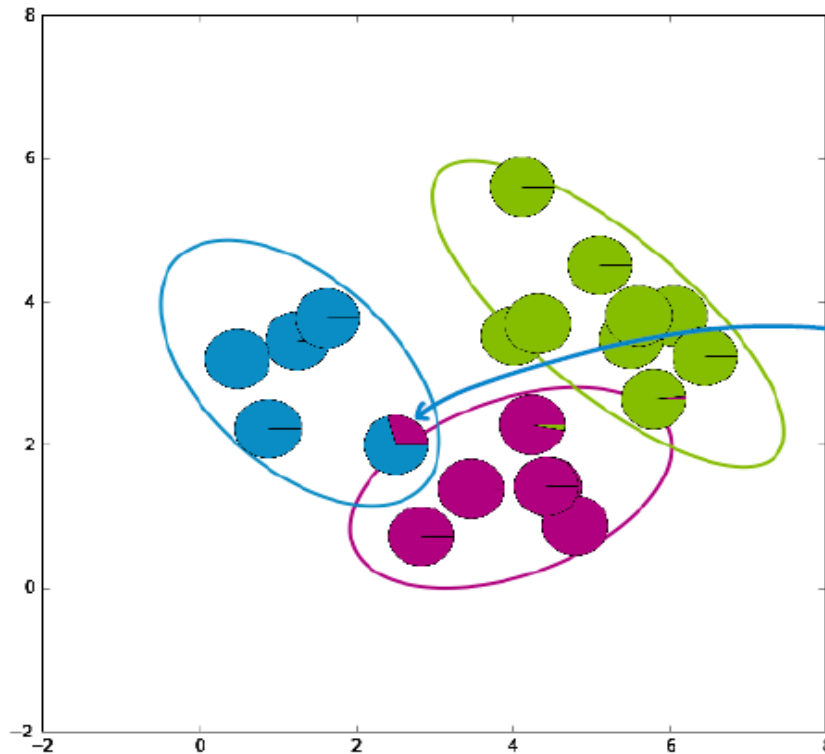


*rinse
+
repeat
until convergence*

Expectation maximization (EM)

96

EM for mixtures of Gaussians in pictures – converged solution

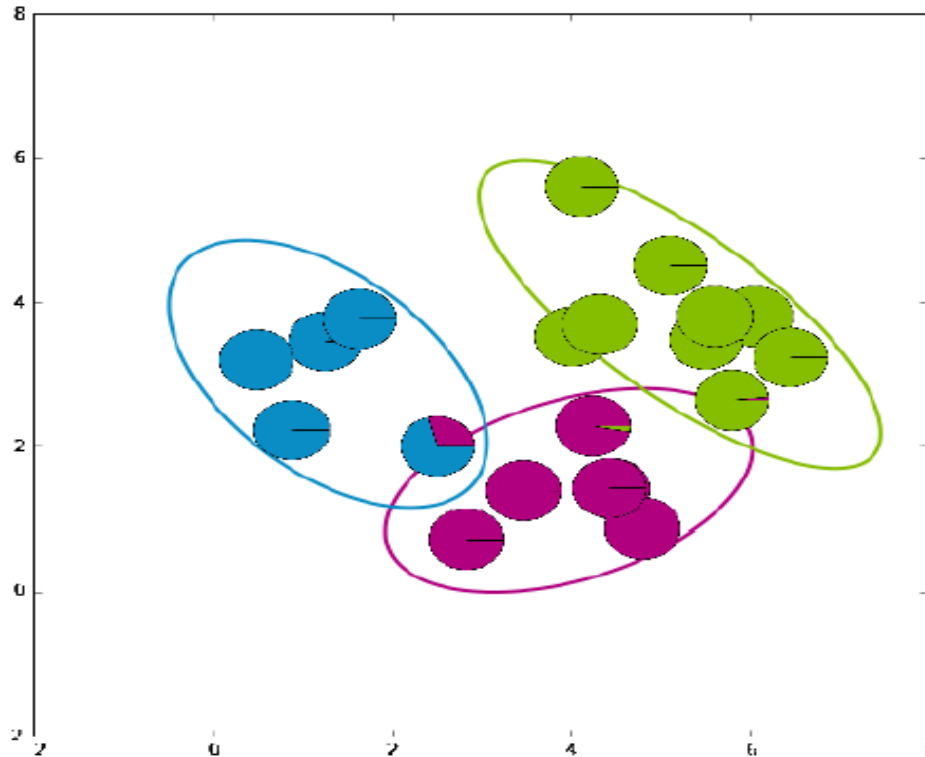


Clearly see
uncertainty in
assignment of obs.
to blue or fuchsia
cluster, even in
final assignments.

Expectation maximization (EM)

97

EM for mixtures of Gaussians
in pictures - [replay](#)



Expectation maximization (ME)

98

Convergence of EM

- EM is a **coordinate-ascent algorithm**
 - Can equate E-and M-steps with alternating maximizations of an objective function
- Converges to a **local mode**
- We will assess via (log) likelihood of data under current parameter and responsibility estimates

Expectation maximization (ME)

99

Initialization

- Many ways to initialize the EM algorithm
- Important for convergence rates and quality of local mode found
- Examples:
 - Choose K observations at random to define K "centroids". Assign other observations to nearest centroid to form initial parameter estimates.
 - Pick centers sequentially to provide good coverage of data like in k -means++
 - Initialize from k -means solution
 - Grow mixture model by splitting (and sometimes removing) clusters until K clusters are formed

Expectation maximization (ME)

100

Overfitting of MLE

Maximizing likelihood can **overfit to data**

Imagine at $K=2$ example with one obs assigned to **cluster 1** and others assigned to **cluster 2**

- **What parameter values maximize likelihood?**



Set center equal to point and shrink variance to 0

Likelihood goes to ∞ !

Expectation maximization (ME)

101

Overfitting in high dims

Doc-clustering example:

Imagine only 1 doc assigned to cluster k has word w
(or all docs in cluster agree on count of word w)

Likelihood maximized by setting $\mu_k[w] = \mathbf{x}_i[w]$ and $\sigma_{w,k}^2 = 0$

Likelihood of any doc with different count on
word w being in cluster k is 0!

Expectation maximization (ME)

102

Simple regularization of M-step for mixtures of Gaussians

Simple fix: Don't let variances $\rightarrow 0$!

Add small amount to diagonal of covariance estimate

Alternatively, take Bayesian approach and place prior on parameters.

Similar idea, but all parameter estimates are "smoothed" via cluster pseudo-observations.

Expectation maximization (ME)

103

Relationship to k-means

Consider Gaussian mixture model with

$$\Sigma = \begin{pmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \sigma^2 & \\ & & & \ddots \end{pmatrix}$$

Spherically symmetric clusters



and let the variance parameter $\sigma \rightarrow 0$

Datapoint gets fully assigned to nearest center, just as in k-means

- Spherical clusters with equal variances, so relative likelihoods just function of distance to cluster center
- As variances $\rightarrow 0$, likelihood ratio becomes 0 or 1
- Responsibilities weigh in cluster proportions, but dominated by likelihood disparity

$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i | \hat{\mu}_k, \sigma^2 I)}{\sum_{j=1}^K \hat{\pi}_j N(x_i | \hat{\mu}_j, \sigma^2 I)}$$

Expectation maximization (ME)

104

Infinitesimally small variance EM = k-means

1. **E-step:** estimate cluster responsibilities given current parameter estimates

$$\hat{r}_{ik} = \frac{\hat{\pi}_k N(x_i | \hat{\mu}_k, \sigma^2 I)}{\sum_{j=1}^K \hat{\pi}_j N(x_i | \hat{\mu}_j, \sigma^2 I)} \in \{0, 1\}$$

Infinitesimally small

Decision based on
distance to nearest
cluster center

2. **M-step:** maximize likelihood over parameters given current responsibilities (**hard assignments!**)

$$\hat{\pi}_k, \hat{\mu}_k \mid \{\hat{r}_{ik}, x_i\}$$

Mixed membership models for documents

Clustering model

106

So far, clustered articles into groups



Clustering goal: discover groups of related docs

Clustering model

107

Are documents about just one thing?



Is this article
just about
science?



Clustering model

108

Soft assignments capture uncertainty



Soft assignment r_{ik} tells us this doc could be about world news or science

But, clustering model still specifies each doc belongs to **1 topic**

Soft assignments

109

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^c, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment of Statistics, University of Washington, Seattle, WA

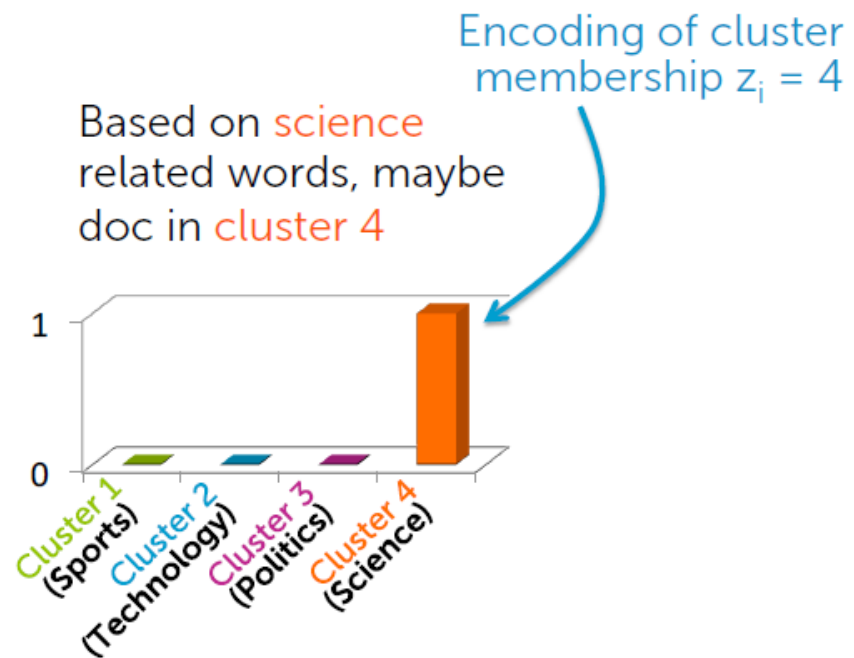
Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Soft assignments

110

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^c, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown seizures. We demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric EEG, factorial hidden Markov model, graphical model, time series

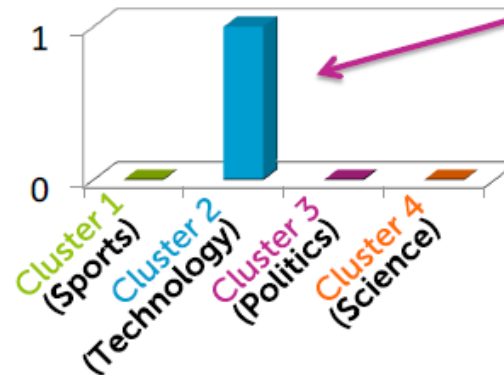
1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

Soft assignments capture uncertainty in $z_i = 2$ or 4

Encoding of cluster membership $z_i = 2$

Or maybe cluster 2 (technology) is a better fit



Soft assignments

111

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^c, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizure events.

into EEG (iEEG) data can switch between states. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

“ z_i ” is both 2 and 4

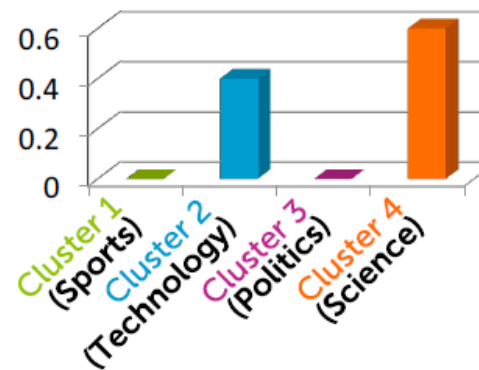
graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

Really, it's about science and technology



Mixed membership models

112

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^c, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

Mixed membership models

Want to discover a **set** of memberships

(In contrast, cluster models aim at discovering a single membership)

Building alternative model

113

An alternative document clustering model



(Back to clustering, not mixed membership modeling)

Building an alternative model

114

So far, we have considered...

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^a, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment of Statistics, University of Washington, Seattle, WA

Abstract

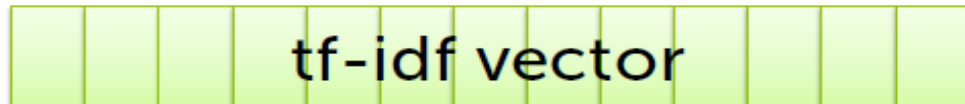
Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

$x_i =$



Building an alternative model

115

Bag-of-words representation

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^c, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the complex dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous transitions between (i) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help facilitate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible



Building an alternative model

116

Bag-of-words representation

Modeling the Complex Dynamics and Changing
Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^c, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

$\mathbf{X}_i = \{ \text{modeling, complex, epilepsy, modeling, Bayesian, clinical, epilepsy, EEG, data, dynamic...} \}$

multiset

= unordered set of words with
duplication of unique elements
mattering

Model for „bag-of-words”

117

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^c, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

A model for bag-of-words representation

As before, the “prior” probability that **doc** i is from **topic** k is:

$$p(z_i = k) = \pi_k$$

$\boldsymbol{\pi} = [\pi_1 \ \pi_2 \ \dots \ \pi_k]$
represents **corpus-wide**
topic prevalence

Model for „bag-of-words”

118

Modeling the Complex Dynamics and Changing Correlations of Epileptic Events

Drausin F. Wulsin^a, Emily B. Fox^c, Brian Litt^{a,b}

^aDepartment of Bioengineering, University of Pennsylvania, Philadelphia, PA

^bDepartment of Neurology, University of Pennsylvania, Philadelphia, PA

^cDepartment of Statistics, University of Washington, Seattle, WA

Abstract

Patients with epilepsy can manifest short, sub-clinical epileptic “bursts” in addition to full-blown clinical seizures. We believe the relationship between these two classes of events—something not previously studied quantitatively—could yield important insights into the nature and intrinsic dynamics of seizures. A goal of our work is to parse these complex epileptic events into distinct dynamic regimes. A challenge posed by the intracranial EEG (iEEG) data we study is the fact that the number and placement of electrodes can vary between patients. We develop a Bayesian nonparametric Markov switching process that allows for (i) shared dynamic regimes between a variable number of channels, (ii) asynchronous regime-switching, and (iii) an unknown dictionary of dynamic regimes. We encode a sparse and changing set of dependencies between the channels using a Markov-switching Gaussian graphical model for the innovations process driving the channel dynamics and demonstrate the importance of this model in parsing and out-of-sample predictions of iEEG data. We show that our model produces intuitive state assignments that can help automate clinical analysis of seizures and enable the comparison of sub-clinical bursts and full clinical seizures.

Keywords: Bayesian nonparametric, EEG, factorial hidden Markov model, graphical model, time series

1. Introduction

Despite over three decades of research, we still have very little idea of what defines a seizure. This ignorance stems both from the complexity of epilepsy as a disease and a paucity of quantitative tools that are flexible

A model for bag-of-words representation

Assuming **doc i** is from **topic k**, words occur with probabilities:

SCIENCE	
patients	0.05
clinical	0.01
epilepsy	0.002
seizures	0.0015
EEG	0.001
...	...

} words in vocab

Model for „bag-of-words”

119

Topic-specific word probabilities

Distribution on words in vocab for **each topic**

SCIENCE		TECH		SPORTS		
experiment	0.1	develop	0.18	player	0.15	
test	0.08	computer	0.09	score	0.07	
discover	0.05	processor	0.032	team	0.06	...
hypothesize	0.03	user	0.027	goal	0.03	
climate	0.01	internet	0.02	injury	0.01	
...	

(table now organized by decreasing probabilities
showing top words in each category)

Model for „bag-of-words”

120

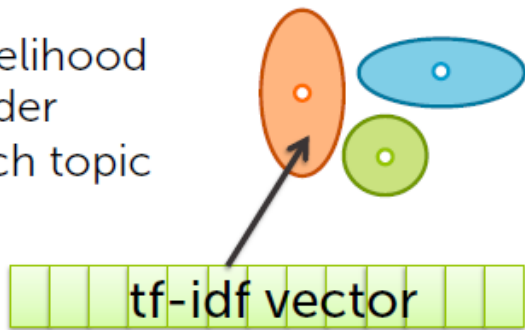
Comparing and contrasting

Previously

Prior topic probabilities

$$p(z_i = k) = \pi_k$$

Likelihood under each topic



compute likelihood of **tf-idf** vector under each **Gaussian**

Now

$$p(z_i = k) = \pi_k$$

	SCIENCE	TECH	SPORTS	
experiment	0.1	develop 0.18	player 0.15	
test	0.08	computer 0.09	score 0.07	
discover	0.05	processor 0.032	team 0.06	...
hypothesize	0.03	user 0.027	goal 0.03	
climate	0.01	internet 0.02	injury 0.01	
...

{modeling, complex, epilepsy, modeling, Bayesian, clinical, epilepsy, EEG, data, dynamic...}

compute likelihood of the **collection of words** in doc under each **topic distribution**

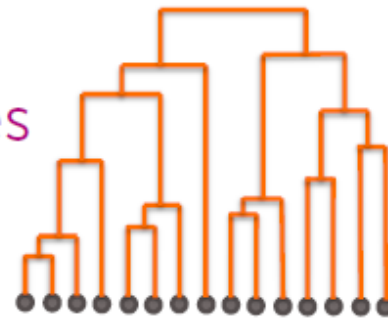
Hierarchical clustering

Why hierarchical clustering

122

- Avoid choosing # clusters beforehand

- **Dendrograms** help visualize different clustering **granularities**
 - No need to rerun algorithm



- Most algorithms allow user to **choose any distance metric**
 - k-means restricted us to Euclidean distance

Why hierarchical clustering

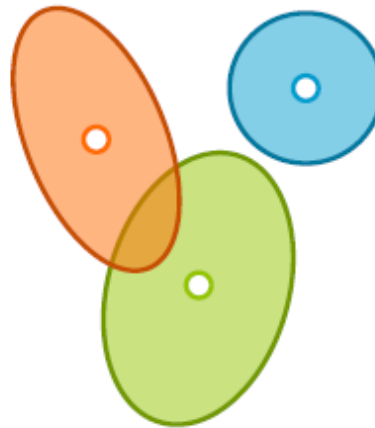
123

Can often find more **complex shapes** than k-means or Gaussian mixture models

k-means: spherical clusters



Gaussian mixtures: ellipsoids

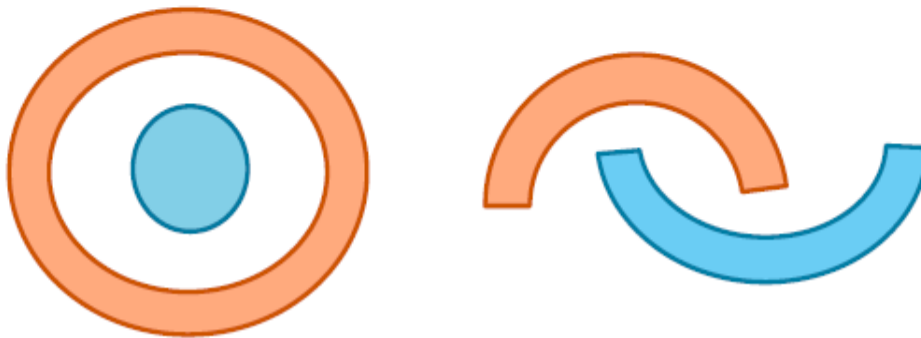


Why hierarchical clustering

124

Can often find more **complex shapes** than k-means or Gaussian mixture models

What about these?



Two main types of algorithms

125

Divisive, *a.k.a. top-down*: Start with all data in one big cluster and recursively split.

- Example: **recursive k-means**

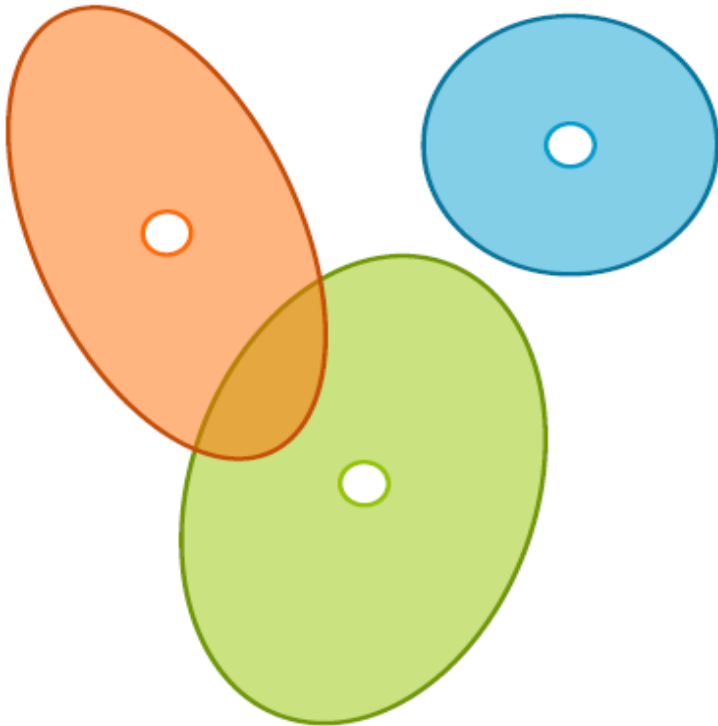
Agglomerative *a.k.a. bottom-up*: Start with each data point as its own cluster. Merge clusters until all points are in one big cluster.

- Example: **single linkage**

Divisive clustering

126

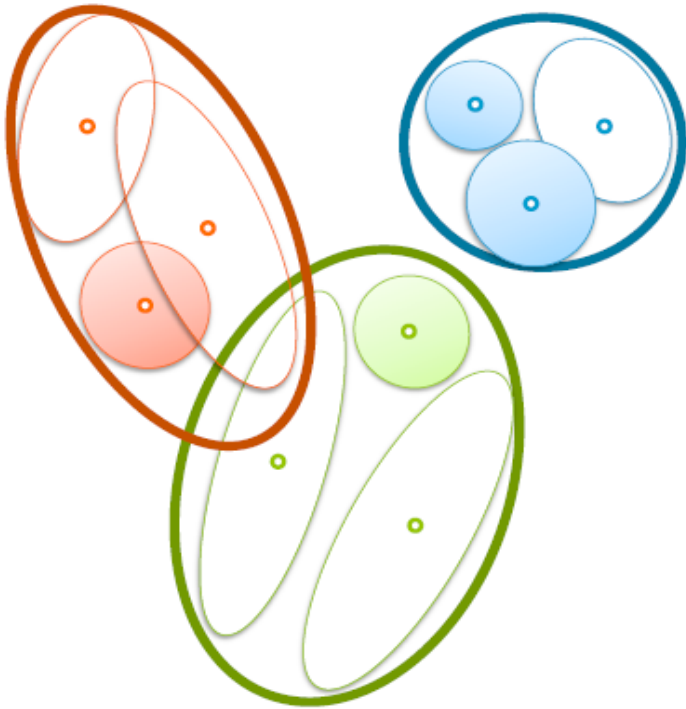
Divisive in pictures – level 1



Divisive clustering

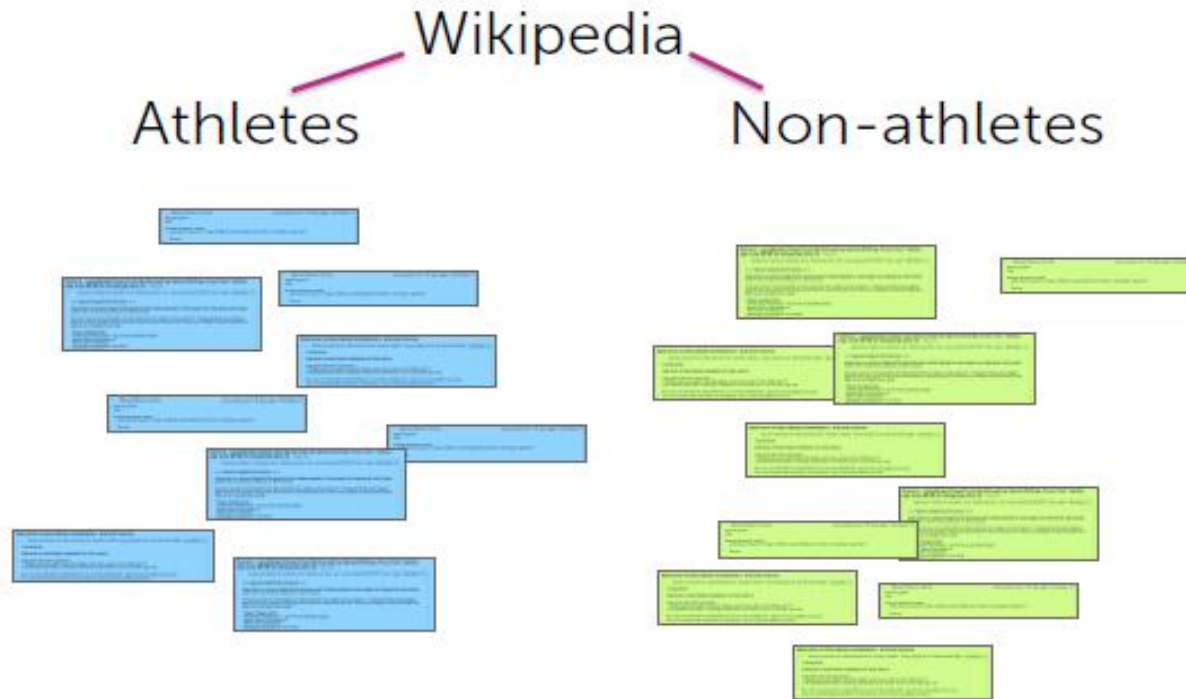
127

Divisive in pictures – level 2



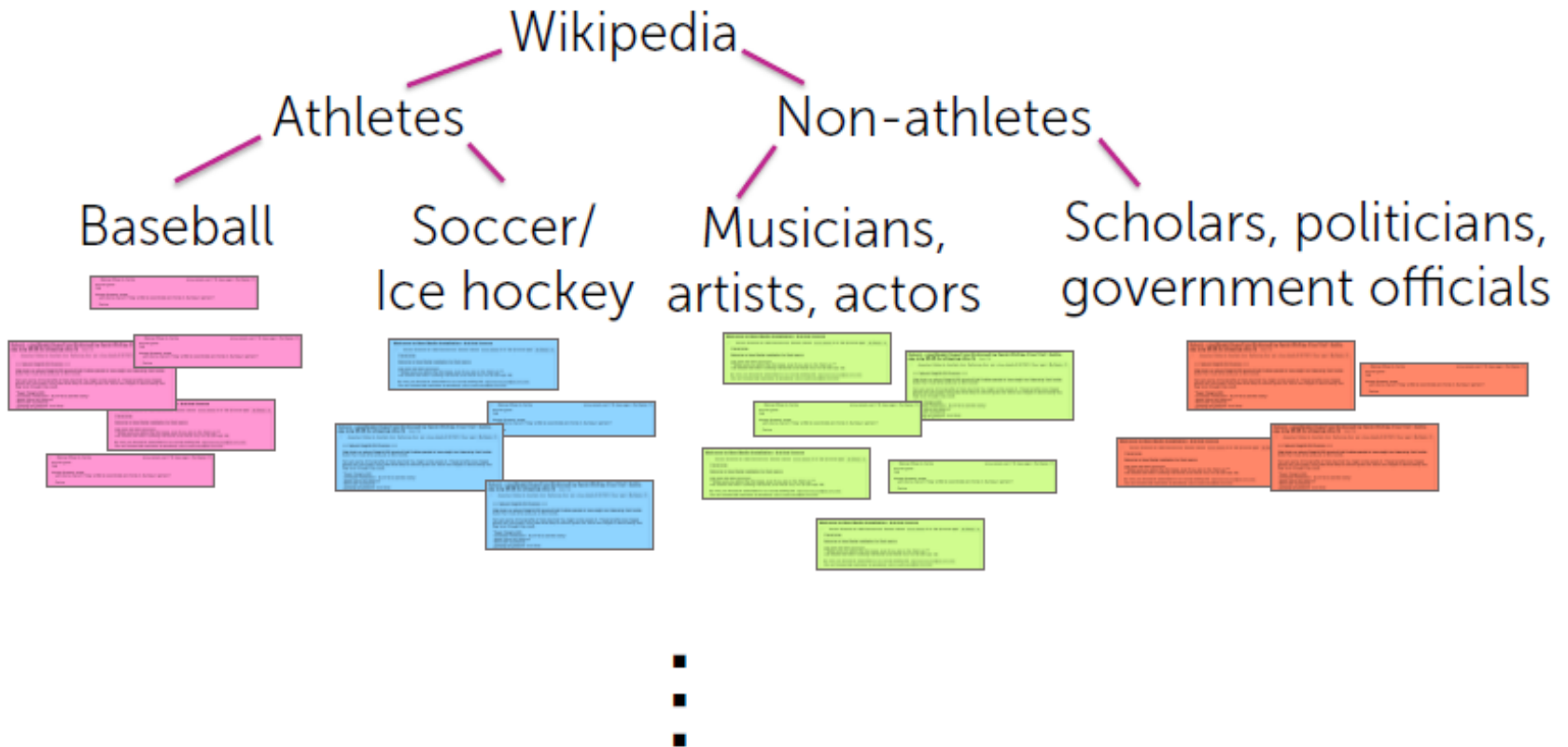
Divisive: Recursive k-means

128



Divisive: Recursive k-means

129



Divisive: choices to be made

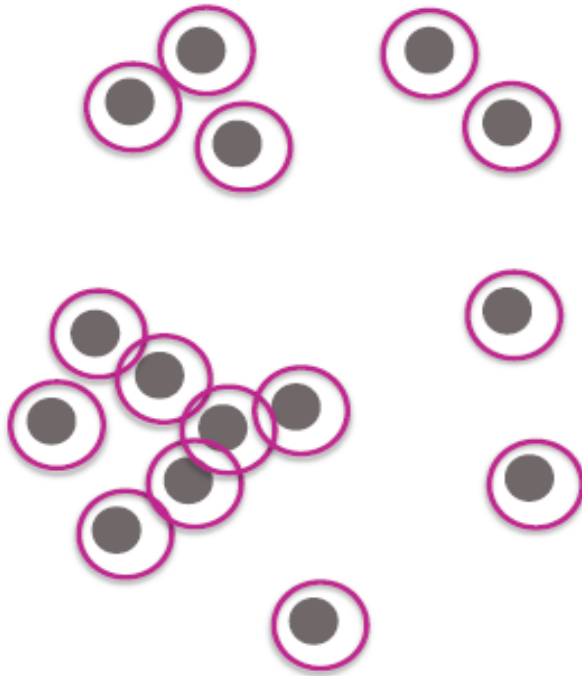
130

- Which algorithm to recurse
- How many clusters per split
- When to split vs. stop
 - **Max cluster size:**
number of points in cluster falls below threshold
 - **Max cluster radius:**
distance to furthest point falls below threshold
 - **Specified # clusters:**
split until pre-specified # clusters is reached

Aglomerative: Single linkage

131

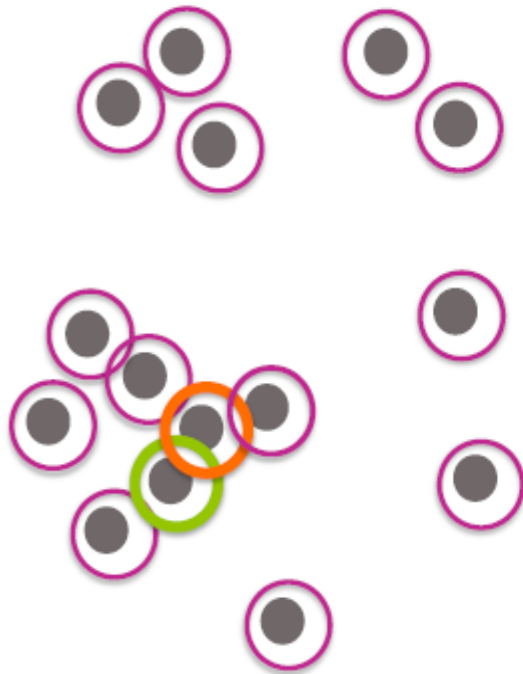
1. Initialize each point to be its own cluster



Aglomerative: Single linkage

132

2. Define distance between clusters to be:



$$\text{distance}(C_1, C_2) =$$

$$\min_{\substack{x_i \in C_1, \\ x_j \in C_2}} d(x_i, x_j)$$

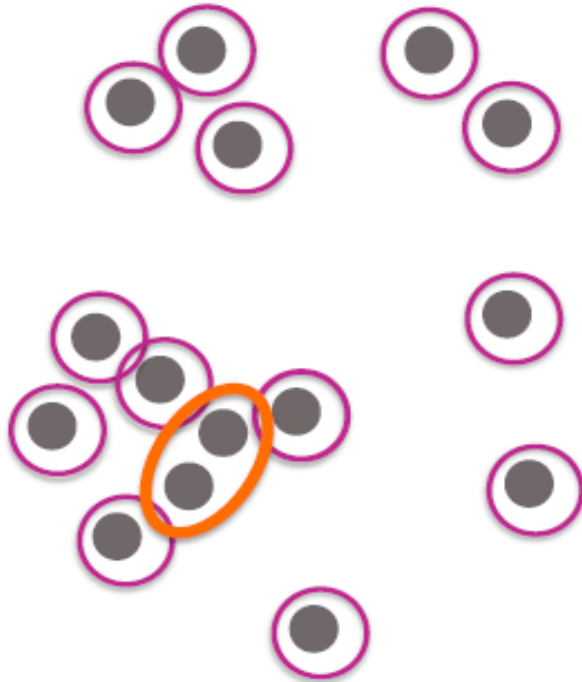
specified pairwise
distance function

Linkage criteria

Aglomerative: Single linkage

133

3. Merge the two closest clusters



Aglomerative: Single linkage

134

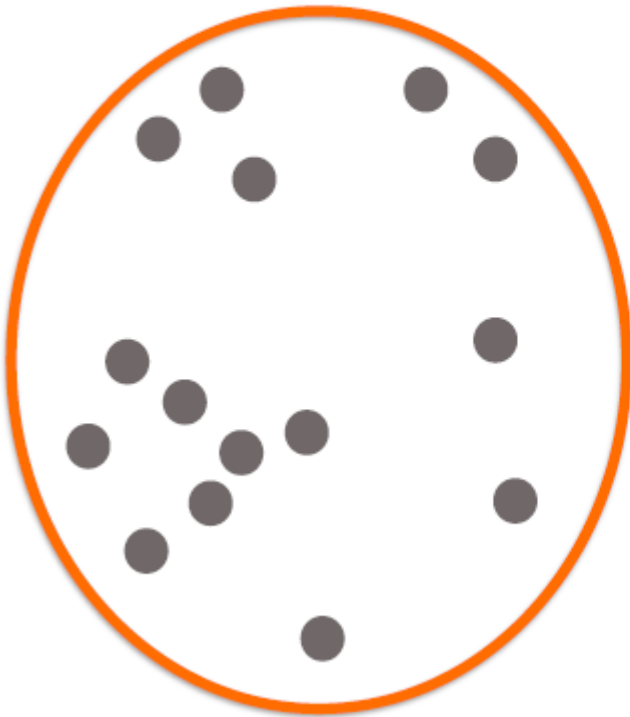
4. Repeat step 3 until all points are in one cluster



Aglomerative: Single linkage

135

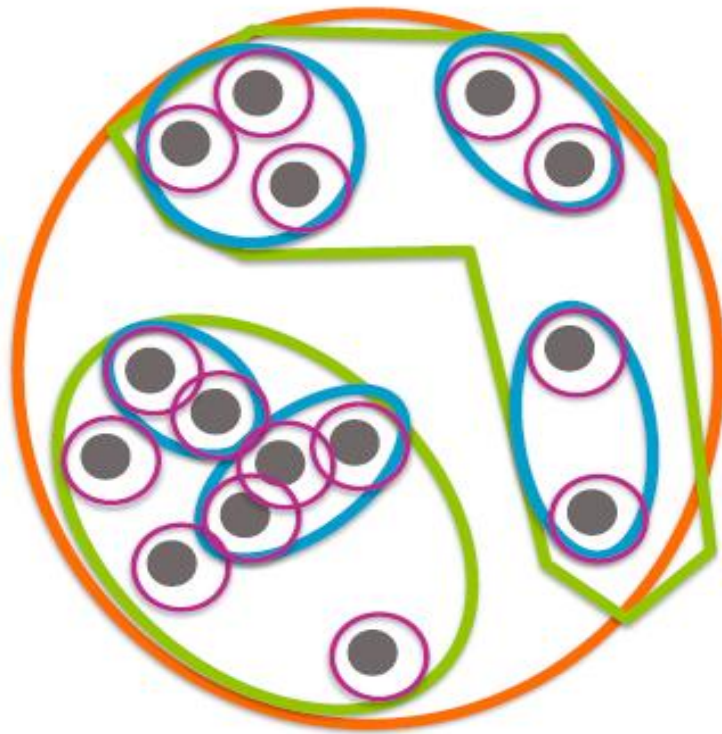
4. Repeat step 3 until all points are in one cluster



Cluster of clusters

136

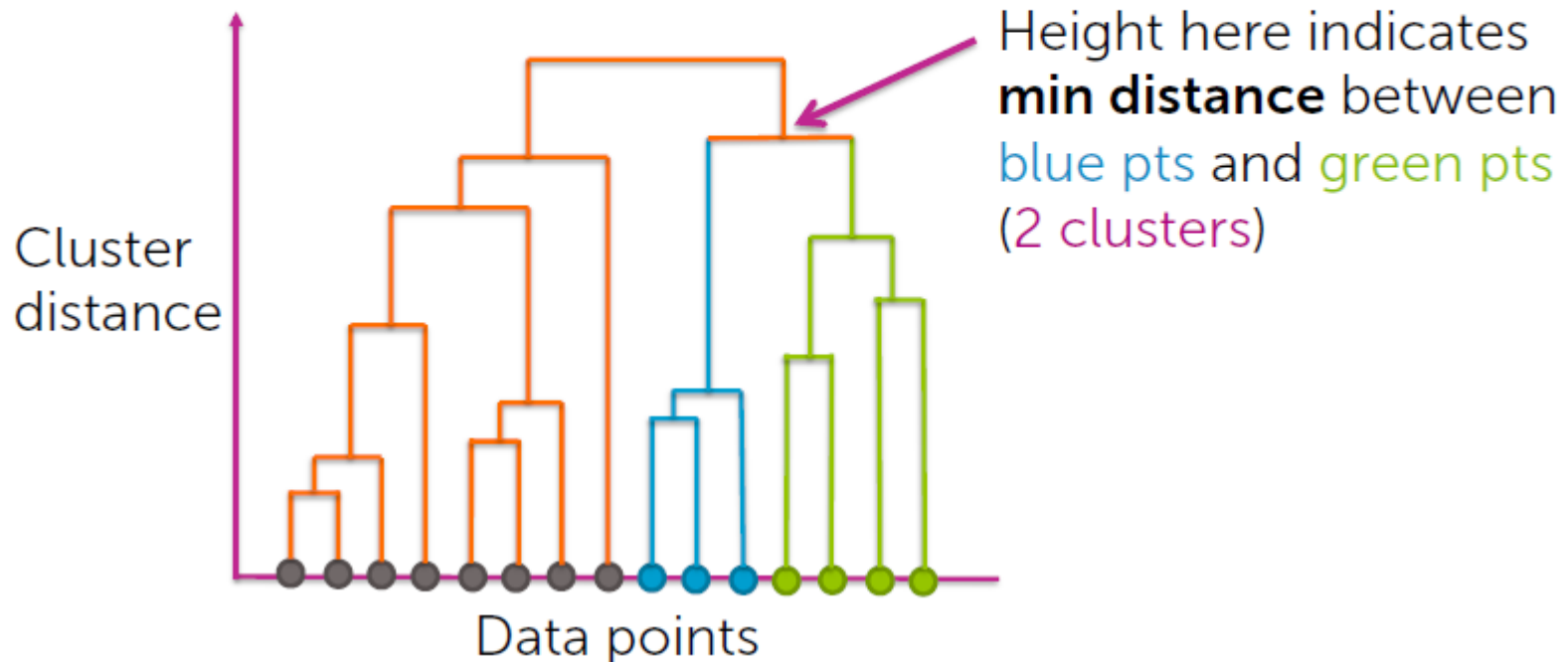
Just like our picture for divisive clustering...



The dendrogram

137

- x axis shows data points (carefully ordered)
- y-axis shows distance between pair of clusters

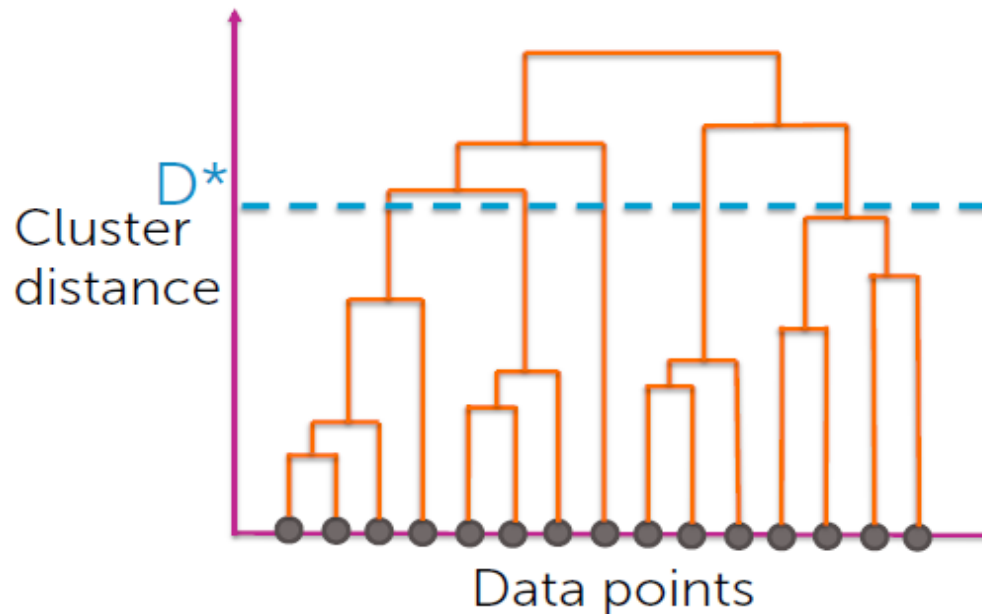


Extracting a partition

138

Choose a distance D^* at which to cut dendrogram

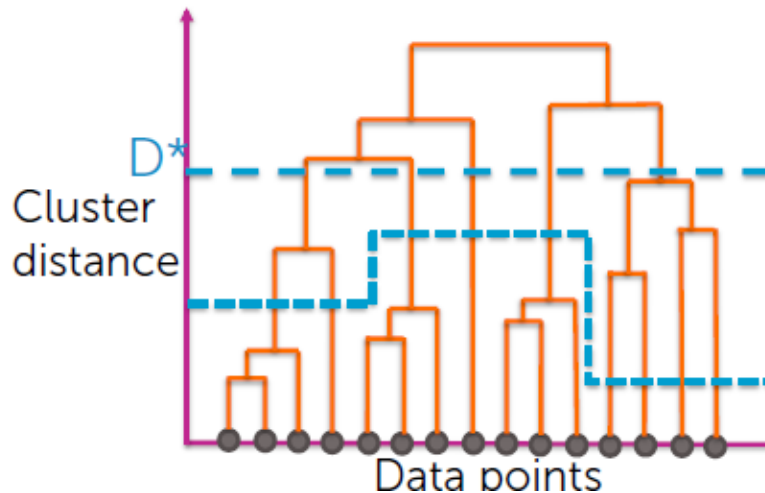
Every branch that crosses D^* becomes a separate cluster



Agglomerative: choices to be made

139

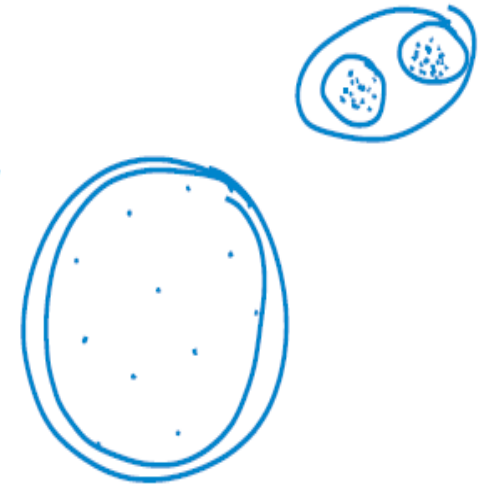
- Distance metric: $d(\mathbf{x}_i, \mathbf{x}_j)$
- Linkage function: e.g., $\min_{\substack{\mathbf{x}_i \in C_1, \\ \mathbf{x}_j \in C_2}} d(\mathbf{x}_i, \mathbf{x}_j)$
- Where and how to cut dendrogram



More on cutting dendrogram


140

- For visualization, smaller # clusters is preferable
- For tasks like outlier detection, cut based on:
 - Distance threshold
 - Inconsistency coefficient
 - Compare height of merge to average merge heights below
 - If top merge is substantially higher, then it is joining two subsets that are relatively far apart compared to the members of each subset internally
 - Still have to **choose a threshold** to cut at, but now in terms of "inconsistency" rather than distance
- No cutting method is "incorrect", some are just more useful than others



Computational considerations

141

- Computing all pairs of distances is **expensive**
 - Brute force algorithm is $O(N^2 \log(N))$
 -  # datapoints
- Smart implementations use triangle inequality to **rule out candidate pairs**
- Best known algorithm is $O(N^2)$