

So far, supervised learning

$h: X \rightarrow \mathbb{R}$ "regression"

$h: X \rightarrow \{0, 1, \dots, K\}$
"classification"

Unsupervised

Clustering K-means

Machine Learning – CSE546

Emily Fox

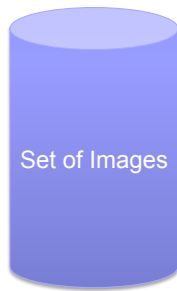
University of Washington

November 4, 2013

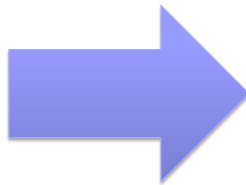
©Carlos Guestrin 2005-2013

1

Clustering images

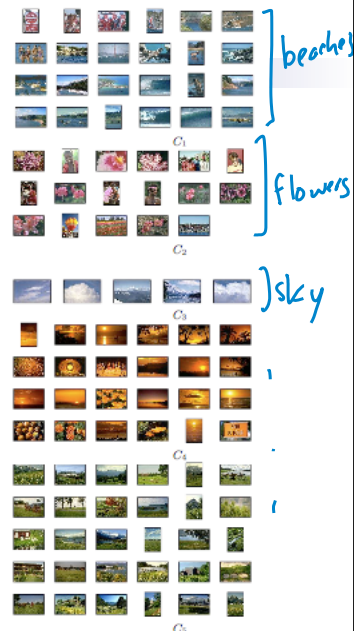


Set of Images



organize into
coherent "themes"

key: no labels given



©Carlos Guestrin 2005-2013

[Goldberger et al.] 2

Clustering web search results

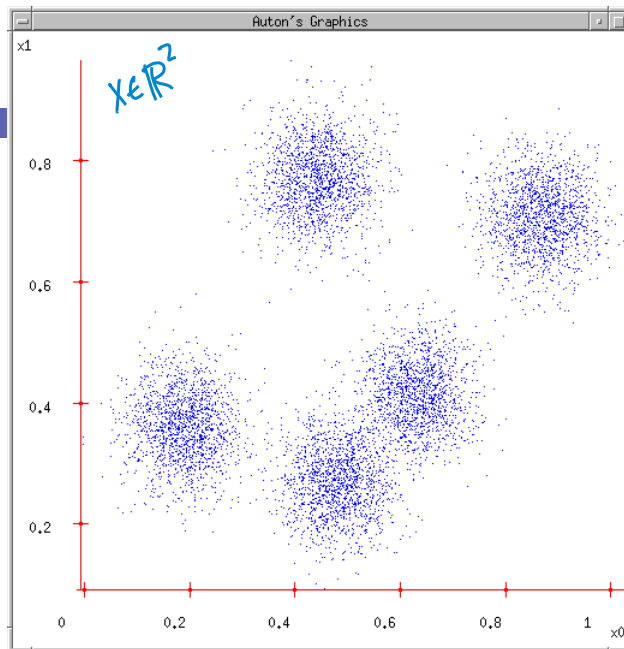
The screenshot shows the Clusty search engine interface. The search term is "race". The sidebar on the left lists various clusters, with "Human" selected. The main content area displays a list of search results, including Wikipedia entries, news articles, and academic papers. A blue arrow points from the "Human" cluster in the sidebar to the first result, which is a Wikipedia article about the classification of human beings. Another blue arrow points from the "Human" cluster to the "Human" cluster in the sidebar. A third blue arrow points from the "Human" cluster to the "Human" cluster in the sidebar.

Cluster Human contains 8 documents.

- Race (classification of human beings) - Wikipedia, the free ...**
The term **race** or racial group usually refers to the concept of dividing **humans** into populations or groups on the basis of various sets of characteristics. The most widely used **human** racial categories are based on visible traits (especially skin color, cranial or facial features and hair texture), and self-identification. Conceptions of **race**, as well as specific ways of grouping **rac**es, vary by culture and over time, and are often controversial for scientific as well as social and political reasons.History - Modern debates - Political and ...
en.wikipedia.org/wiki/Race_(classification_of_human_beings) - [cache] - Live, Ask
- Race - Wikipedia, the free encyclopedia**
General: **Racing** competitions The **Race** (yachting **race**), or La course du millénaire, a no-rules round-the-world sailing event; **Race** (biology), classification of flora and fauna; **Race** (classification of human beings) **Race** and ethnicity in the United States Census, official definitions of "race" used by the US Census Bureau; **Race** and genetics, notion of racial classifications based on genetics. Historical definitions of **race**; **race** (bearing), the inner and outer rings of a rolling-element bearing. **RACE** in molecular biology "Rapid ... General - Surnames - Television - Music - Literature - Video games
en.wikipedia.org/wiki/Race - [cache] - Live, Ask
- Publications | Human Rights Watch**
The use of torture, unlawful rendition, secret prisons, unfair trials, ... Risks to Migrants, Refugees, and Asylum Seekers in Egypt and Israel ... In the run-up to the Beijing Olympics in August 2008, ...
www.hrw.org/background/usa/race - [cache] - Ask
- Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich ...**
Amazon.com: **Race: The Reality Of Human Differences: Vincent Sarich, Frank Miele:** Books ... From Publishers Weekly Sarich, a Berkeley emeritus anthropologist, and Miele, an editor ...
www.amazon.com/Race-Reality-Differences-Vincent-Sarich/dp/0813340881 - [cache] - Live
- AAPA Statement on Biological Aspects of Race**
AAPA Statement on Biological Aspects of **Race** ... Published in the American Journal of Physical Anthropology, vol. 101, pp 569-570, 1996 ... PREAMBLE As scientists who study human evolution and variation, ...
www.physanth.org/positions/race.html - [cache] - Ask
- race, Definition from Answers.com**
race n. A local geographic or global human population distinguished as a more or less distinct group by genetically transmitted physical
www.answers.com/topic/race-1 - [cache] - Live
- Dopefish.com**
Site for newbies as well as experienced Dopefish followers, chronicing the birth of the Dopefish, its numerous appearances in several computer games, and its eventual take-over of the human **race**. Maintained by Mr. Dopefish himself, Joe Siegler of Apogee Software.
www.dopefish.com - [cache] - Open Directory

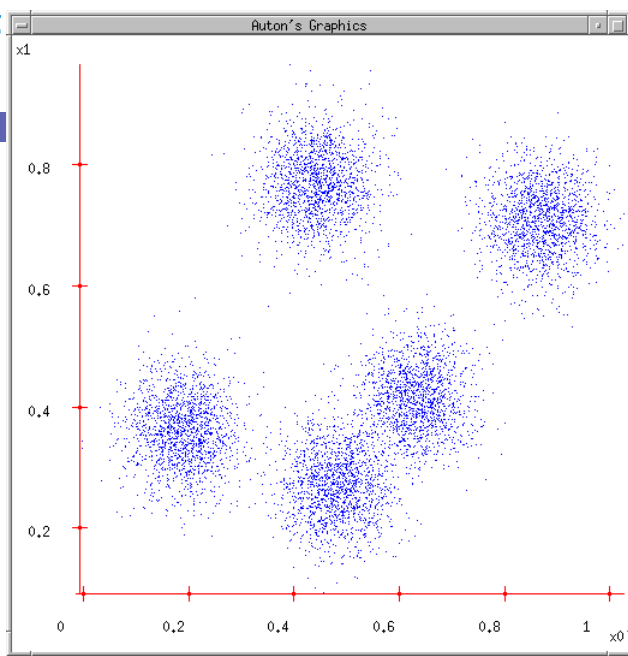
©Carlos Guestrin 2005-2013 3

Some Data



K-means

1. Ask user how many clusters they'd like. (e.g. $k=5$)

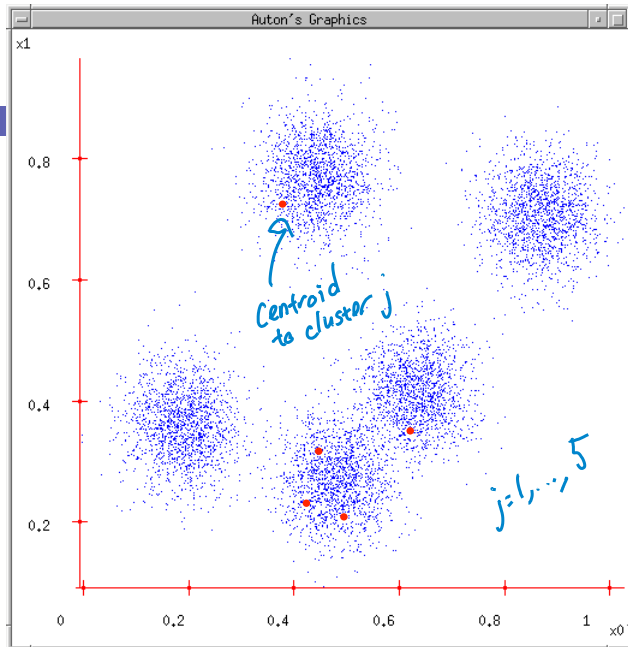


©Carlos Guestrin 2005-2013

5

K-means

1. Ask user how many clusters they'd like. (e.g. $k=5$)
2. Randomly guess k cluster Center locations



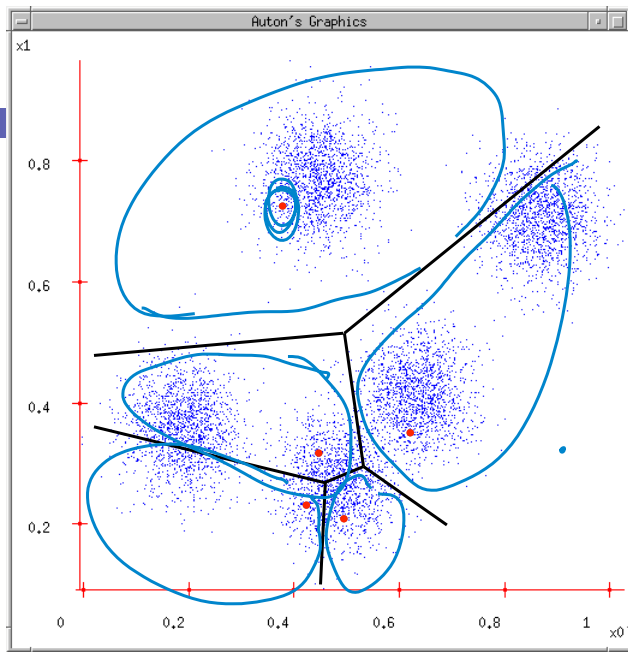
©Carlos Guestrin 2005-2013

6

K-means

1. Ask user how many clusters they'd like. (e.g. $k=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)

Voronoi tessellation

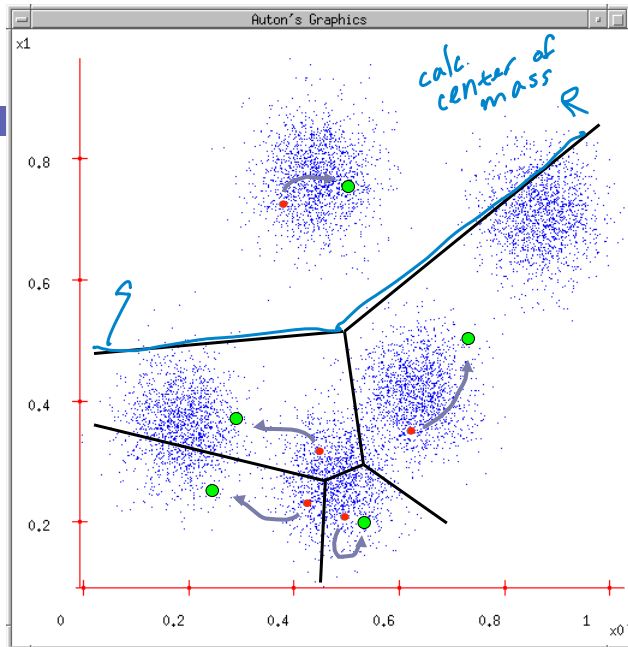


©Carlos Guestrin 2005-2013

7

K-means

1. Ask user how many clusters they'd like. (e.g. $k=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns

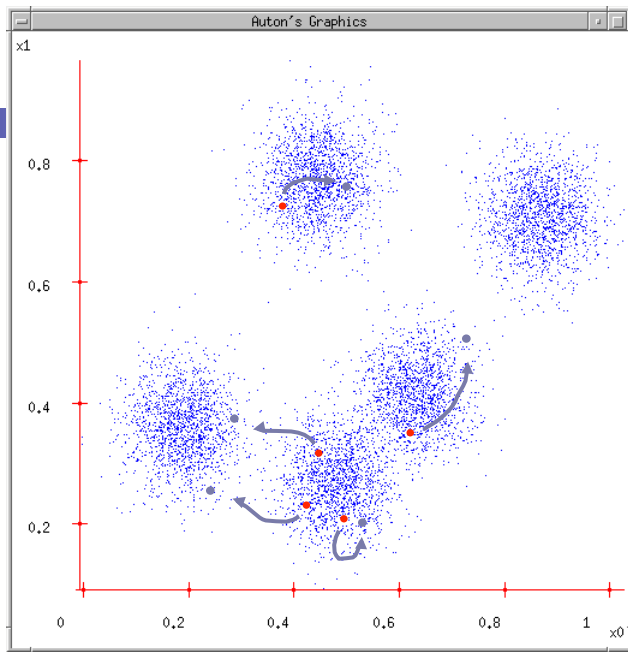


©Carlos Guestrin 2005-2013

8

K-means

1. Ask user how many clusters they'd like. (e.g. $k=5$)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!



©Carlos Guestrin 2005-2013

9

K-means

- Randomly initialize k centers (or "smartly")
 - $\mu^{(0)} = \mu_1^{(0)}, \dots, \mu_k^{(0)}$ *iteration*
- **Classify:** Assign each point $j \in \{1, \dots, N\}$ to nearest center:
 - $C^{(t)}(j) \leftarrow \arg \min_i \|\mu_i - x_j\|^2$ *fix μ , opt. C*
 - $C(j)=k \Rightarrow j$ th obs. is assoc. w/ cluster k*
- **Recenter:** μ_i becomes centroid of its point: *fix C , opt. μ*
 - $\mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j: C(j)=i} \|\mu - x_j\|^2 \leftarrow \mu_i = \frac{\sum_{j: C(j)=i} x_j}{|\{j: C(j)=i\}|}$
 - Equivalent to $\mu_i \leftarrow$ average of its points! *$|\{j: C(j)=i\}|$*

©Carlos Guestrin 2005-2013

10

What is K-means optimizing?

- Potential function $F(\mu, C)$ of centers μ and point allocations C :

$$\square F(\mu, C) = \sum_{j=1}^N \|\mu_{C(j)} - x_j\|^2$$

\nearrow \uparrow
k centers allocations

- Optimal K-means:

$$\square \min_{\mu} \min_C F(\mu, C)$$

\curvearrowright optimization is hard, but iteratively easy $\uparrow\uparrow$ coordinate descent!

Does K-means converge??? Part 1

- Optimize potential function:

$$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j: C(j)=i} \|\mu_i - x_j\|^2$$

- Fix μ , optimize C

$$\min_{C: (c(1), \dots, c(N))} \sum_{j=1}^N \|\mu_{c(j)} - x_j\|^2 = \min_{c(1)} \min_{c(2)} \dots \min_{c(N)} \sum_{j=1}^N \|\mu_{c(j)} - x_j\|^2$$

$$= \sum_{j=1}^N \min_{c(j)} \|\mu_{c(j)} - x_j\|^2$$

\uparrow ind. minimizations

exactly the "classification step"

Does K-means converge??? Part 2

- Optimize potential function:

$$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j:C(j)=i} \|\mu_i - x_j\|^2$$

- Fix C, optimize μ

$$\begin{aligned} \min_{\mu: \mu_1, \dots, \mu_k} \sum_{j=1}^k \|\mu_{C(j)} - x_j\|^2 &= \min_{\mu: \mu_1, \dots, \mu_k} \sum_{i=1}^k \sum_{j:C(j)=i} \|\mu_i - x_j\|^2 \\ &= \sum_{i=1}^k \min_{\mu_i} \sum_{j:C(j)=i} \|\mu_i - x_j\|^2 \\ &\text{"recenter step"} \quad \mu_i = \text{center of pts} = \text{avg} = \frac{\sum_{j:C(j)=i} x_j}{|\{j:C(j)=i\}|} \end{aligned}$$

©Carlos Guestrin 2005-2013

13

Coordinate descent algorithms

$$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j:C(j)=i} \|\mu_i - x_j\|^2 \geq 0$$

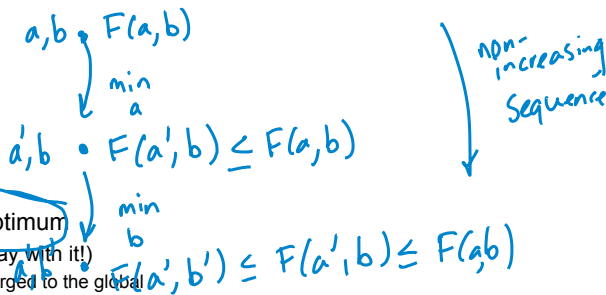
- Want: $\min_a \min_b F(a, b)$

- Coordinate descent:

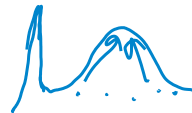
- fix a, minimize b
- fix b, minimize a
- repeat

- Converges!!!

- if F is bounded
- to a (often good) local optimum
 - as we saw in applet (play with it!)
 - (For LASSO it converged to the global optimum, because of convexity)



Random restarts help



- K-means is a coordinate descent algorithm!

©Carlos Guestrin 2005-2013

14

model that can be used for clustering, density est. :

Mixtures of Gaussians

Machine Learning – CSE546
Emily Fox
University of Washington
November 4, 2013
©Carlos Guestrin 2005-2013

15

(One) bad case for k-means

- Clusters may overlap
- Some clusters may be "wider" than others

shape

centers

params defining clusters

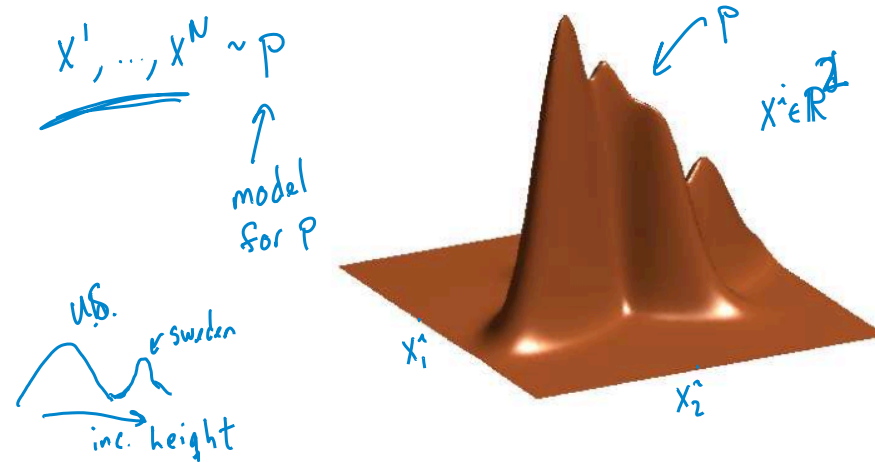
so centers alone don't tell the whole story

©Carlos Guestrin 2005-2013

16

Density Estimation

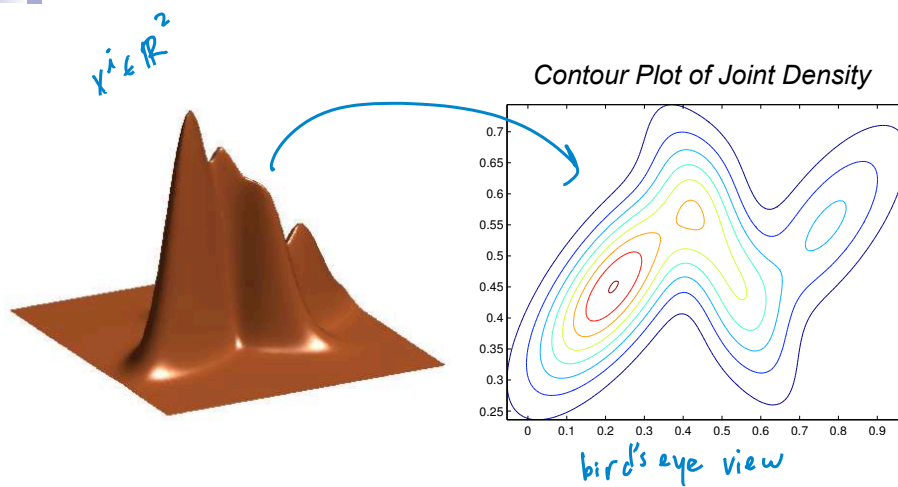
- Estimate a density based on x^1, \dots, x^N



©Emily Fox 2013

17

Density Estimation



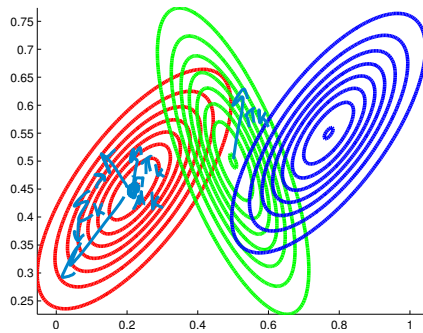
©Emily Fox 2013

18

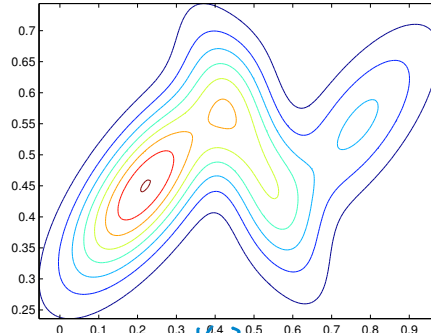
Density as Mixture of Gaussians

- Approximate density with a mixture of Gaussians

Mixture of K Gaussians



Contour Plot of Joint Density



Each Gaussian has weight π_k and shape params μ_k, Σ_k w/ $\sum_{k=1}^K \pi_k = 1$

©Emily Fox 2013

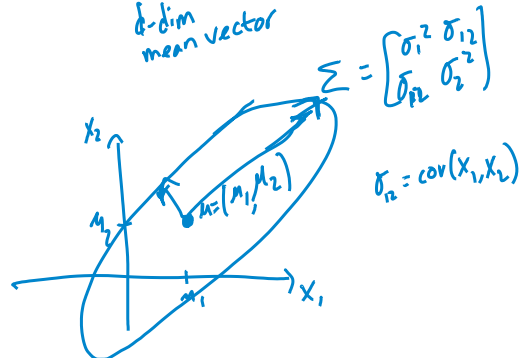
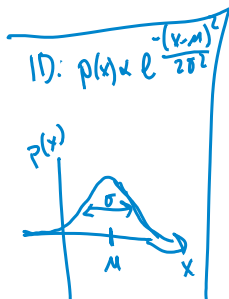
19

Gaussians in d Dimensions

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} \|\Sigma\|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right]$$

$d \times d$ covariance matrix Σ

d -dim mean vector μ



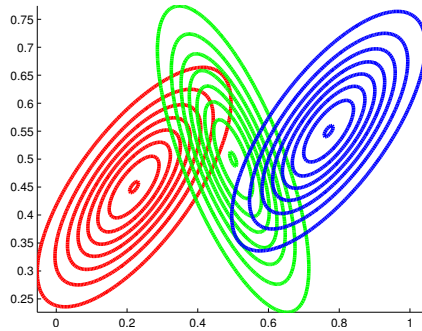
©Carlos Guestrin 2005-2013

20

Density as Mixture of Gaussians

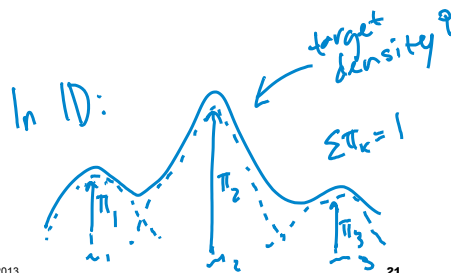
- Approximate density with a mixture of Gaussians

Mixture of 3 Gaussians



$$p(x^i | \pi, \mu, \Sigma) = \sum_{k=1}^K \pi_k N(x^i | \mu_k, \Sigma_k)$$

Handwritten notes: $\{\pi_1, \dots, \pi_K\}$ and $\{\mu_k, \Sigma_k\}$ are indicated by arrows pointing to the parameters in the equation.



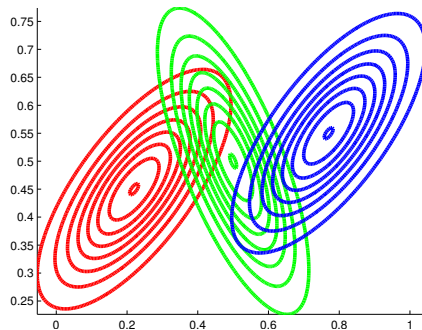
©Emily Fox 2013

21

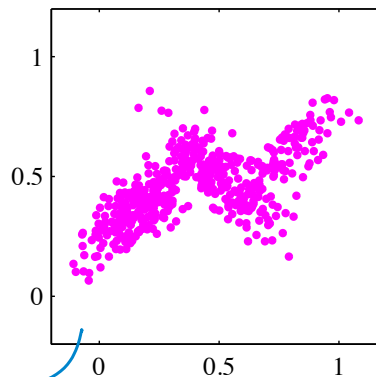
Density as Mixture of Gaussians

- Approximate with density with a mixture of Gaussians

Mixture of 3 Gaussians



Our actual observations

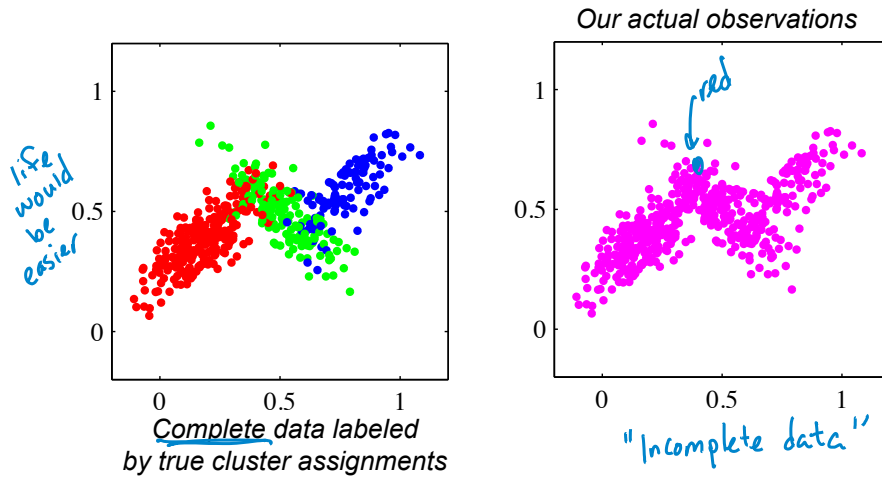


Handwritten notes: "How???" with an arrow pointing from the observations plot to the mixture plot, and "from obs., est. model params" below it.

C. Bishop, Pattern Recognition & Machine Learning

Clustering our Observations

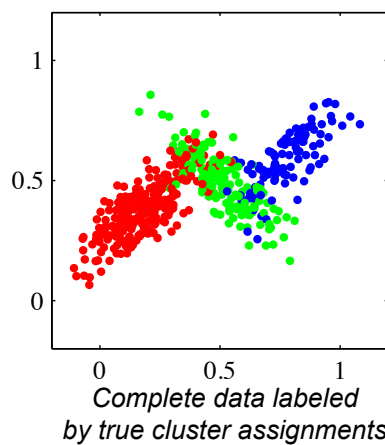
- Imagine we have an assignment of each x^i to a Gaussian



C. Bishop, Pattern Recognition & Machine Learning

Clustering our Observations

- Imagine we have an assignment of each x^i to a Gaussian

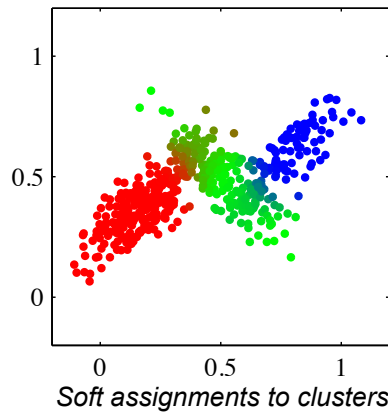


- Introduce latent cluster indicator variable z^i
 $z^i \in \{1, \dots, K\} \equiv (i) \rightarrow z^i$
 $\Pr(z^i = k) = \pi_k$
- Then we have
 $p(x^i | z^i = k, \mu, \Sigma) = N(x^i | \mu_k, \Sigma_k)$
 param est. is easy if we have $\{z^i\}$
 \Rightarrow decouples into K Gauss. est.

C. Bishop, Pattern Recognition & Machine Learning

Clustering our Observations

- We must infer the cluster assignments from the observations



- Posterior probabilities of assignments to each cluster *given* model parameters:

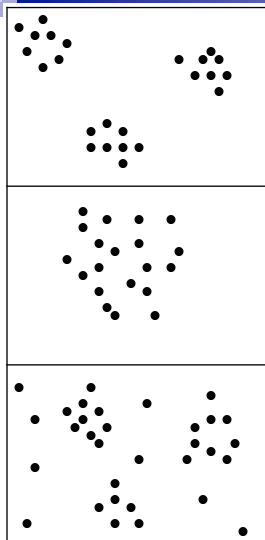
$$r_{ik} = p(z^i = k | x^i, \pi, \mu, \Sigma) =$$

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

motivates an iterative alg.

C. Bishop, Pattern Recognition & Machine Learning

Unsupervised Learning: not as hard as it looks



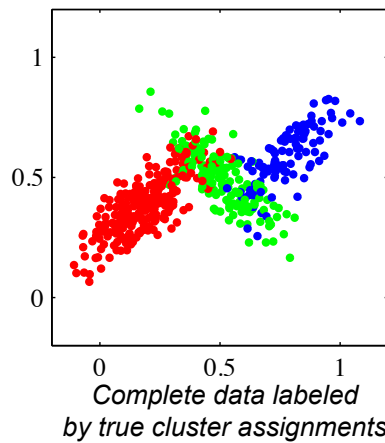
Sometimes easy

Sometimes impossible

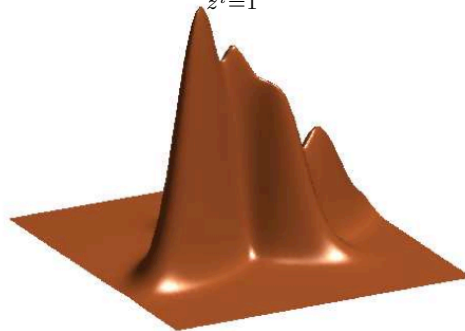
and sometimes in between

Summary of GMM Concept

- Estimate a density based on x^1, \dots, x^N



$$p(x^i | \pi, \mu, \Sigma) = \sum_{z^i=1}^K \pi_{z^i} \mathcal{N}(x^i | \mu_{z^i}, \Sigma_{z^i})$$



Surface Plot of Joint Density, Marginalizing Cluster Assignments

©Emily Fox 2013

27

Summary of GMM Components

- Observations $x^i \in \mathbb{R}^d, \quad i = 1, 2, \dots, N$
- Hidden cluster labels $z_i \in \{1, 2, \dots, K\}, \quad i = 1, 2, \dots, N$
- Hidden mixture means $\mu_k \in \mathbb{R}^d, \quad k = 1, 2, \dots, K$
- Hidden mixture covariances $\Sigma_k \in \mathbb{R}^{d \times d}, \quad k = 1, 2, \dots, K$
- Hidden mixture probabilities $\pi_k, \quad \sum_{k=1}^K \pi_k = 1$

Gaussian mixture marginal and conditional likelihood :

$$p(x^i | \pi, \mu, \Sigma) = \sum_{z^i=1}^K \pi_{z^i} p(x^i | z^i, \mu, \Sigma)$$

$$p(x^i | z^i, \mu, \Sigma) = \mathcal{N}(x^i | \mu_{z^i}, \Sigma_{z^i})$$

©Emily Fox 2013

28