Clustering

CS 229: Machine Learning
Emily Fox
Stanford University
February 28, 2024

©2024 Emily Fox

1

Motivating clustering approaches

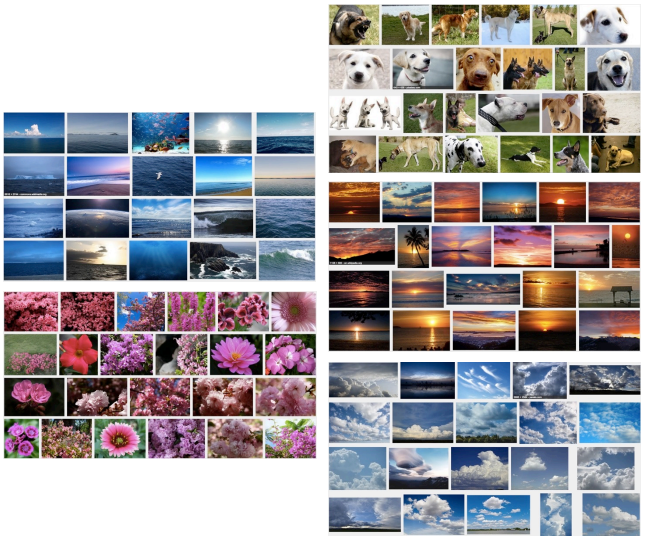
©2024 Emily Fox

2

Clustering images

Discover groups of similar images:

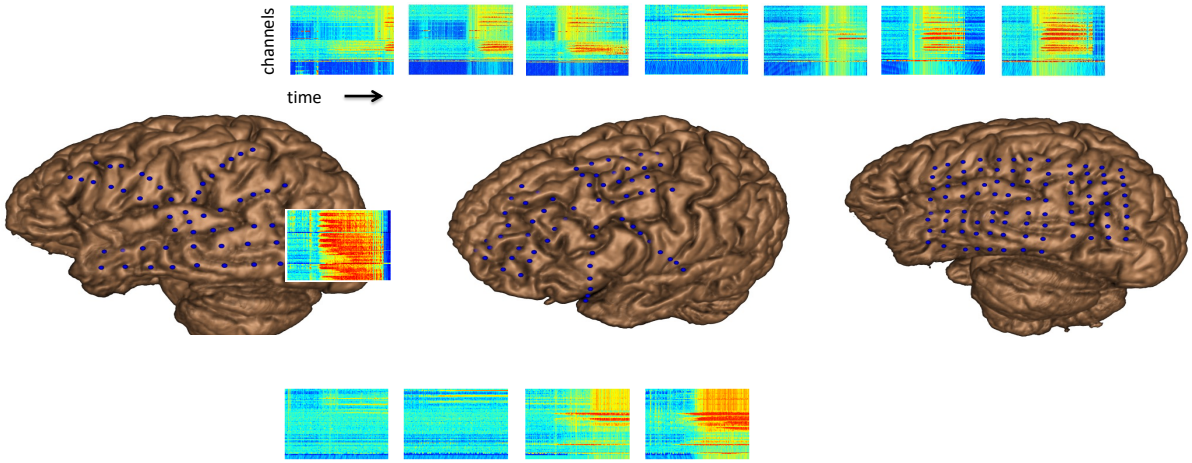
- Ocean
- Pink flower
- Dog
- Sunset
- Clouds
- …



©2024 Emily Fox

3

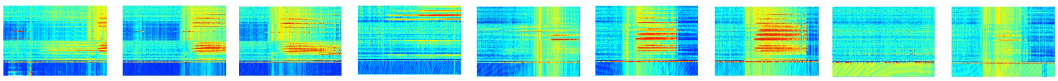
Characterizing patients and seizure types



©2024 Emily Fox

4

Characterizing patients and seizure types



©2024 Emily Fox

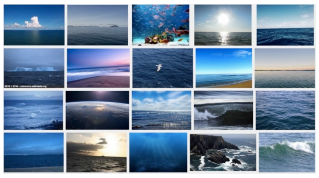
5

Clustering: An unsupervised learning task

©2024 Emily Fox

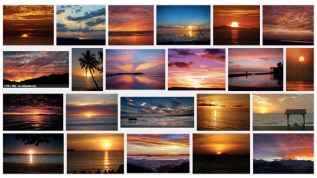
6

What if labels are known for training data?



Ocean

Dog



Sunset

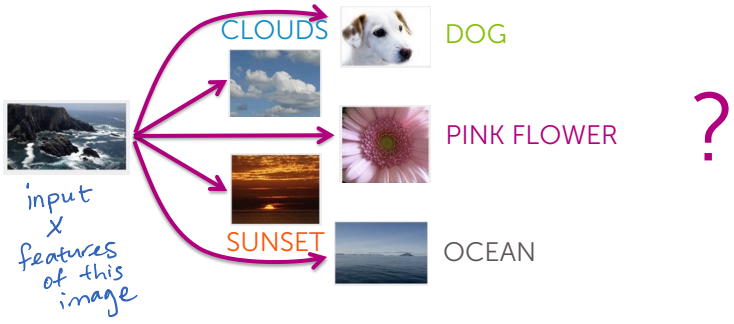
Pink flower

Clouds

©2024 Emily Fox

7

Multiclass classification problem



Example of **supervised learning**

©2024 Emily Fox

8

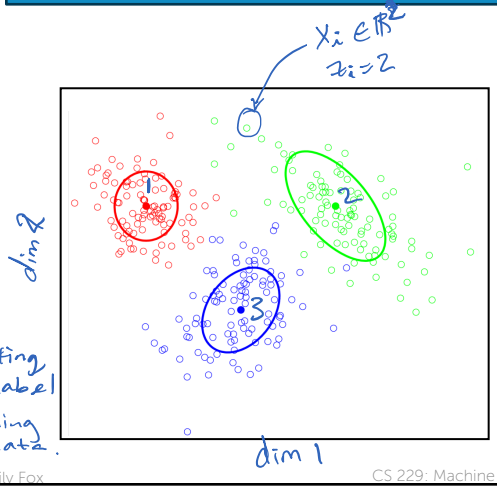
Clustering

No labels provided
…uncover cluster structure from input alone

Input: image features $\mathbf{x}_i$
Output: cluster labels z_i

If supervised setting have output label y_i in training data.

An **unsupervised learning** task



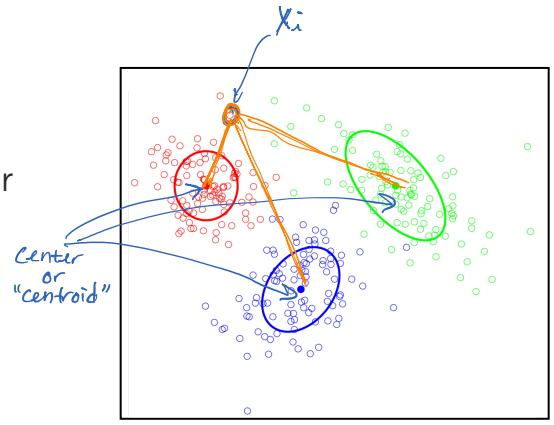
©2024 Emily Fox

9

What defines a cluster?

Cluster defined by center & shape/spread

Assign observation $\mathbf{x}_i$ to cluster k if
- Score under cluster k is higher than under others
- For simplicity, often define score as distance to cluster center (ignoring shape)



©2024 Emily Fox

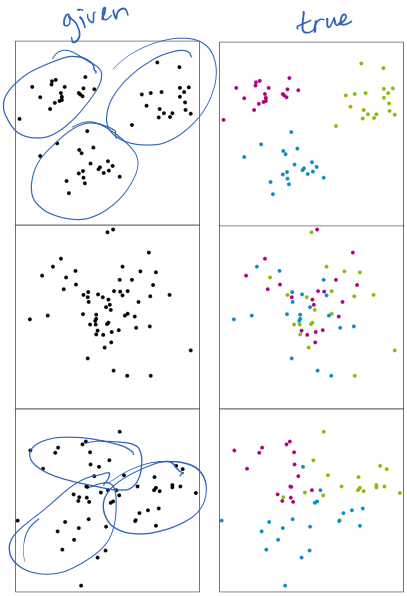
10

Hope for unsupervised learning

Easy

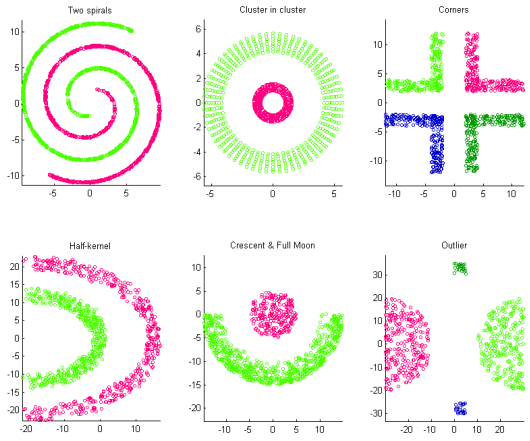
Impossible

In between



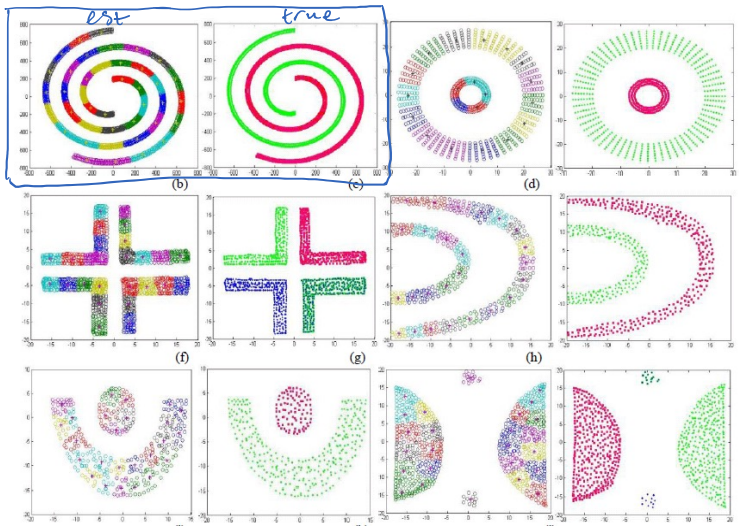
©2024 Emily Fox

Other (challenging!) clusters to discover…



©2024 Emily Fox

Other (challenging!) clusters to discover…



©2024 Emily Fox

13

k-means: A clustering algorithm

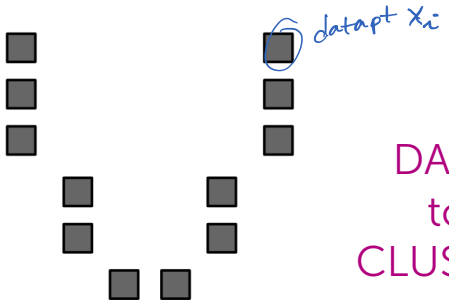
©2024 Emily Fox

14

k-means

Assume

- Score= distance to cluster center (smaller better)



DATA to CLUSTER

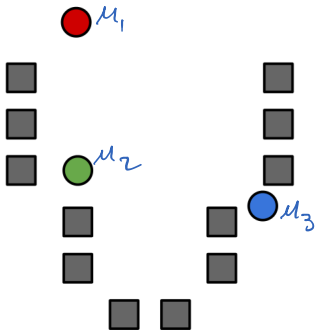
©2024 Emily Fox

15

k-means algorithm

0. Initialize cluster centers

$$\mu_1, \mu_2, \ldots, \mu_k$$



(K=3 clusters)

©2024 Emily Fox

16

k-means algorithm

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

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

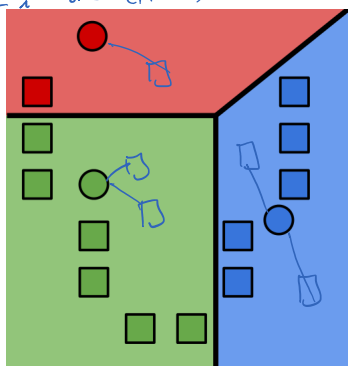
Inferred label for obs i, whereas supervised learning has given label y_i

jth cluster center (fixed for now)

ith obs (fixed)

dist. b/t x_i and μ_j

Select the cluster index closest to x_i



Voronoi tesselation

©2024 Emily Fox

17

k-means algorithm

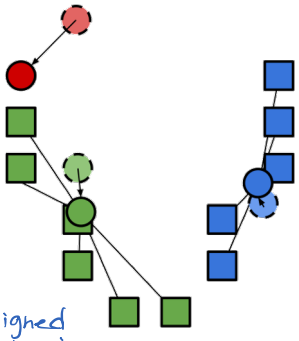
0. Initialize cluster centers
1. Assign observations to closest cluster center
2. Revise cluster centers as mean of assigned observations

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

center of mass per cluster

total # of obs in cluster j

all obs. assigned to cluster j

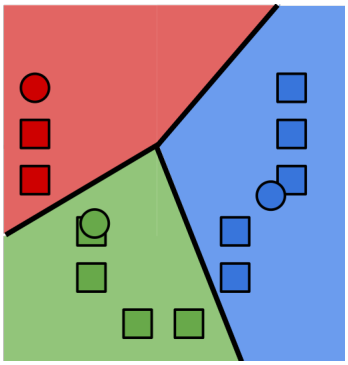


©2024 Emily Fox

18

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



©2024 Emily Fox

19

k-means as coordinate descent

©2024 Emily Fox

20

Recall: Coordinate descent

Goal: Minimize some function g

$\min_{\mathbf{w}} g(\mathbf{w})$

$g(\mathbf{w}) = g(w_0, w_1, \ldots, w_D)$

when keeping others fixed

Often, hard to find minimum for all coordinates, but easy for each coordinate

Coordinate descent:

Initialize $\hat{\mathbf{w}} = 0$ (or smartly…)

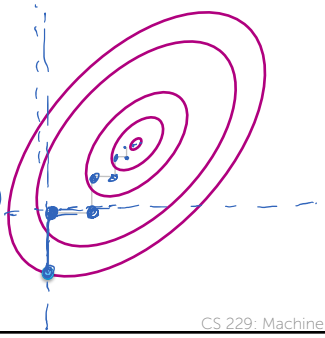
while not converged

pick a coordinate j

$\hat{w}_j \leftarrow \min_{w} g(\hat{w}_0, \hat{w}_1, \ldots, \hat{w}_{j-1}, w, w_{j+1}, \ldots, \hat{w}_D)$

values from prev. iter

just min jth coord



k-means as a coordinate descent algorithm

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

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

equivalent to

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

k-means as a coordinate descent algorithm

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

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

©2024 Emily Fox

k-means as a coordinate descent algorithm

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**

©2024 Emily Fox

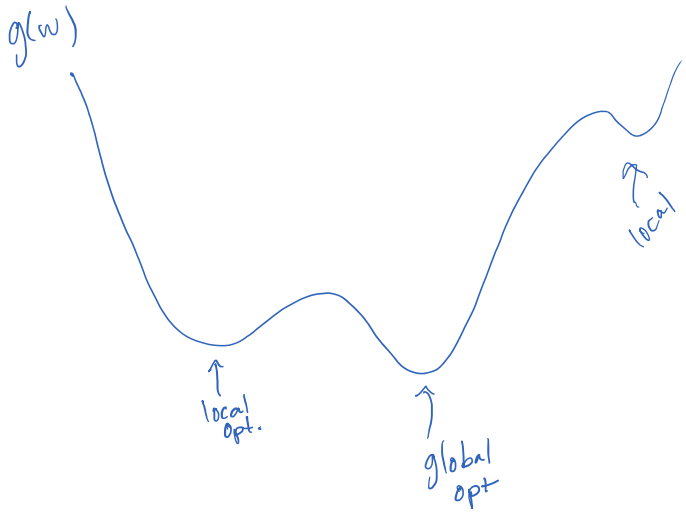
Convergence of k-means

Converges to:

- ~~Global optimum~~

- Local optimum

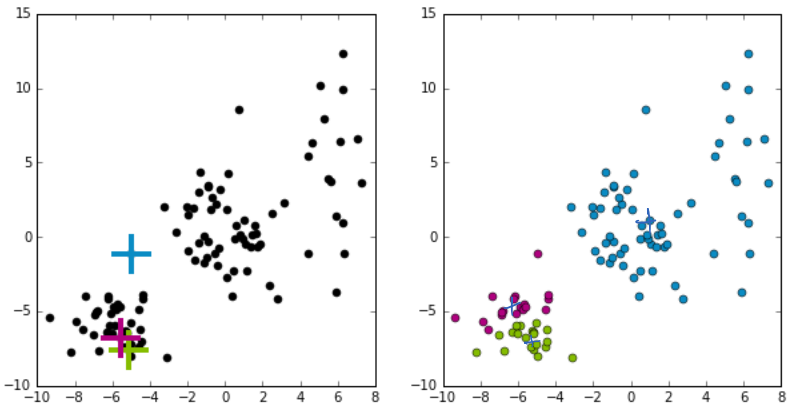
- ~~neither~~



©2024 Emily Fox

25

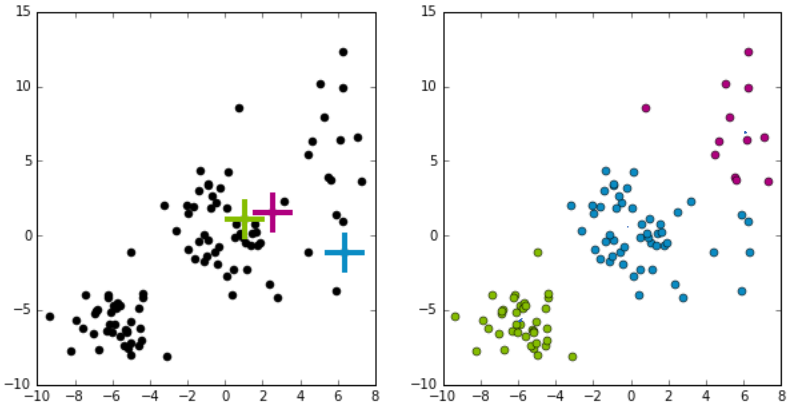
Convergence of k-means to local mode



©2024 Emily Fox

26

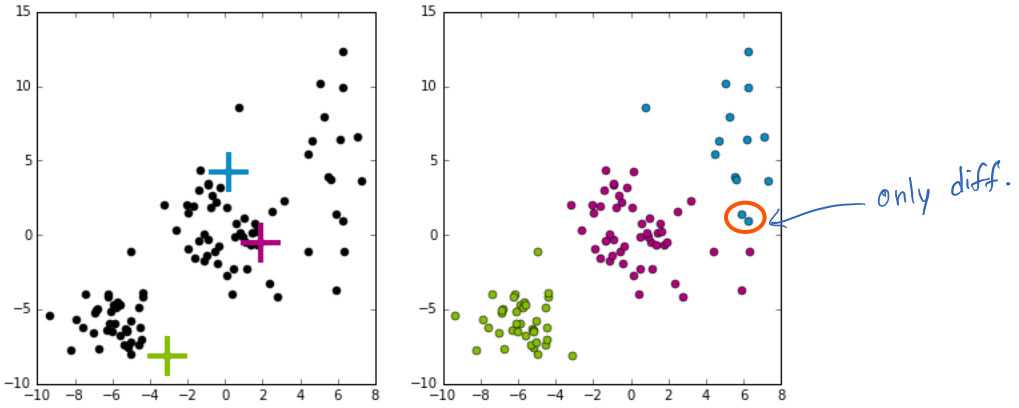
Convergence of k-means to local mode



©2024 Emily Fox

27

Convergence of k-means to local mode



©2024 Emily Fox

28

Smart initialization with k-means++

©2024 Emily Fox

29

k-means++ overview

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

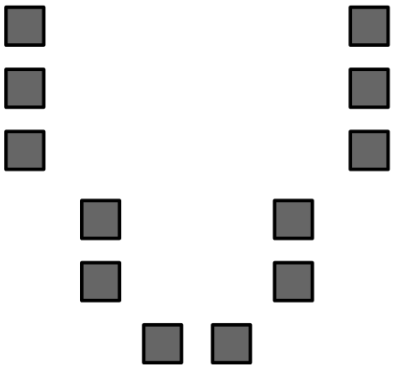
Smart initialization:

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

©2024 Emily Fox

30

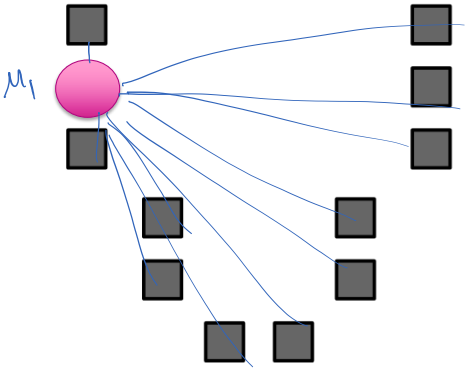
k-means++ visualized



©2024 Emily Fox

31

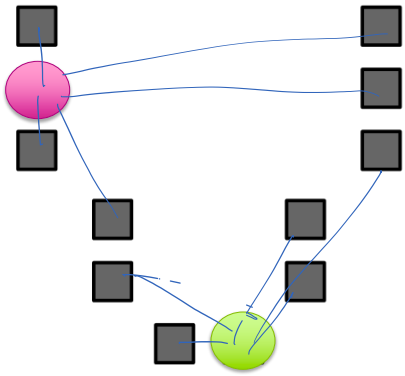
k-means++ visualized



©2024 Emily Fox

32

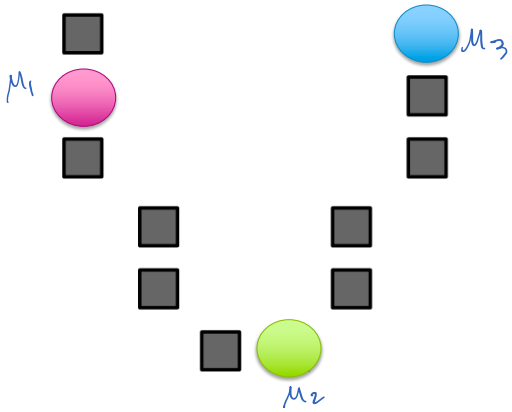
k-means++ visualized



©2024 Emily Fox

33

k-means++ visualized



©2024 Emily Fox

34

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

©2024 Emily Fox

35

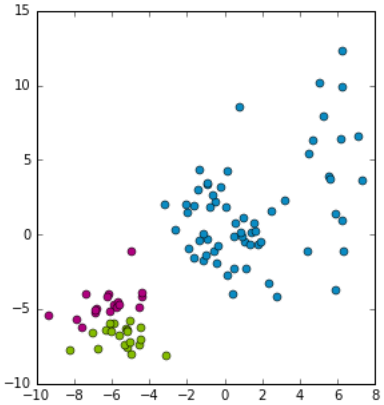
Assessing quality of the clustering and choosing the # of clusters

©2024 Emily Fox

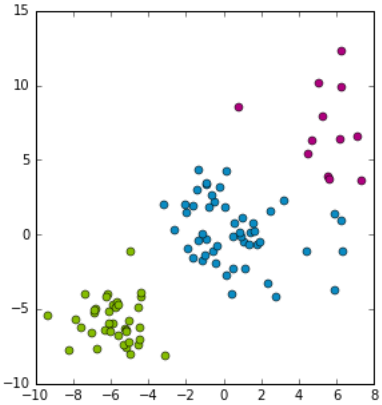
36

Which clustering does k-means prefer?

Option A



Option B



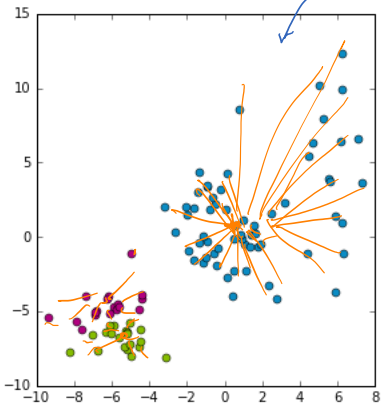
©2024 Emily Fox

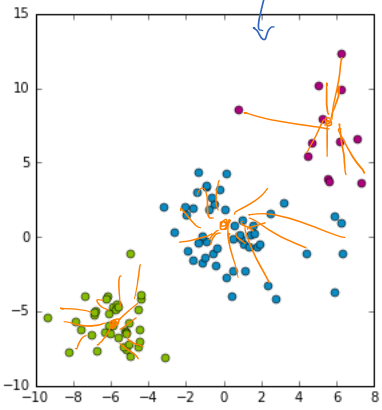
37

k-means objective

more heterogeneous

more homogeneous





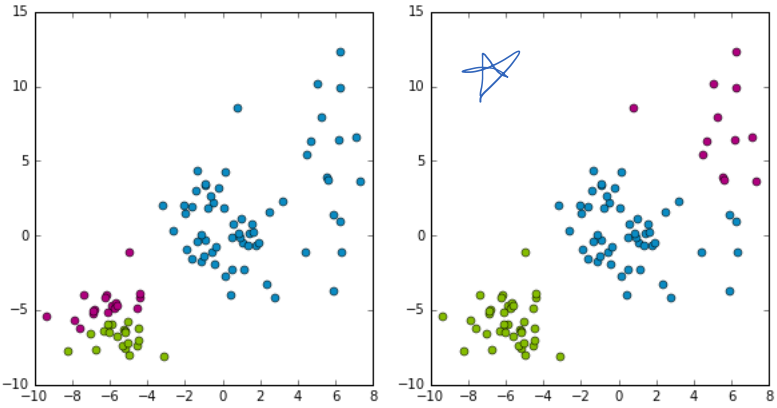
k-means is trying to minimize the sum of squared distances:

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

©2024 Emily Fox

38

Cluster heterogeneity



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!

©2024 Emily Fox

39

What happens as k increases?

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 ! (all distances to nearest cluster center are = 0)

Lowest possible cluster heterogeneity decreases with increasing k

©2024 Emily Fox

40

How to choose k?

©2024 Emily Fox

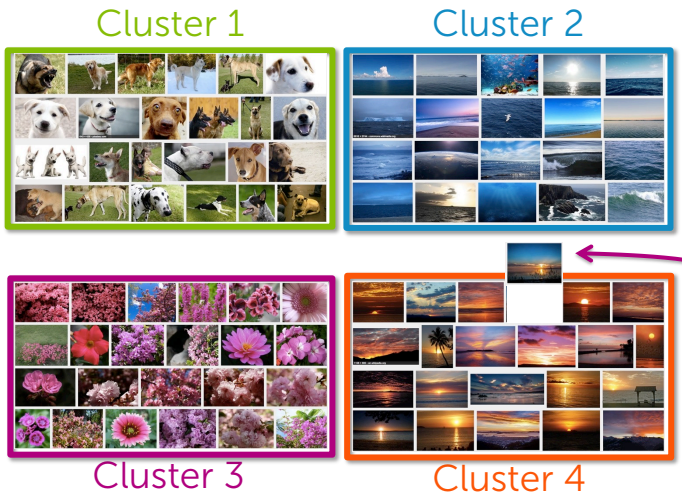
41

Limitations & failure modes of k-means

©2024 Emily Fox

42

Uncertainty in cluster assignments



Hard assignments don't tell full story

Slightly closer to Cluster 4 than Cluster 2, but count fully for Cluster 4?

©2024 Emily Fox

43

Other limitations of k-means

Assign observations to closest cluster center

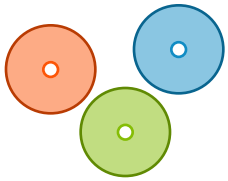
$$z_i \leftarrow \arg\min_j ||\mu_j - \mathbf{x}_i||_2^2$$

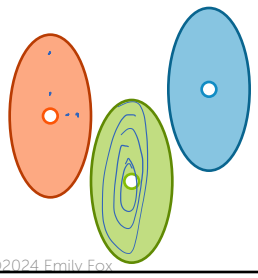
Can use weighted Euclidean, but requires *known* weights

Only center matters

Still assumes all clusters have the same axis-aligned ellipses

Equivalent to assuming spherically symmetric clusters

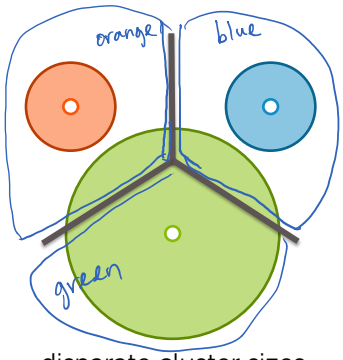




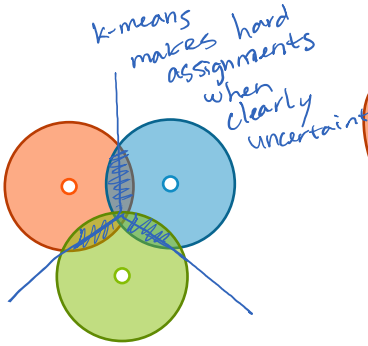
©2024 Emily Fox

44

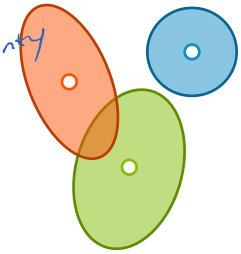
Failure modes of k-means



disparate cluster sizes



overlapping clusters



different shaped/
oriented clusters

©2024 Emily Fox

45

Motivates probabilistic model: **Mixture model**

- Provides **soft assignments** of observations to clusters (uncertainty in assignment)
  - e.g., 54% chance image is clouds, 45% sunset, 1% dog, and 0% pink flower
- Accounts for cluster shapes not just centers
- Enables learning weightings of dimensions
  - e.g., how much to weight each feature when computing cluster assignment

©2024 Emily Fox

46

Summary for k-means

47

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 observations using k-means
- Cast k-means as a coordinate descent algorithm and discuss convergence
- Describe potential applications of clustering

48

Mixture Models:
Model-Based Clustering

CS 229: Machine Learning
Emily Fox
Stanford University
February 28, 2024

©2024 Emily Fox

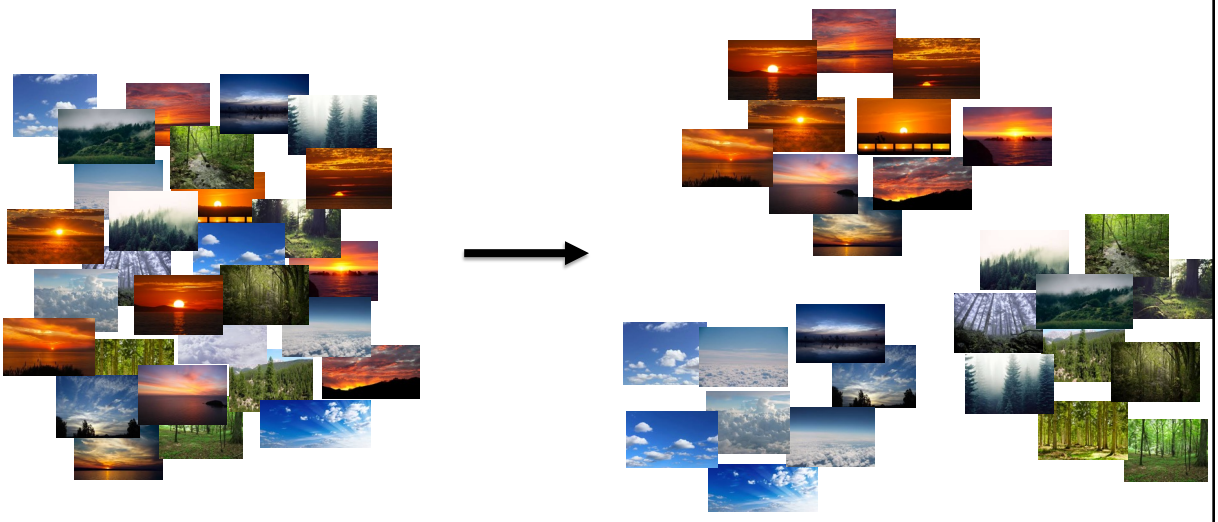
49

Again: Clustering images

©2024 Emily Fox

50

Case study: Clustering images



©2024 Emily Fox

51

Simple image representation

Consider average red, green, blue pixel intensities

$x_i \in \mathbb{R}^3$

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

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

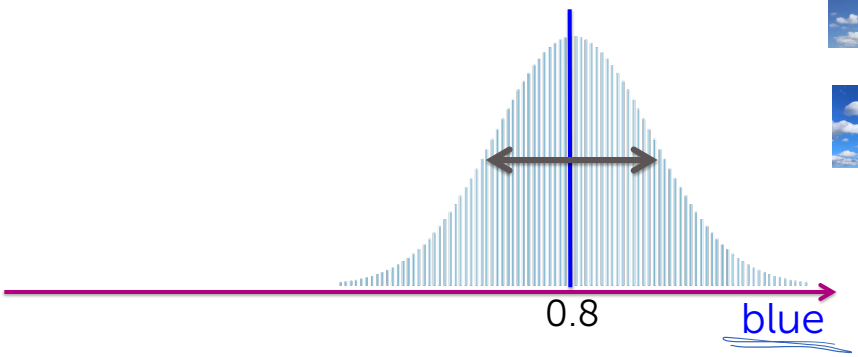
[R = 0.02, G = 0.95, B = 0.4]

©2024 Emily Fox

52

Distribution over all cloud images

Let's look at just the blue dimension

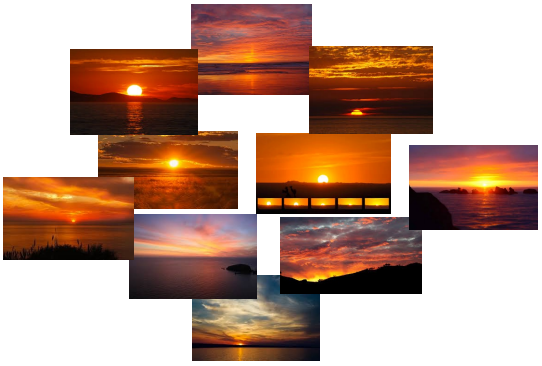


©2024 Emily Fox

53

Distribution over all sunset images

Let's look at just the blue dimension

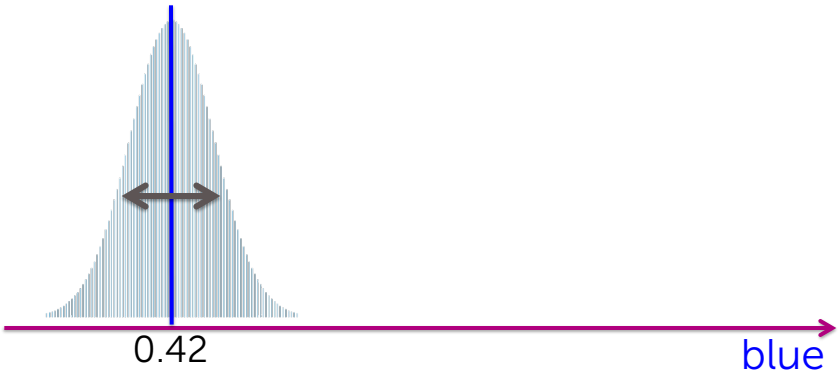


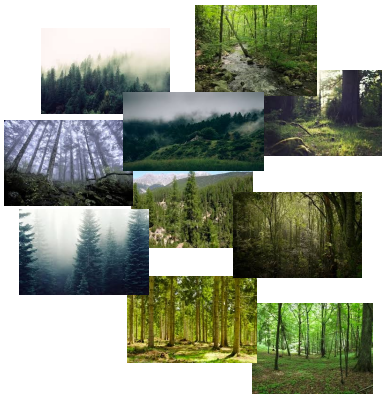
©2024 Emily Fox

54

Distribution over all forest images

Let's look at just the blue dimension

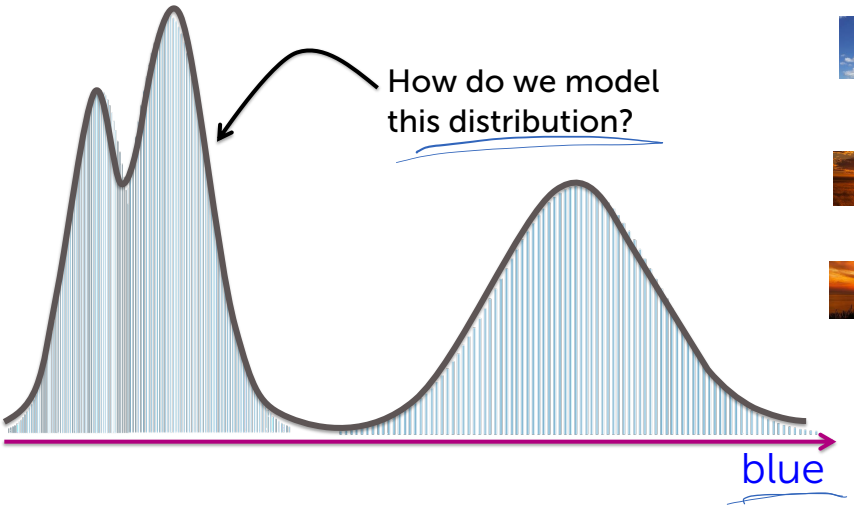




©2024 Emily Fox

55

Distribution over all images



©2024 Emily Fox

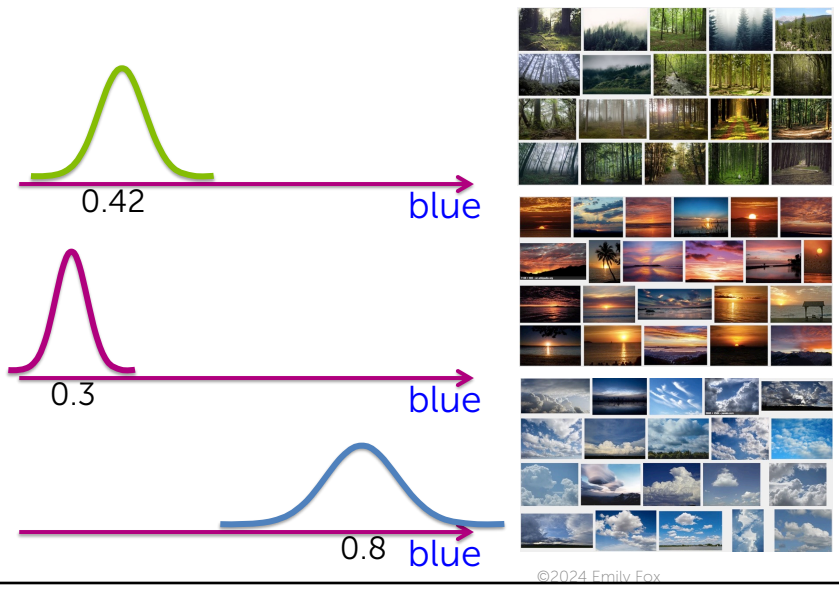
56

Mixture of Gaussians

©2024 Emily Fox

57

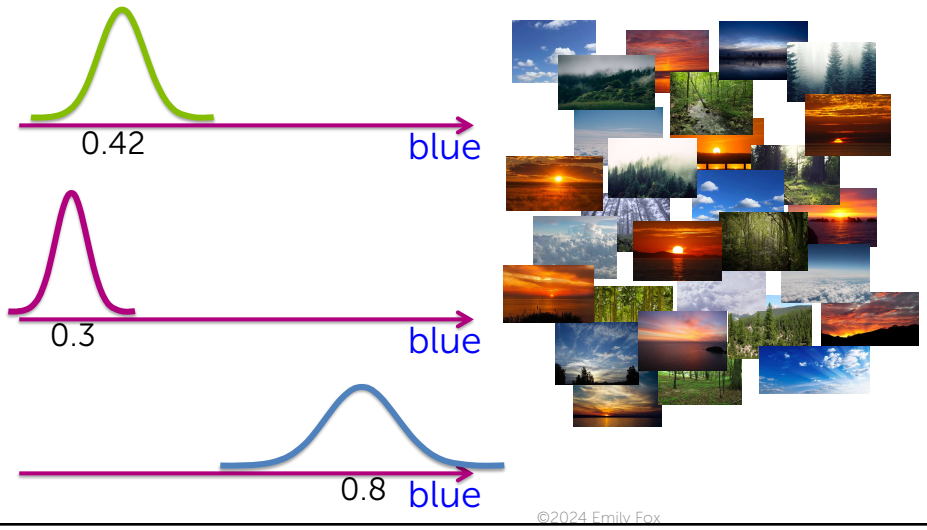
Model as Gaussian per cluster



©2024 Emily Fox

58

Model of jumble of unlabeled images

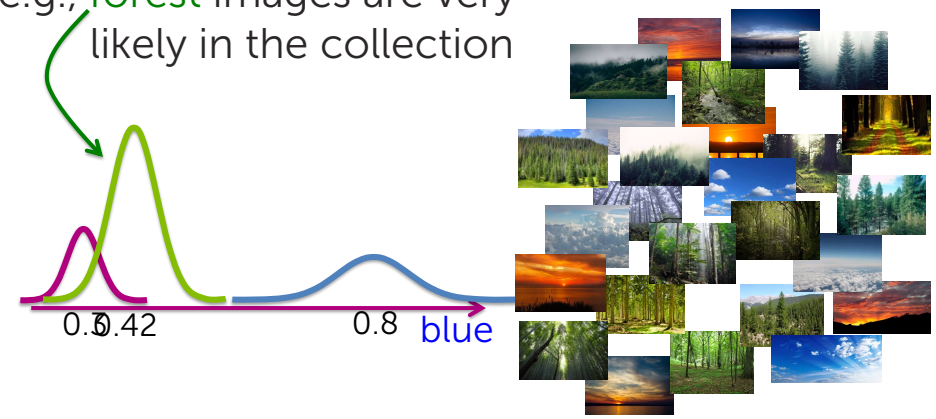


©2024 Emily Fox

59

What if image types not equally represented?

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



©2024 Emily Fox

60

Combination of weighted Gaussians

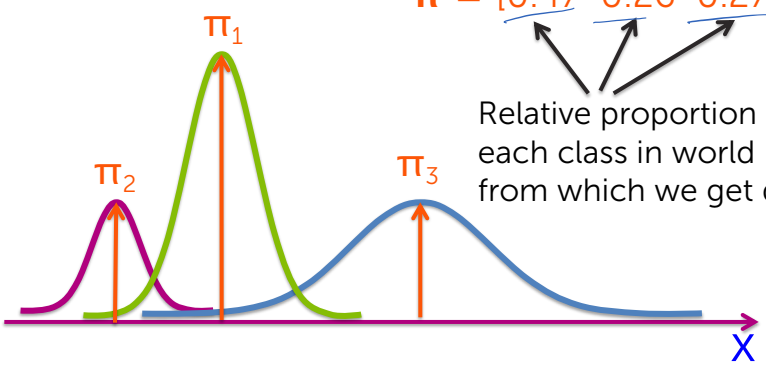
Associate a weight π_k with each Gaussian component

$\boldsymbol{\pi} = [\pi_1 \; \pi_2 \; \pi_3] = [0.47 \; 0.26 \; 0.27]$

Relative proportion of each class in world from which we get data

$0 \le \pi_k \le 1$

$\sum_{k=1}^{K} \pi_k = 1$

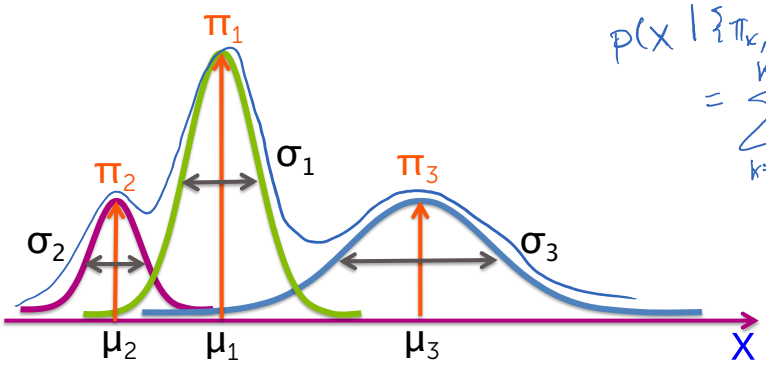


61

Mixture of Gaussians (1D)

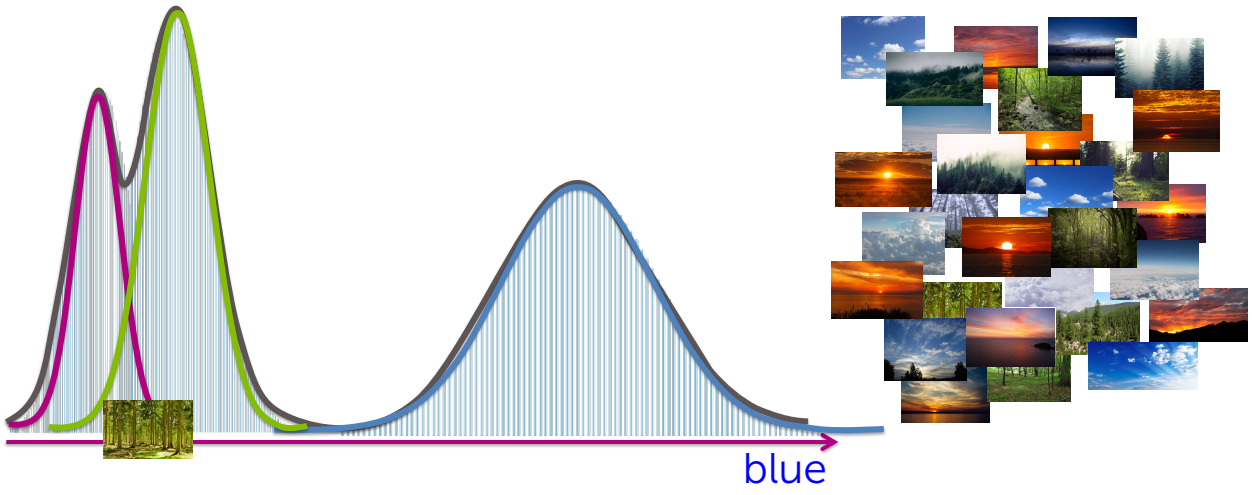
Each mixture component represents a unique cluster specified by: $\{\pi_k, \mu_k, \sigma_k^2\}$

$p(x \mid \{\pi_k, \mu_k, \sigma_k^2\}) = \sum_{k=1}^{K} \pi_k \, N(x \mid \mu_k, \sigma_k^2)$



62

Distribution over all images

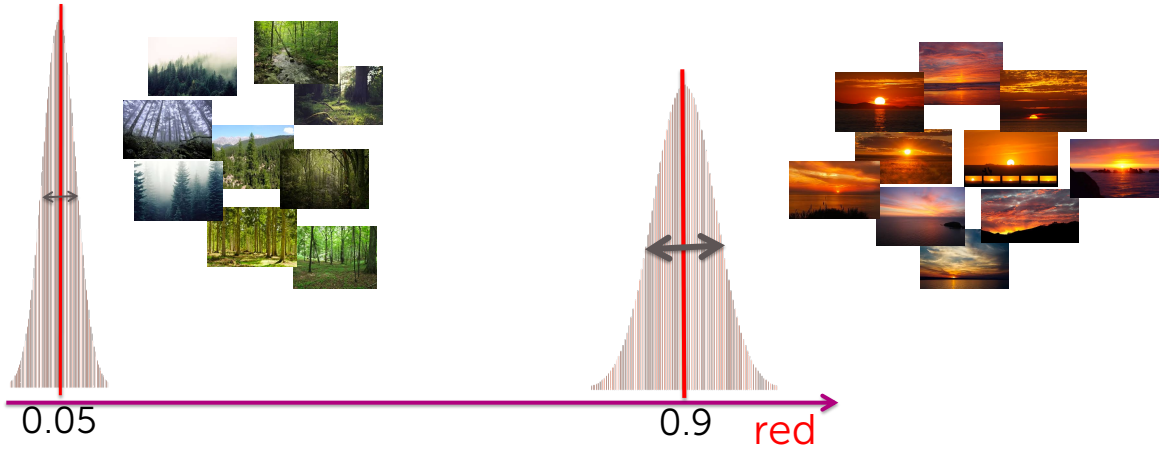


©2024 Emily Fox

63

Can be distinguished along other dim

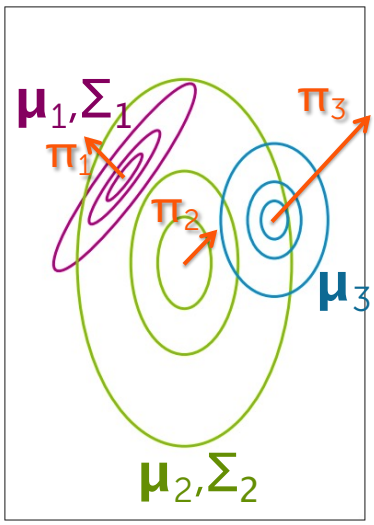
Now look at the red dimension



©2024 Emily Fox

64

Mixture of Gaussians (general)



Each mixture component represents a unique cluster specified by: $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$

$$p(x \mid \{\pi_k, \mu_k, \Sigma_k\}) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

©2024 Emily Fox

65

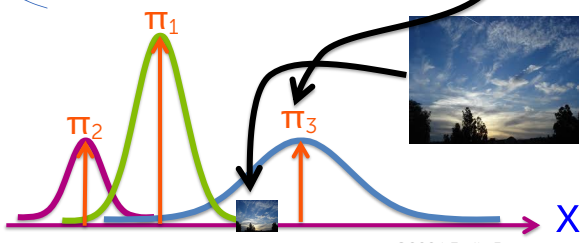
According to the model…

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")

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



©2024 Emily Fox

66

Summary for mixture models

©2024 Emily Fox

67

What you can do now…

- Interpret a probabilistic model-based approach to clustering using mixture models
- Describe model parameters
- Motivate the utility of soft assignments and describe what they represent
- Compare and contrast mixtures of Gaussians and k-means

©2024 Emily Fox

68

Background: Gaussian distributions

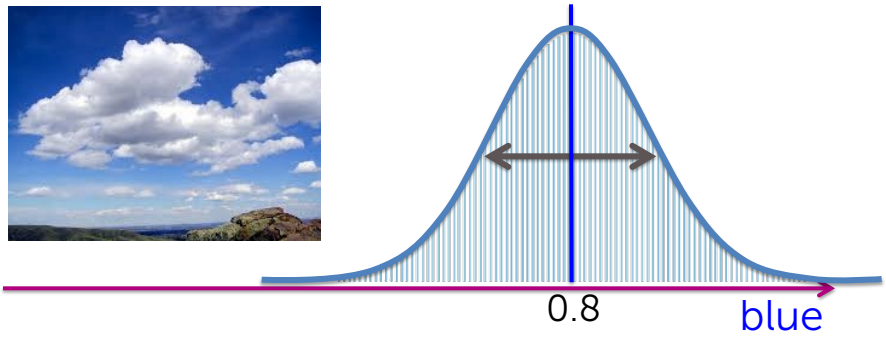
OPTIONAL

©2024 Emily Fox

69

Model for a given image type

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

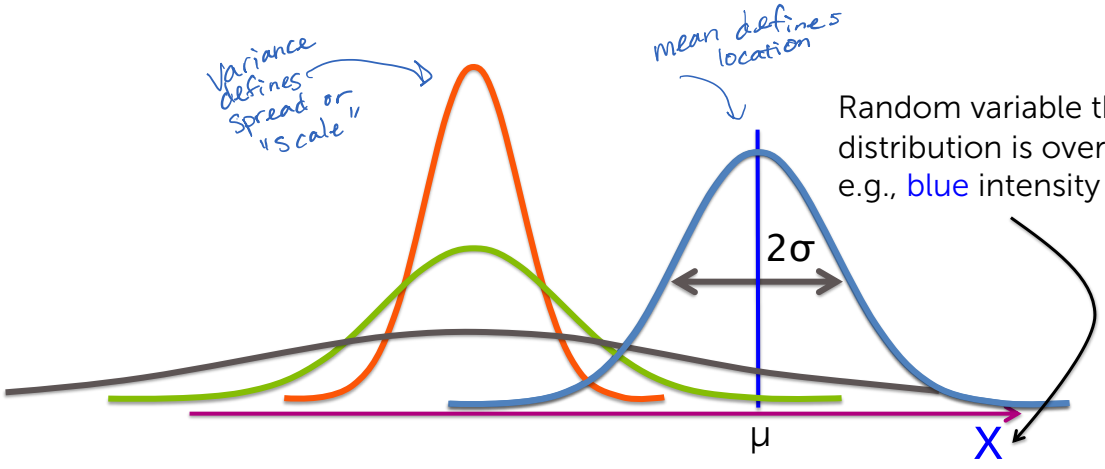


©2024 Emily Fox

70

1D Gaussians

Fully specified by **mean** μ and **variance** σ^2 (or **st. dev.** σ)



©2024 Emily Fox

71

Notating a 1D Gaussian distribution

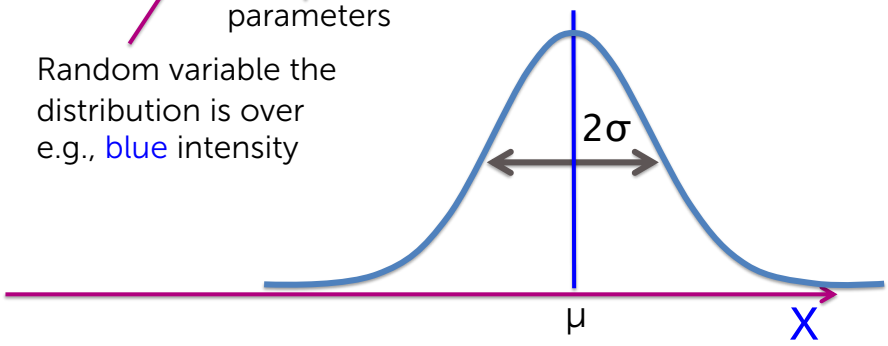
given

$N(x \mid \mu, \sigma^2)$

$(N(x; \mu, \sigma^2))$

parameters

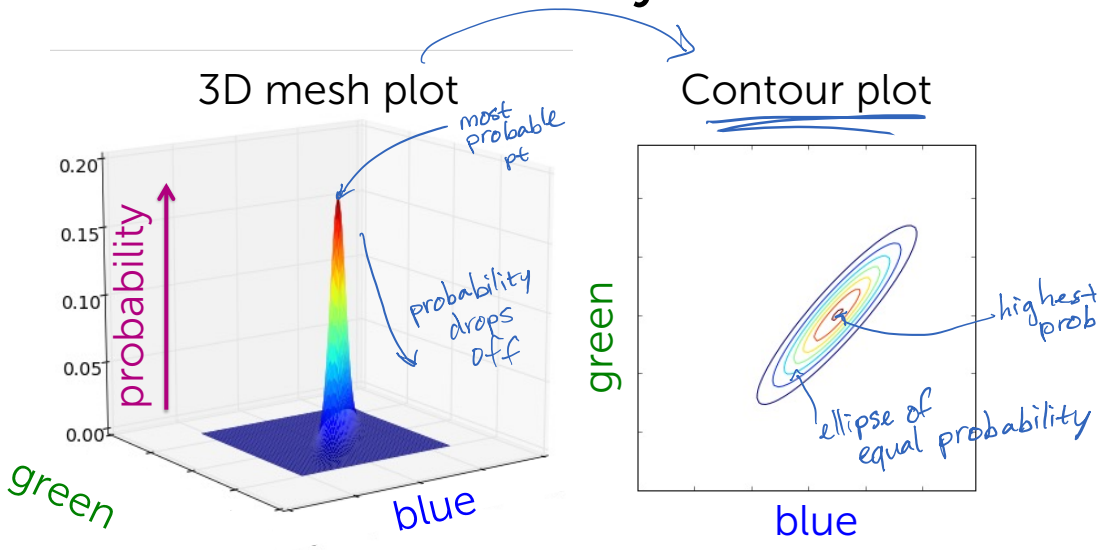
Random variable the distribution is over e.g., blue intensity



©2024 Emily Fox

72

2D Gaussians – Bird's eye view



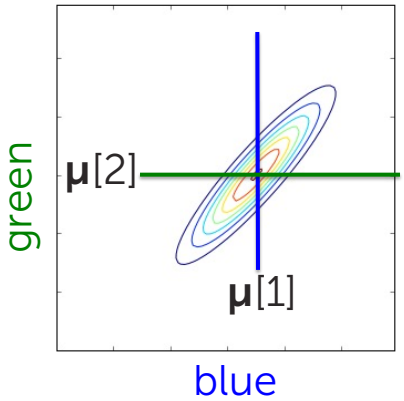
©2024 Emily Fox

2D Gaussians – Parameters

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

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

mean centers the distribution in 2D



©2024 Emily Fox

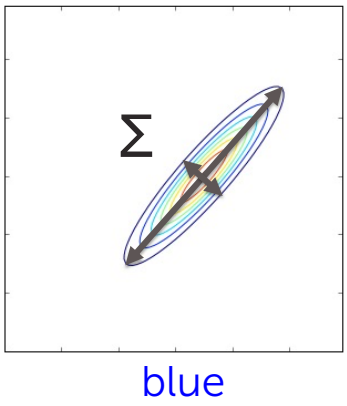
2D Gaussians – Parameters

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

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

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

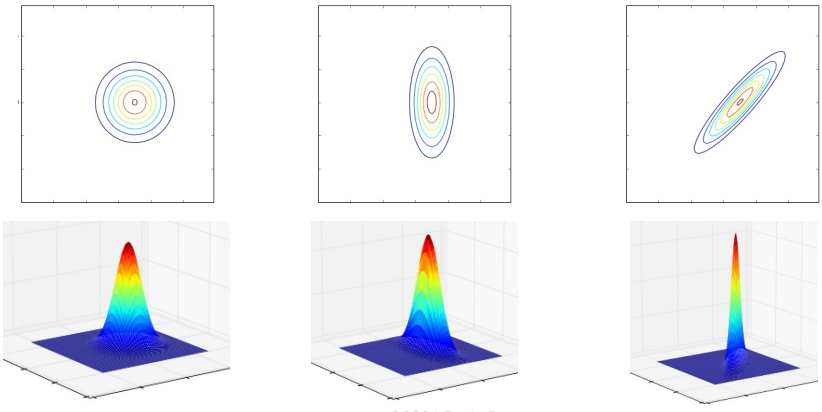
covariance determines orientation + spread



©2024 Emily Fox

Covariance structures

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_B^2 & 0 \\ 0 & \sigma_G^2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_B^2 & \sigma_{B,G} \\ \sigma_{G,B} & \sigma_G^2 \end{pmatrix}$$



©2024 Emily Fox

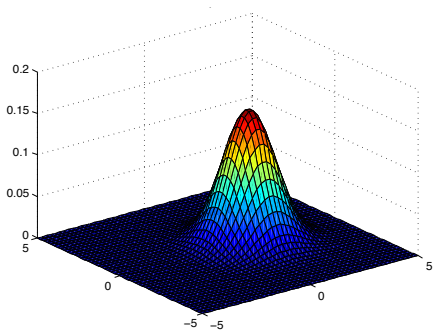
Notating a multivariate Gaussian

$\in \mathbb{R}^d$ $\in \mathbb{R}^d$ $\in \mathbb{R}^{d \times d}$ (pos. semidef.)

$$N(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma)$$

parameters

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



©2024 Emily Fox

77

Parameter estimation for 1D Gaussian distributions (and more background)

OPTIONAL

©2024 Emily Fox

78

Some properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \rightarrow Y \sim N(a\mu+b, a^2\sigma^2)$

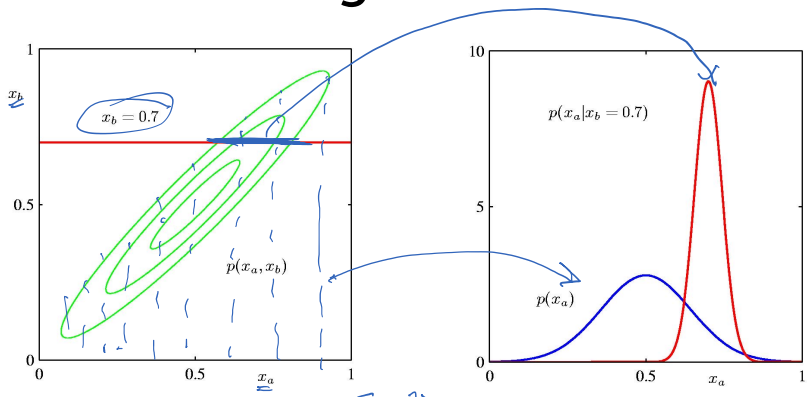
same affine trans: $E[aX+b] = aE[x]+b = a\mu+b$

known (deterministic) quantities (scalars)

- Sum of independent Gaussian random variables
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

©2024 Emily Fox

Conditional & marginal distributions



$\begin{pmatrix} X_a \\ X_b \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \right)$

Marg: $p(x_a) = \int p(x_a, x_b)\, dx_b \rightarrow X_a \sim N(\mu_a, \Sigma_{aa})$

Cond: $p(x_a|x_b) = \frac{p(x_a,x_b)}{p(x_b)}$ ← $\int p(x_a,x_b)\, dx_a$

$\rightarrow X_a|X_b \sim N(\mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b), \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})$

fixed

©2024 Emily Fox

Learning a 1D Gaussian

- Collect a bunch of data
  - Assume i.i.d. samples
- Learn parameters
  - Mean $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$ ← Why? MLE
  - Variance $\hat{\sigma}^2$

$$p(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

©2024 Emily Fox

MLE for 1D Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_N\}$:

$$p(D \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

datapt x_i

$$\hat{\mu}^{MLE}, \hat{\sigma}^{MLE} = \arg\max_{\mu,\sigma} p(D \mid \mu, \sigma) = \arg\max_{\mu,\sigma} \ln p(D \mid \mu, \sigma)$$

- Log-likelihood of data:

$$\ln p(D \mid \mu, \sigma) = \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right]$$

$$= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

©2024 Emily Fox

MLE for mean of a 1D Gaussian

- What's MLE for the mean?

$$\frac{d}{d\mu} \ln p(D \mid \mu, \sigma) = \frac{d}{d\mu}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}\right] = 0$$

no depend. on μ

$$\frac{d}{d\mu} \ln p(D \mid \mu, \sigma) = -\sum_{i=1}^{N} \frac{d}{d\mu} \frac{(x_i-\mu)^2}{2\sigma^2} = \sum_{i=1}^{N} \frac{x_i - \mu}{\sigma^2} = 0$$

$$N \hat{\mu}^{MLE} = \sum_{i=1}^{N} x_i$$

$$\hat{\mu}^{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i$$ ← MLE doesn't depend on choice of σ^2

MLE for variance

- Again, set derivative to zero:

$$\frac{d}{d\sigma} \ln p(D \mid \mu, \sigma) = \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

$$= \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N} \frac{d}{d\sigma}\left[\frac{(x_i - \mu)^2}{2\sigma^2}\right] = 0$$

$-\frac{N}{\sigma}$ $\quad -\frac{(x_i-\mu)^2}{\sigma^3}$

$$\Rightarrow -\frac{N}{\sigma} + \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{\sigma^3} = 0$$

$$\Rightarrow \hat{\sigma}^{2\,MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu}^{MLE})^2$$

use $\mu = \hat{\mu}^{MLE}$ bc optimal choice of μ doesn't depend on σ

Learning 1D Gaussian parameters

- MLE: $\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{MLE})^2$$

- FYI, MLE for the variance of a Gaussian is **biased**
  - Expected value of estimator is **not** true parameter!
  - Unbiased variance estimator:

$$\hat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu}_{MLE})^2$$

©2024 Emily Fox