

In question 2, you solve for  $\alpha \in \mathbb{R}^n$ :  $\hat{f}(x) = \sum_{i=1}^n \alpha_i K(x, x_i)$  ← This is a function that can be evaluated on an arbitrarily fine grid

# Announcements

$$\operatorname{argmin}_a \mathbb{E}[(z-a)^2] = \mathbb{E}[z] \quad \mathbb{E}[(z - \mathbb{E}[z])^2] \geq 0$$

- Homework 3 due tonight! Milestones graded.
- HW 4 will be posted tonight. Start early.

$$\mathbb{E}[Y|X=x] \stackrel{\operatorname{argmin}_f}{=} \mathbb{E}[(Y-f(x))^2] \quad \begin{bmatrix} x \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right) \quad \leftarrow |x|$$

$\begin{matrix} dx/d \\ \downarrow \\ \Sigma_{xx} \end{matrix}$ 
 $\begin{matrix} dx/d \\ \downarrow \\ \Sigma_{yy} \end{matrix}$

If  $(x, Y)$  Gaussian it turns out that  $\mathbb{E}[Y|X=x] = w^T x + b$  for  $w \in \mathbb{R}^d$ ,  $b \in \mathbb{R}$

Find  $w, b$ .  $\mathbb{E}[Y|X=x] = \mu_y + \Sigma_{yx}^T \Sigma_{xx}^{-1} (x - \mu_x)$

$$\min_{w, b} \mathbb{E}[(Y - w^T x - b)^2] = \min_w \mathbb{E}[\underbrace{(Y - \mu_y - w^T (x - \mu_x))^2}_{l(w)}]$$

$$\begin{aligned} \nabla_w l(w) &= \nabla_w \mathbb{E}[-2w^T (x - \mu_x)(Y - \mu_y) + w^T (x - \mu_x)(x - \mu_x)w] \leftarrow \\ &= \mathbb{E}[-2(x - \mu_x)(Y - \mu_y) + 2(x - \mu_x)(x - \mu_x)w] \\ &= -2 \Sigma_{yx} + 2 \Sigma_{xx} w = 0 \quad w = \Sigma_{xx}^{-1} \Sigma_{yx} \end{aligned}$$



# Clustering

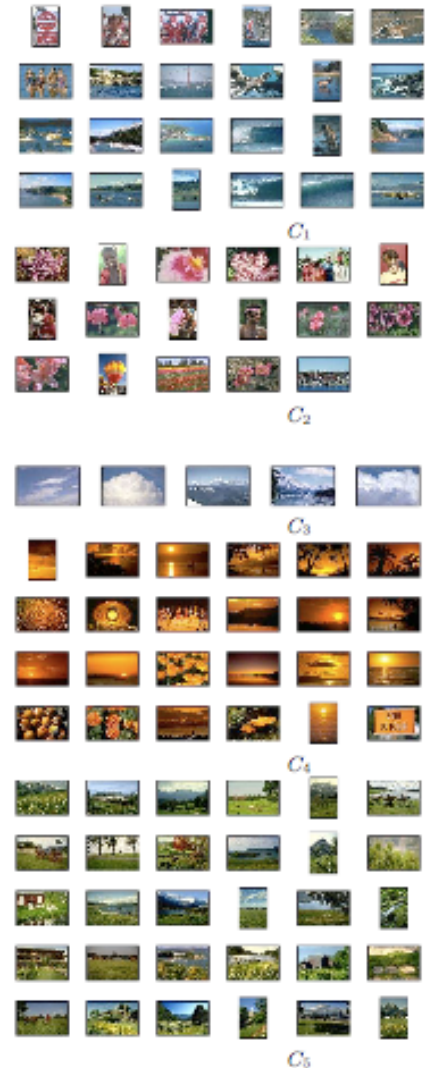
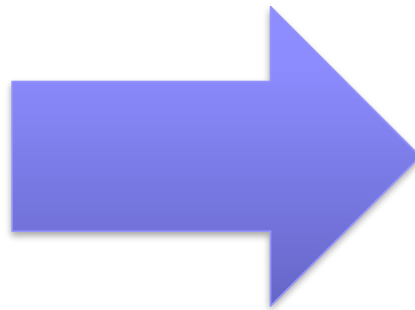
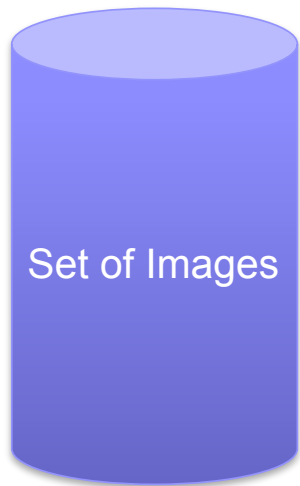
Machine Learning – CSE546

Kevin Jamieson

University of Washington

November 21, 2016

# Clustering images



# Clustering web search results

The screenshot shows the Clusty search engine interface. At the top, there is a navigation bar with links for 'web', 'news', 'images', 'wikipedia', 'blogs', 'jobs', and 'more'. Below this is a search bar containing the word 'race' and a 'Search' button. To the right of the search bar are links for 'advanced preferences'. On the left side, there is a sidebar with a 'clusters' tab selected, showing a list of clusters for 'race'. The main content area displays a list of search results, each with a title, a brief description, and a URL. The results are numbered 1 through 7.

web news images wikipedia blogs jobs more »

Clusty

race Search advanced preferences

clusters sources sites

All Results (238) remix

- Car (20)
- Race cars (7)
- Photos, Races Scheduled (10)
- Game (4)
- Track (3)
- Nascar (2)
- Equipment And Safety (2)
- Other Topics (7)
- Photos (22)
- Game (14)
- Definition (13)
- Team (18)
- Human (8)**
  - Classification Of Human (2)
  - Statement, Evolved (2)
  - Other Topics (4)
- Weekend (8)
- Ethnicity And Race (7)
- Race for the Cure (8)
- Race Information (8)

more | all clusters

find in clusters:  Find

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 **races**, 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/en/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/0813340861](#) - [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, chronicling 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

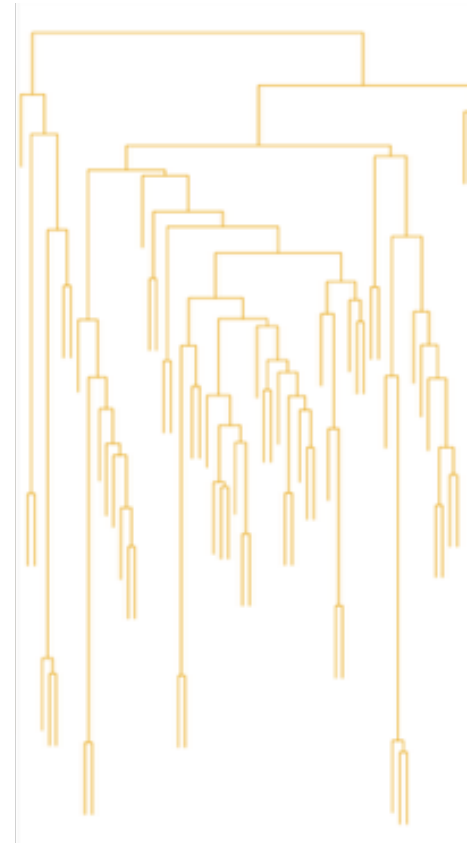
# Hierarchical Clustering

Pick one:

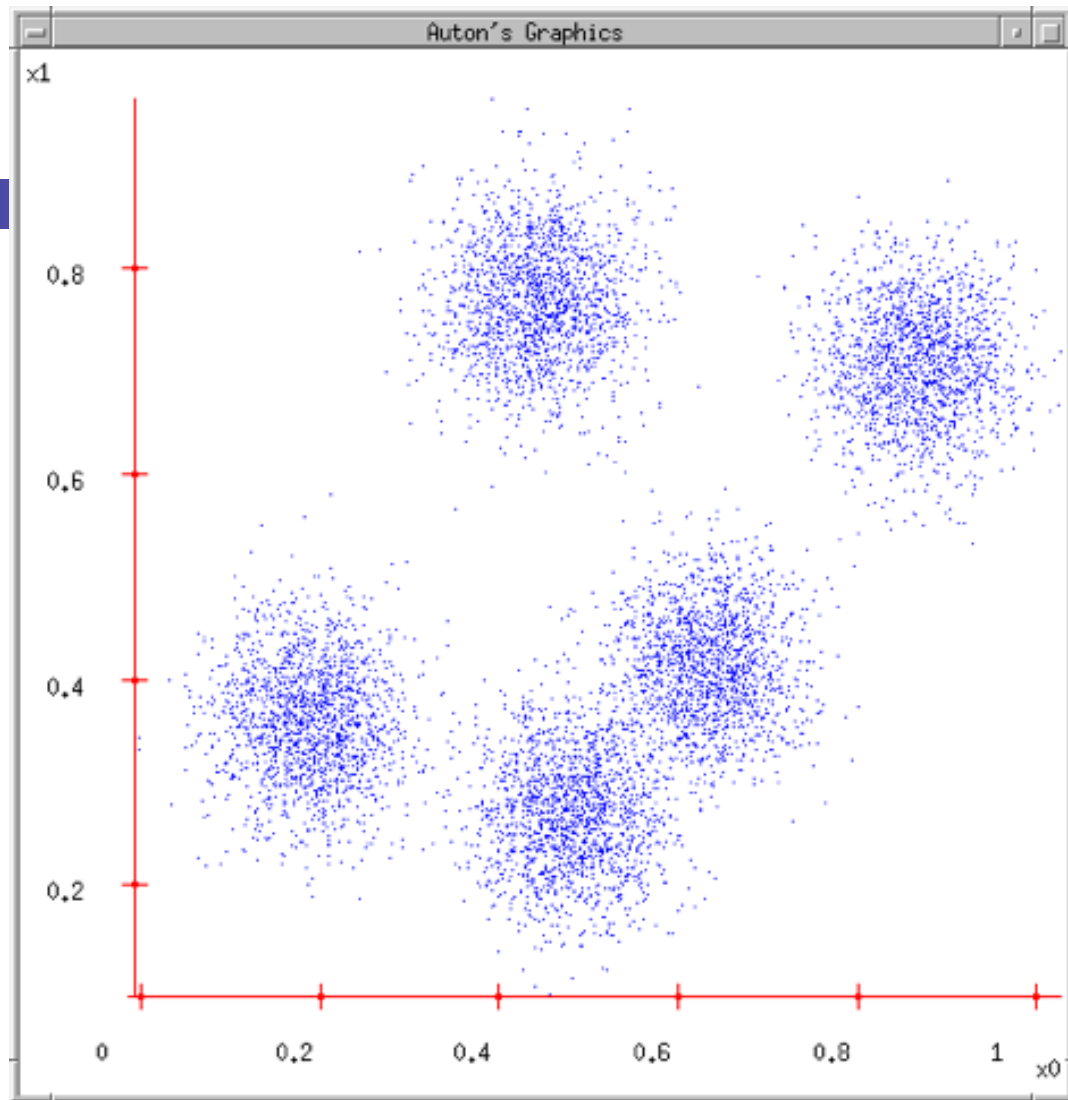
- Bottom up: start with every point as a cluster and merge
- Top down: start with a single cluster containing all points and split

Different rules for splitting/merging, no “right answer”

Gives apparently interpretable tree representation.  
However, warning: even random data with no structure will produce a tree that “appears” to be structured.

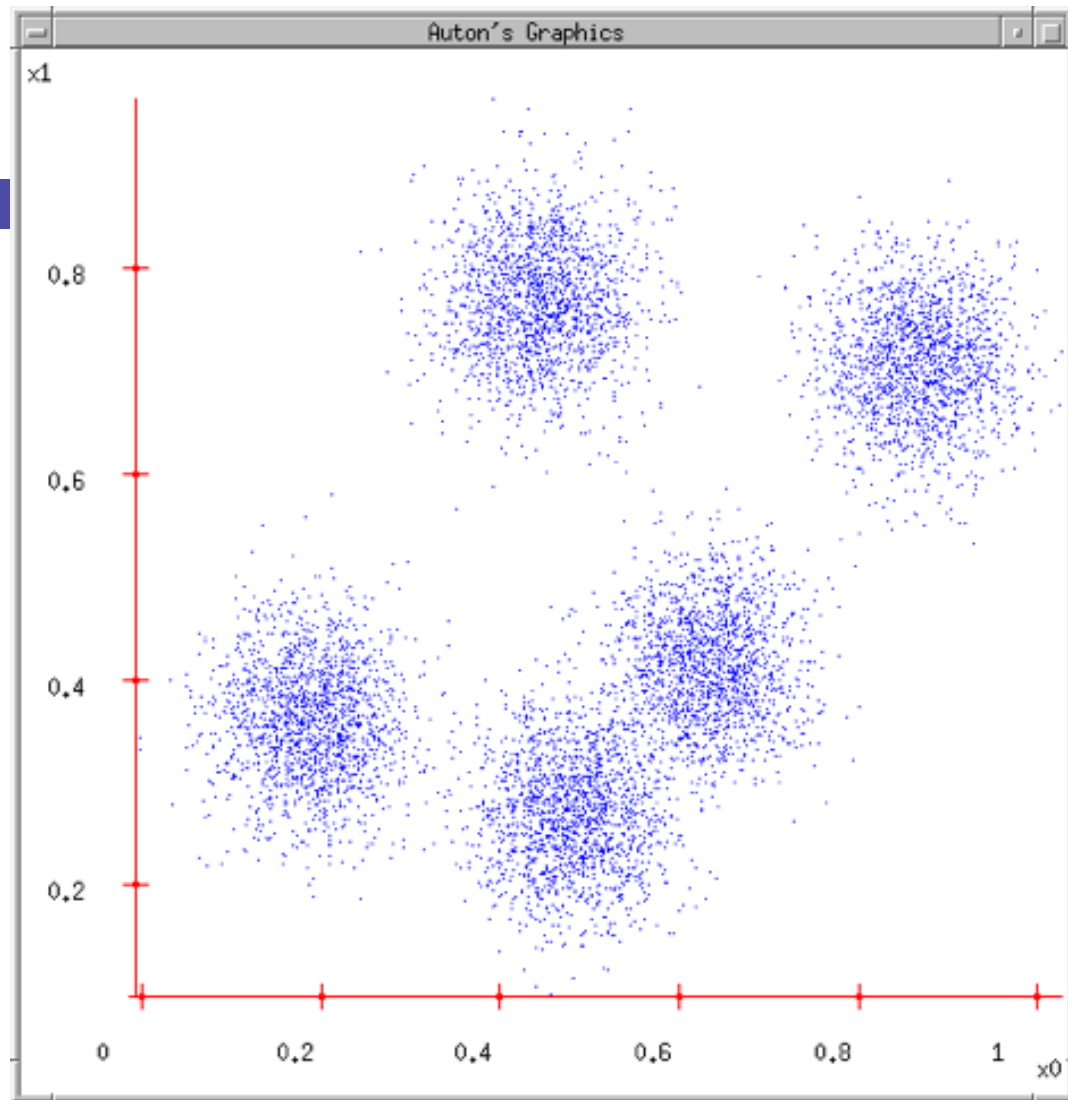


# Some Data



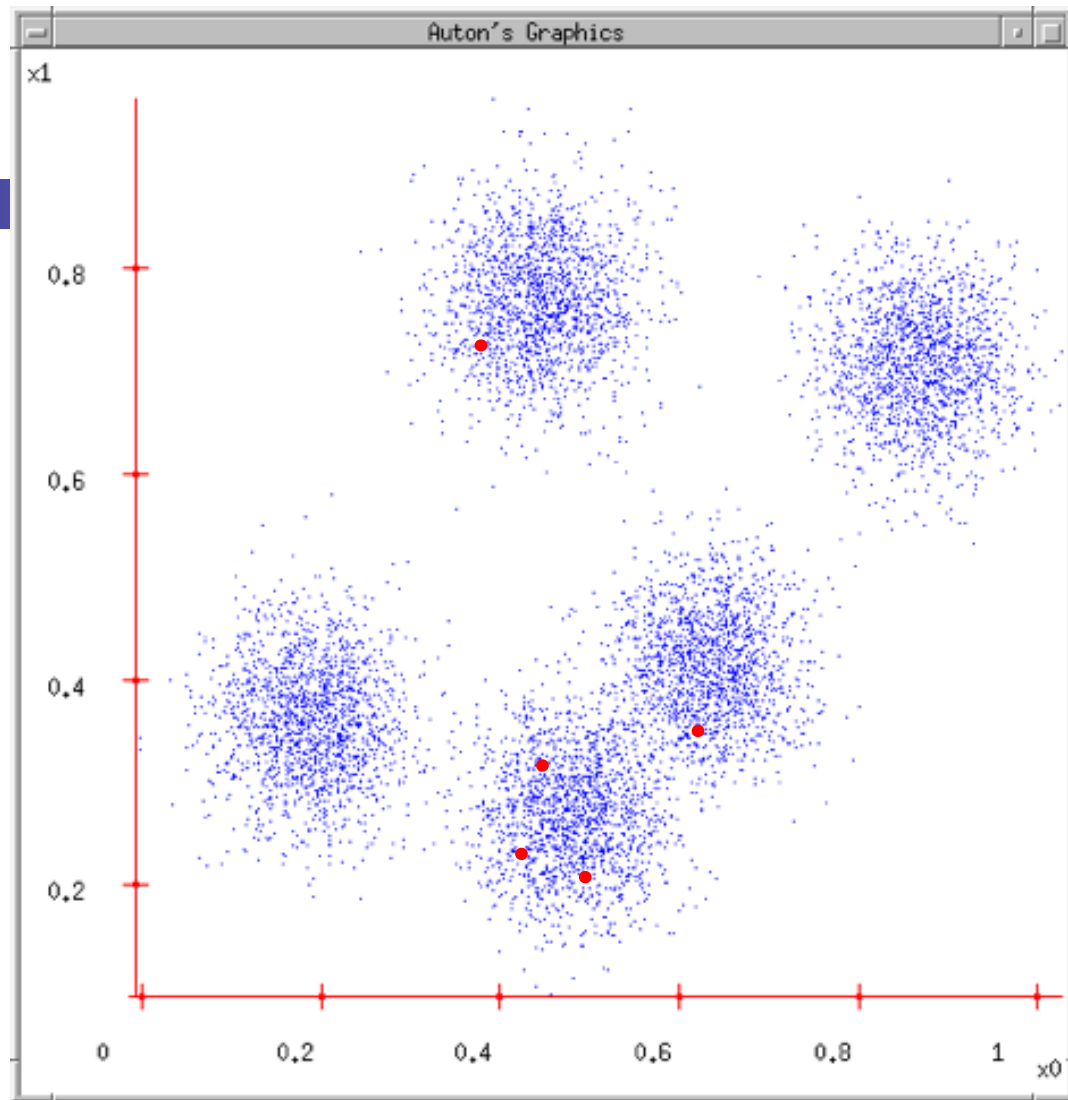
# K-means

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



# K-means

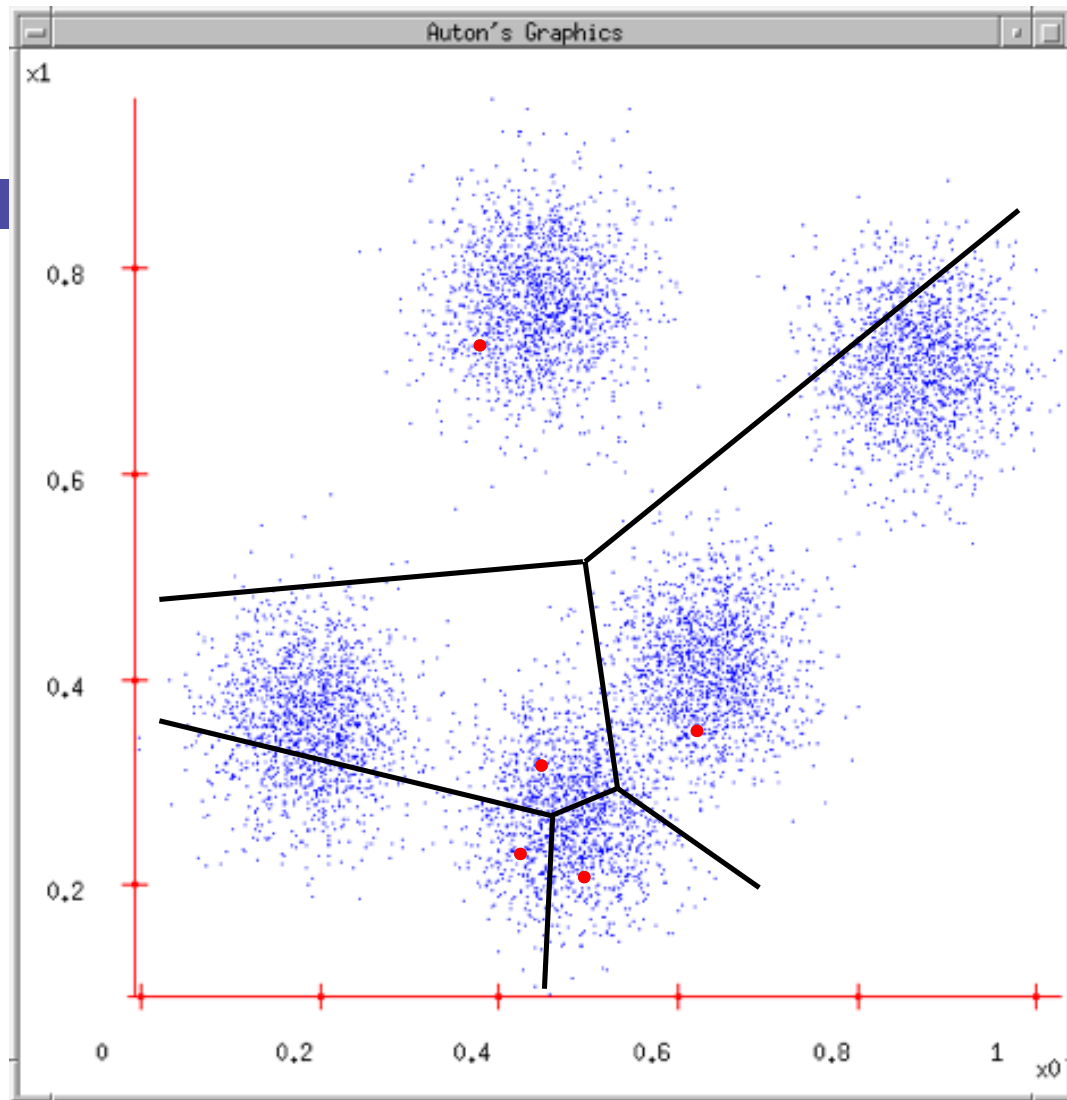
1. Ask user how many clusters they'd like.  
(e.g.  $k=5$ )
2. Randomly guess  $k$  cluster Center locations





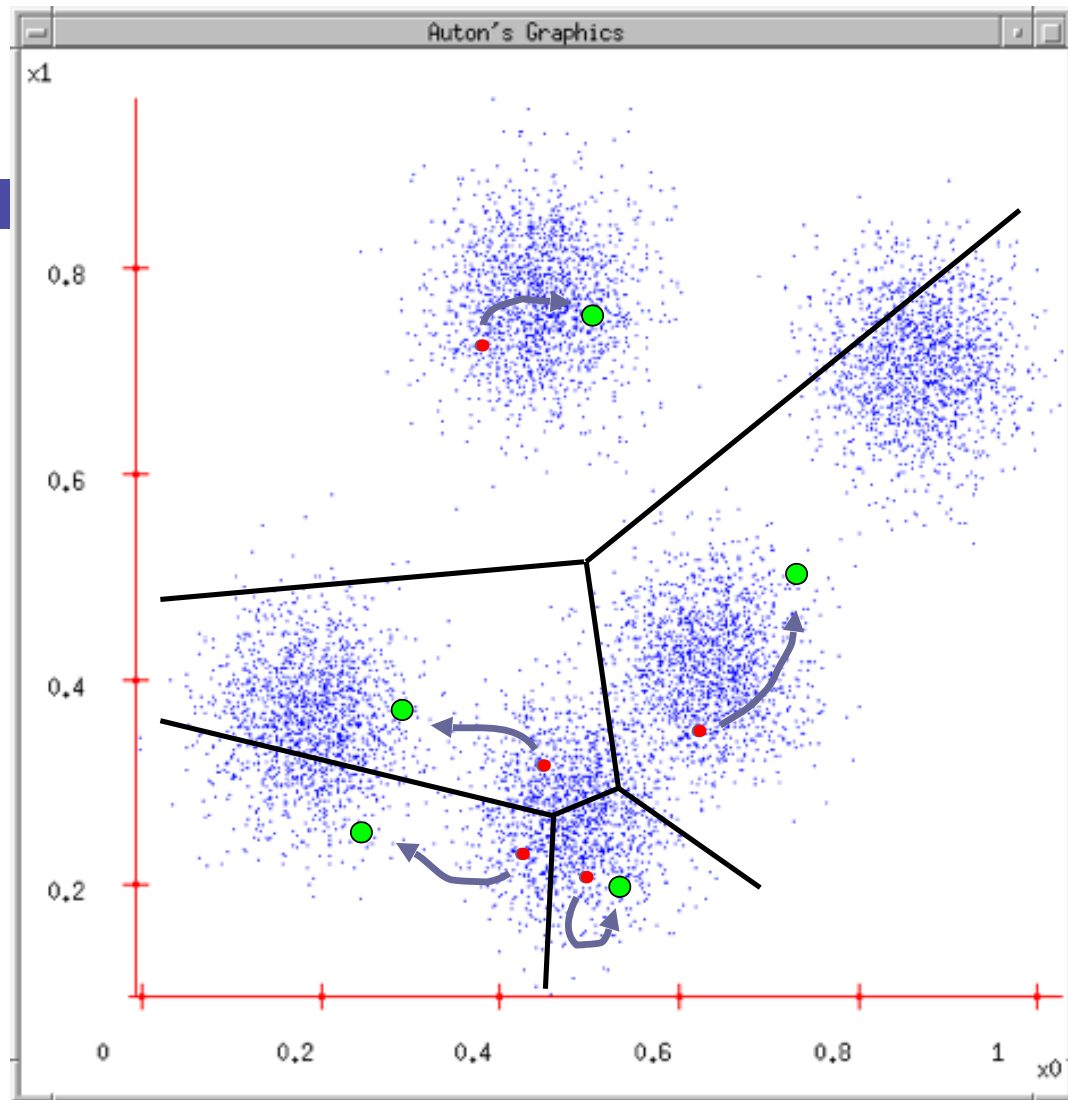
# K-means

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)



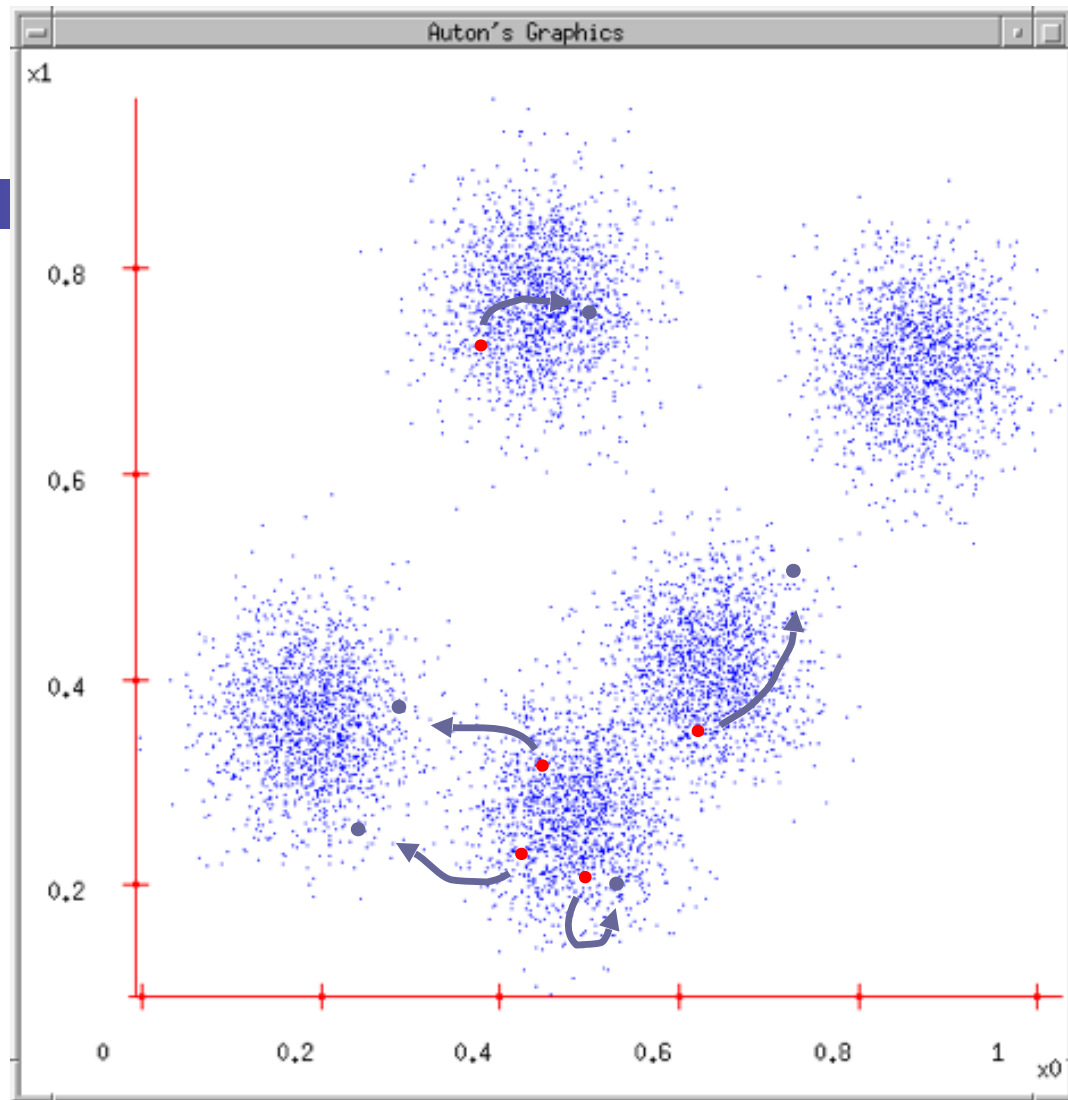
# K-means

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



# 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!



# K-means

- Randomly initialize  $k$  centers
  - $\mu^{(0)} = \mu_1^{(0)}, \dots, \mu_k^{(0)}$
- **Classify:** Assign each point  $j \in \{1, \dots, N\}$  to nearest center:
  - $C^{(t)}(j) \leftarrow \arg \min_i \|\mu_i - x_j\|^2$
- **Recenter:**  $\mu_i$  becomes centroid of its point:
  - $\mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j: C(j)=i} \|\mu - x_j\|^2$
  - Equivalent to  $\mu_i \leftarrow$  average of its points!

# What is K-means optimizing?

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

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

- Optimal K-means:
  - $\min_{\mu} \min_C F(\mu, C)$

# 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  $\mu$ , optimize C

# 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  $\mu$

# Vector Quantization, Fisher Vectors

## Vector Quantization (for compression)

1. Represent image as grid of patches
2. Run k-means on the patches to build code book
3. Represent each patch as a code word.



**FIGURE 14.9.** Sir Ronald A. Fisher (1890 – 1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a  $1024 \times 1024$  grayscale image at 8 bits per pixel. The center image is the result of  $2 \times 2$  block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel



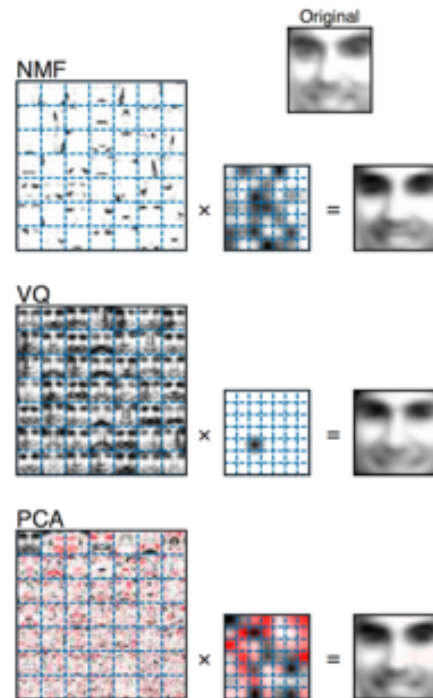
# Vector Quantization, Fisher Vectors

## Vector Quantization (for compression)

1. Represent image as grid of patches
2. Run k-means on the patches to build code book
3. Represent each patch as a code word.



**FIGURE 14.9.** Sir Ronald A. Fisher (1890 – 1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a  $1024 \times 1024$  grayscale image at 8 bits per pixel. The center image is the result of  $2 \times 2$  block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel



# Vector Quantization, Fisher Vectors

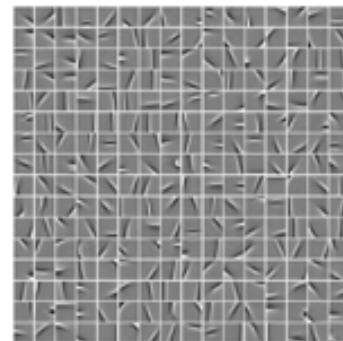
## Vector Quantization (for compression)

1. Represent image as grid of patches
2. Run k-means on the patches to build code book
3. Represent each patch as a code word.



**FIGURE 14.9.** Sir Ronald A. Fisher (1890 – 1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a  $1024 \times 1024$  grayscale image at 8 bits per pixel. The center image is the result of  $2 \times 2$  block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel

Typical output of k-means  
on patches



Similar reduced representation can be used as a feature vector

Coates, Ng, *Learning Feature Representations with K-means*, 2012

# Spectral Clustering

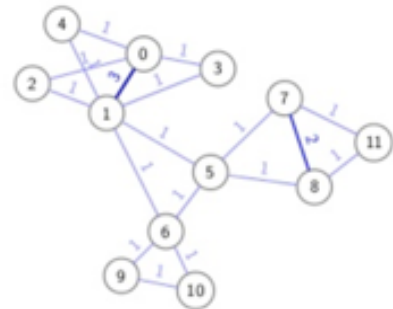
Adjacency matrix:  $\mathbf{W}$

$\mathbf{W}_{i,j}$  = weight of edge  $(i, j)$

$$\mathbf{D}_{i,i} = \sum_{j=1}^n \mathbf{W}_{i,j} \quad \mathbf{L} = \mathbf{D} - \mathbf{W}$$

Given feature vectors, could construct:

- k-nearest neighbor graph with weights in  $\{0, 1\}$
- weighted graph with arbitrary *similarities*  $\mathbf{W}_{i,j} = e^{-\gamma \|x_i - x_j\|^2}$



Let  $f \in \mathbb{R}^n$  be a function over the nodes

$$\begin{aligned} \mathbf{f}^T \mathbf{L} \mathbf{f} &= \sum_{i=1}^N g_i f_i^2 - \sum_{i=1}^N \sum_{i'=1}^N f_i f_{i'} w_{ii'} \\ &= \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N w_{ii'} (f_i - f_{i'})^2. \end{aligned}$$

# Spectral Clustering

Adjacency matrix:  $\mathbf{W}$

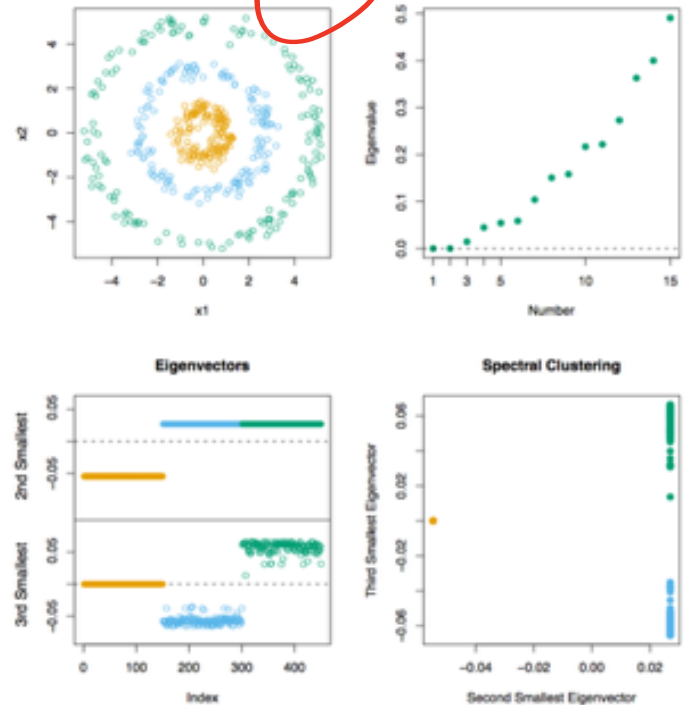
$\mathbf{W}_{i,j}$  = weight of edge  $(i, j)$

$$\mathbf{D}_{i,i} = \sum_{j=1}^n \mathbf{W}_{i,j} \quad \mathbf{L} = \mathbf{D} - \mathbf{W}$$

Given feature vectors, could construct:

- (k=10)-nearest neighbor graph with weights in  $\{0,1\}$

Popular to use the Laplacian  $\mathbf{L}$  or its normalized form  $\tilde{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}$  as a regularizer for learning over graphs





# Mixtures of Gaussians

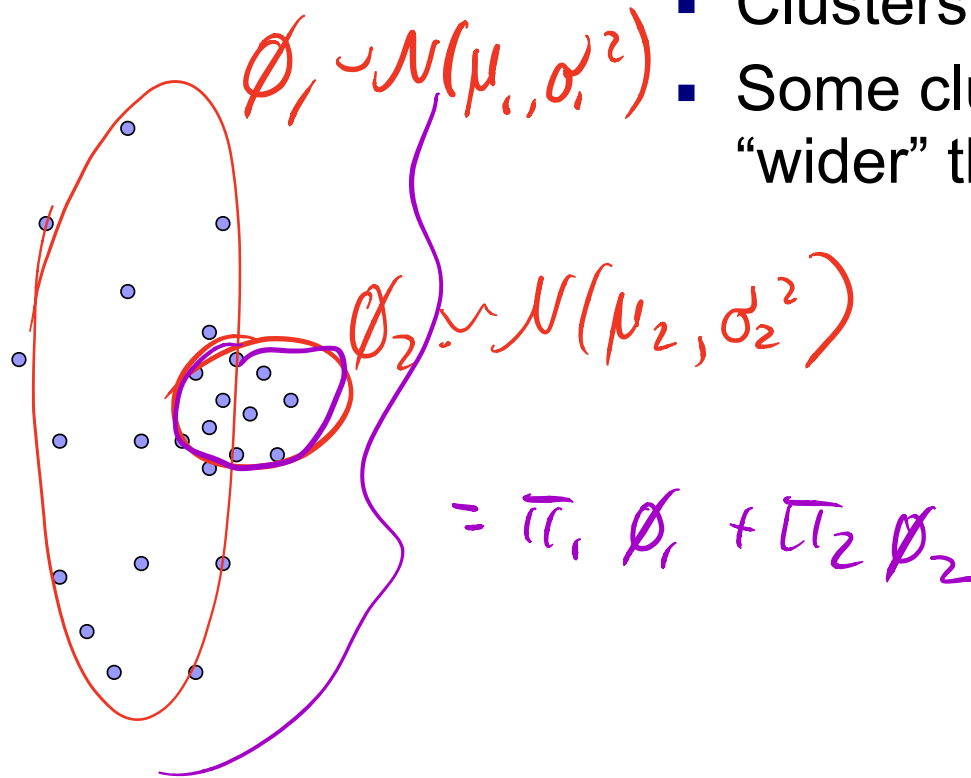
Machine Learning – CSE546

Kevin Jamieson

University of Washington

November 21, 2016

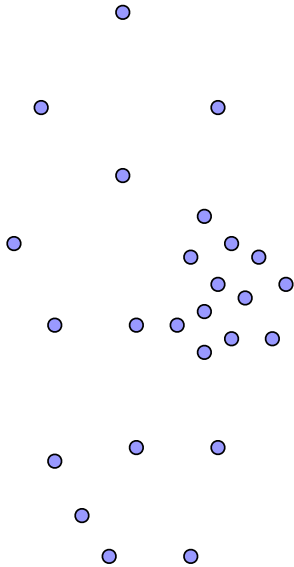
# (One) bad case for k-means



- Clusters may overlap
- Some clusters may be “wider” than others

# (One) bad case for k-means

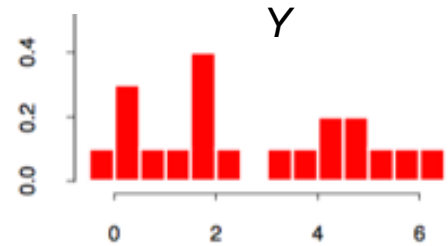
- Clusters may overlap
- Some clusters may be “wider” than others



# Mixture models

$$\begin{aligned} Y_1 &\sim N(\mu_1, \sigma_1^2), & \phi_{\theta_1} &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \\ Y_2 &\sim N(\mu_2, \sigma_2^2), \\ Y &= (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2, \end{aligned}$$

$$\Delta \in \{0, 1\} \text{ with } \Pr(\Delta = 1) = \pi$$



$\mathbf{Z} = \{y_i\}_{i=1}^n$  is observed data

If  $\phi_{\theta}(x)$  is Gaussian density with parameters  $\theta = (\mu, \sigma^2)$  then

$$\ell(\theta; \mathbf{Z}) = \sum_{i=1}^n \log[(1 - \pi)\phi_{\theta_1}(y_i) + \pi\phi_{\theta_2}(y_i)]$$



# Mixture models

$$\ell(\theta | x_i = 1) = \theta \quad \ell(\theta; x) = \theta^{x_i} (1-\theta)^{1-x_i}$$

$$\ell(\theta | x_i = 0) = 1 - \theta$$

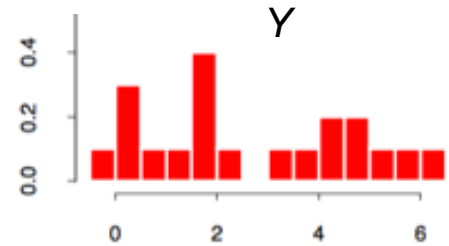
$$Y_1 \sim N(\mu_1, \sigma_1^2),$$

$$Y_2 \sim N(\mu_2, \sigma_2^2),$$

$$Y = (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2,$$

$$\Delta \in \{0, 1\} \text{ with } \Pr(\Delta = 1) = \pi$$

$$\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$



$\mathbf{Z} = \{y_i\}_{i=1}^n$  is observed data

$\mathbf{\Delta} = \{\Delta_i\}_{i=1}^n$  is unobserved data

If  $\phi_\theta(x)$  is Gaussian density with parameters  $\theta = (\mu, \sigma^2)$  then

$$\ell(\theta; y_i, \underline{\Delta_i = 0}) = \log(\phi_{\theta_1}(y_i) (1 - \pi))$$

$$\ell(\theta; y_i, \underline{\Delta_i = 1}) = \log(\phi_{\theta_2}(y_i) \pi)$$

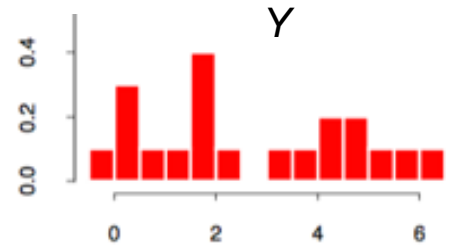
$$\ell(\theta; y_i, \Delta_i) = (1 - \Delta_i) \log((1 - \pi) \phi_{\theta_1}(y_i)) + \Delta_i \log(\pi \phi_{\theta_2}(y_i))$$

# Mixture models

$$\begin{aligned}
 Y_1 &\sim N(\mu_1, \sigma_1^2), \\
 Y_2 &\sim N(\mu_2, \sigma_2^2), \\
 Y &= (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2,
 \end{aligned}$$

$$\Delta \in \{0, 1\} \text{ with } \Pr(\Delta = 1) = \pi$$

$$\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$



$\mathbf{Z} = \{y_i\}_{i=1}^n$  is observed data

$\Delta = \{\Delta_i\}_{i=1}^n$  is unobserved data

If  $\phi_\theta(x)$  is Gaussian density with parameters  $\theta = (\mu, \sigma^2)$  then

$$\ell(\theta; \mathbf{Z}, \Delta) = \sum_{i=1}^n (1 - \Delta_i) \log[(1 - \pi)\phi_{\theta_1}(y_i)] + \Delta_i \log(\pi\phi_{\theta_2}(y_i))$$

If we knew  $\Delta$ , how would we choose  $\theta$ ?

$$\hat{\mu}_1 = \frac{1}{\sum_{i=1}^n (1 - \Delta_i)} \sum_{i=1}^n (1 - \Delta_i) y_i$$

# Mixture models

$$\begin{aligned}
 Y_1 &\sim N(\mu_1, \sigma_1^2), \\
 Y_2 &\sim N(\mu_2, \sigma_2^2), \\
 Y &= (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2,
 \end{aligned}$$

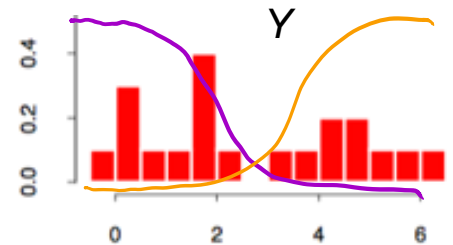
$$\Delta \in \{0, 1\} \text{ with } \Pr(\Delta = 1) = \pi$$

$$\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

If  $\phi_\theta(x)$  is Gaussian density with parameters  $\theta = (\mu, \sigma^2)$  then

$$\ell(\theta; \mathbf{Z}, \Delta) = \sum_{i=1}^n (1 - \Delta_i) \log[(1 - \pi)\phi_{\theta_1}(y_i)] + \Delta_i \log(\pi\phi_{\theta_2}(y_i))$$

If we knew  $\theta$ , how would we choose  $\Delta$ ?  $\mathbb{E}[\Delta_i | \theta, z] = \mathbb{P}(\Delta_i = 1 | \theta, z)$

$$= \frac{\pi \phi_2(y_i)}{(1 - \pi)\phi_1(y_i) + \pi \phi_2(y_i)}$$


$\mathbf{Z} = \{y_i\}_{i=1}^n$  is observed data

$\Delta = \{\Delta_i\}_{i=1}^n$  is unobserved data

# Mixture models

$$\begin{aligned}Y_1 &\sim N(\mu_1, \sigma_1^2), \\Y_2 &\sim N(\mu_2, \sigma_2^2), \\Y &= (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2,\end{aligned}$$

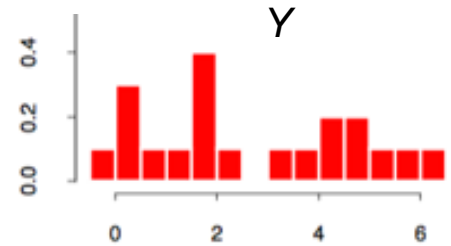
$$\Delta \in \{0, 1\} \text{ with } \Pr(\Delta = 1) = \pi$$

$$\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

If  $\phi_\theta(x)$  is Gaussian density with parameters  $\theta = (\mu, \sigma^2)$  then

$$\ell(\theta; \mathbf{Z}, \mathbf{\Delta}) = \sum_{i=1}^n (1 - \Delta_i) \log[(1 - \pi)\phi_{\theta_1}(y_i)] + \Delta_i \log(\pi\phi_{\theta_2}(y_i))$$

$$\gamma_i(\theta) = \mathbb{E}[\Delta_i | \theta, \mathbf{Z}] =$$



$\mathbf{Z} = \{y_i\}_{i=1}^n$  is observed data

$\mathbf{\Delta} = \{\Delta_i\}_{i=1}^n$  is unobserved data

# Mixture models

---

**Algorithm 8.1** *EM Algorithm for Two-component Gaussian Mixture.*

---

1. Take initial guesses for the parameters  $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$  (see text).
2. *Expectation Step*: compute the responsibilities

$$E[\mathbb{1}_{\{i \in \mathcal{C}_1\}} | \theta, \mathbf{z}] = \hat{\gamma}_i = \frac{\hat{\pi} \phi_{\hat{\theta}_2}(y_i)}{(1 - \hat{\pi}) \phi_{\hat{\theta}_1}(y_i) + \hat{\pi} \phi_{\hat{\theta}_2}(y_i)}, \quad i = 1, 2, \dots, N. \quad (8.42)$$

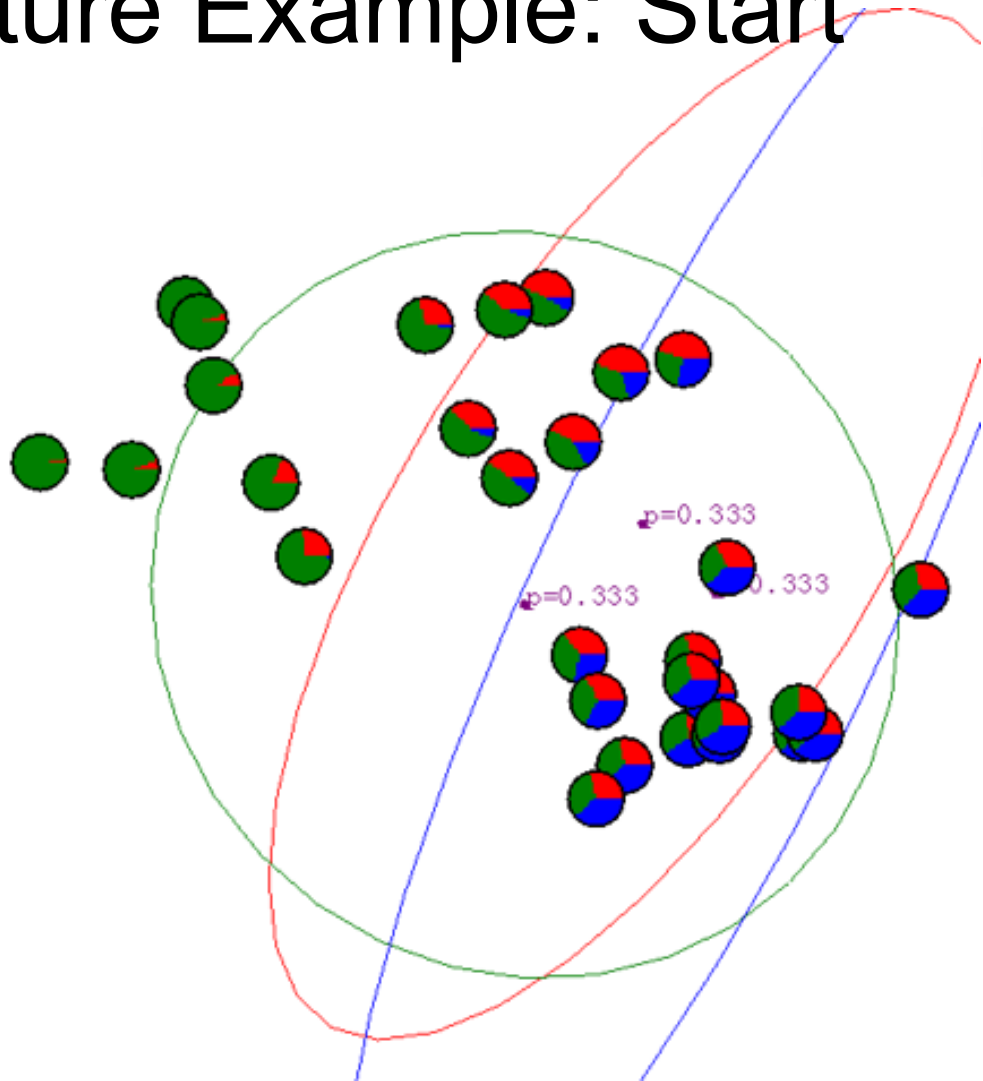
3. *Maximization Step*: compute the weighted means and variances:

$$\begin{aligned} \hat{\mu}_1 &= \frac{\sum_{i=1}^N (1 - \hat{\gamma}_i) y_i}{\sum_{i=1}^N (1 - \hat{\gamma}_i)}, & \hat{\sigma}_1^2 &= \frac{\sum_{i=1}^N (1 - \hat{\gamma}_i) (y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1 - \hat{\gamma}_i)}, \\ \hat{\mu}_2 &= \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, & \hat{\sigma}_2^2 &= \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i}, \end{aligned}$$

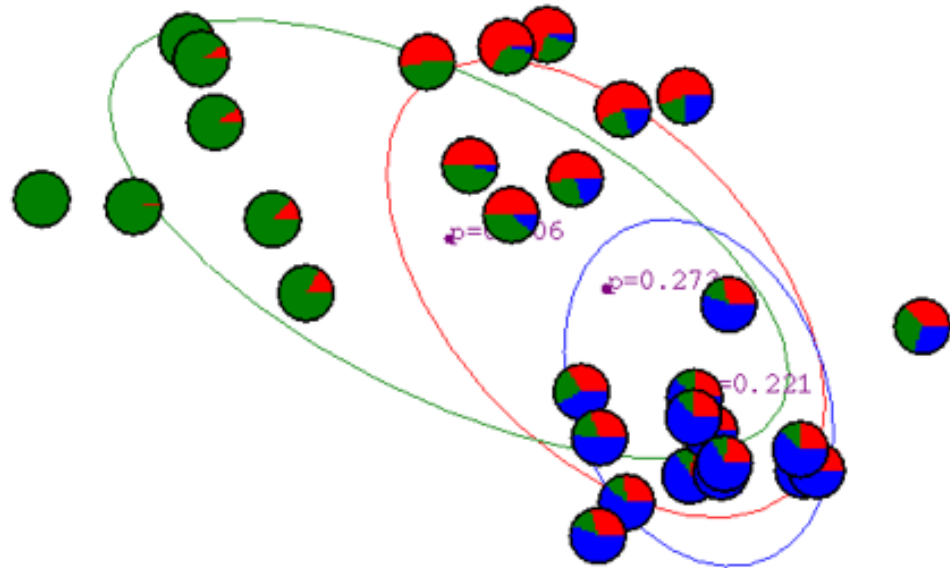
and the mixing probability  $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$ .

4. Iterate steps 2 and 3 until convergence.
-

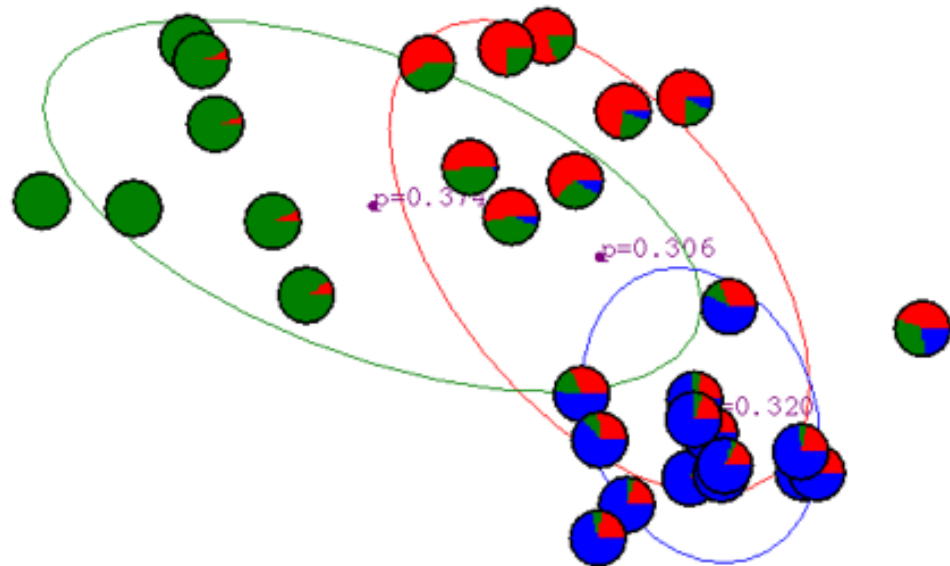
# Gaussian Mixture Example: Start



# After first iteration

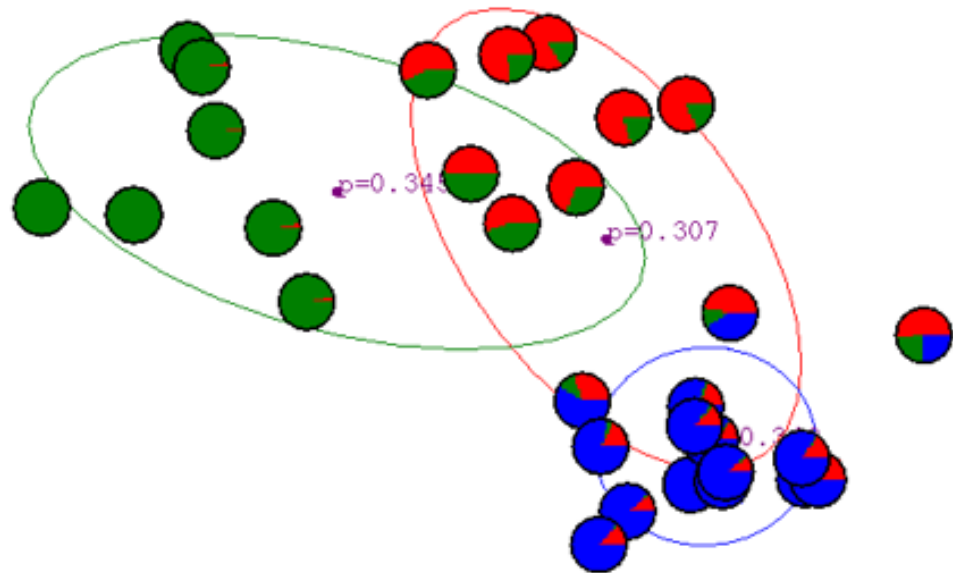


# After 2nd iteration

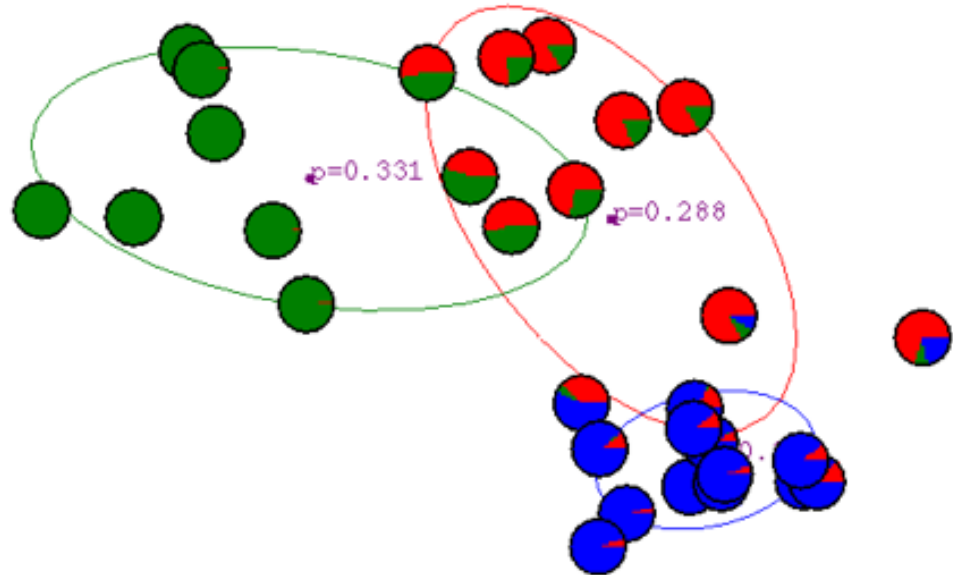




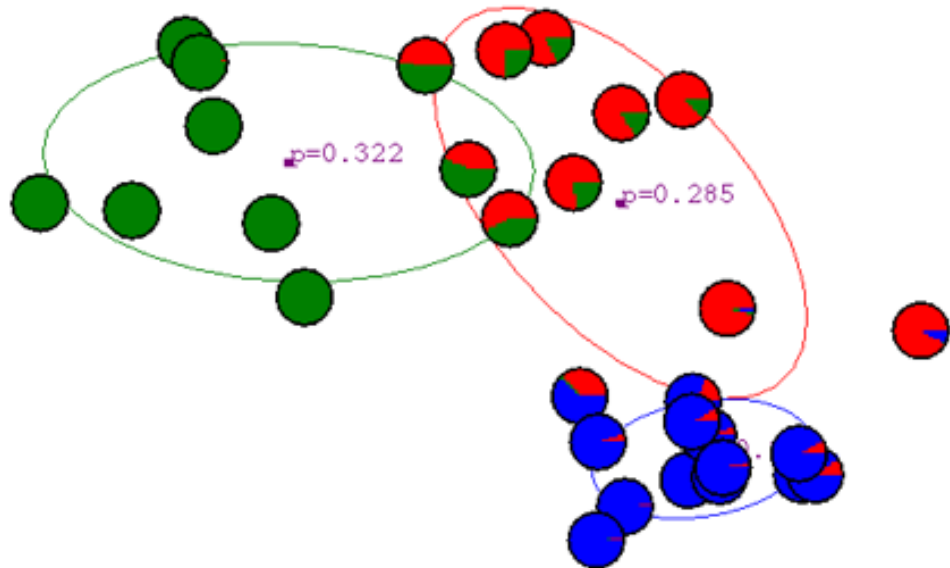
# After 3rd iteration



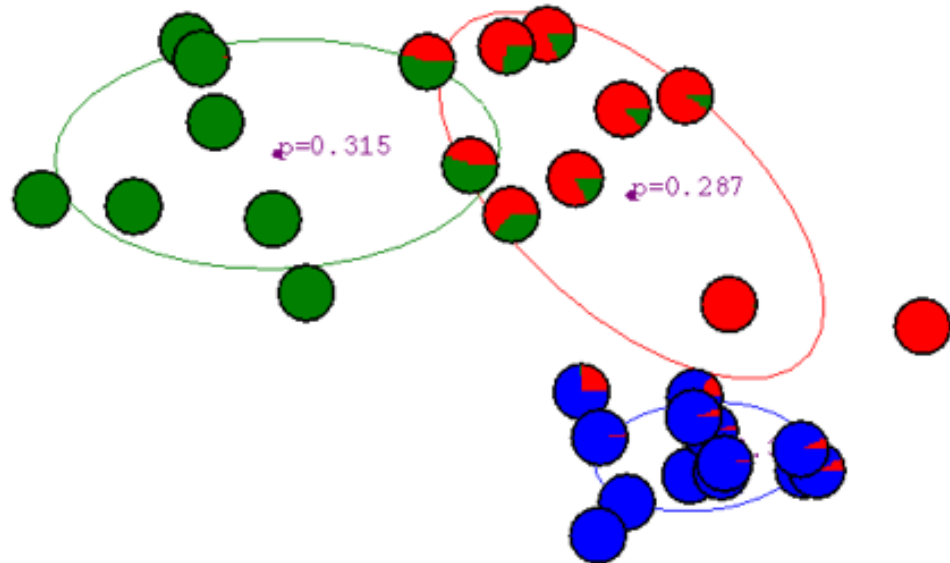
# After 4th iteration



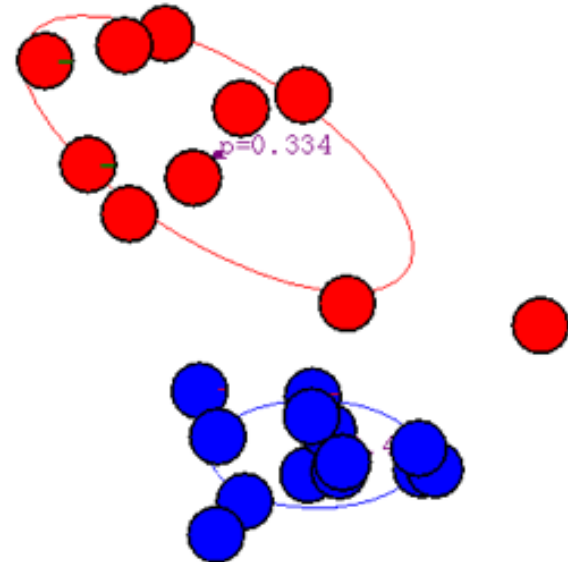
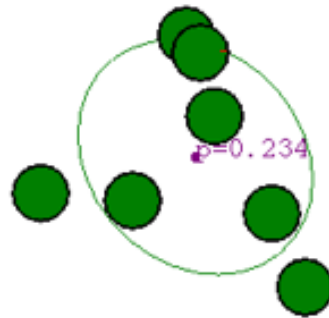
# After 5th iteration



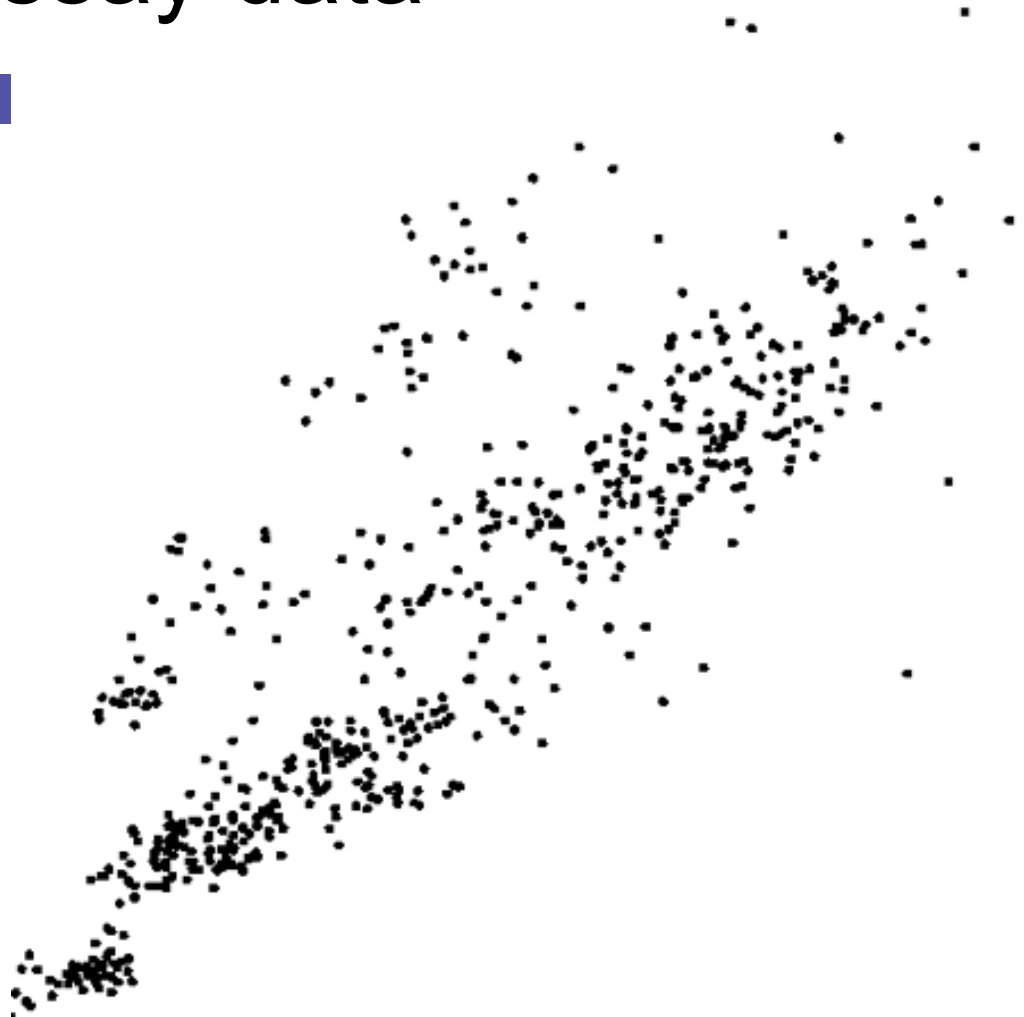
# After 6th iteration



# After 20th iteration



# Some Bio Assay data

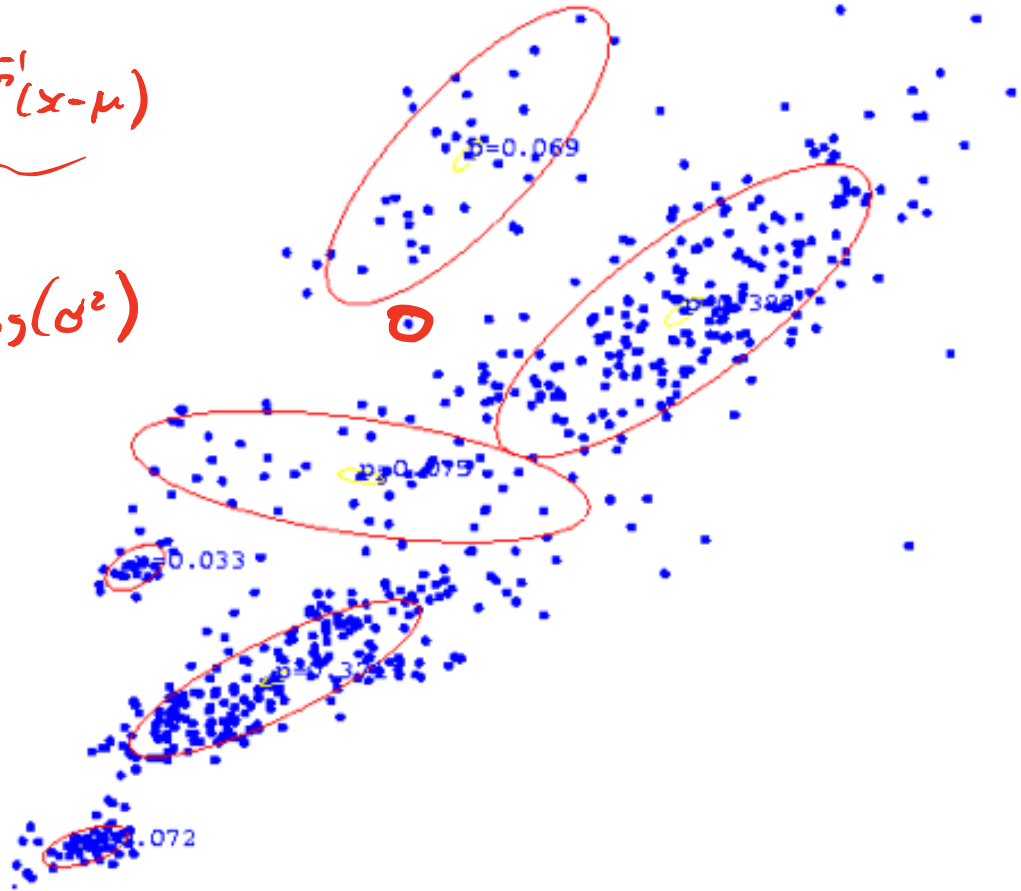


# GMM clustering of the assay data

$$\frac{1}{2} \log(2\pi |\Sigma|) - \frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)$$

↓

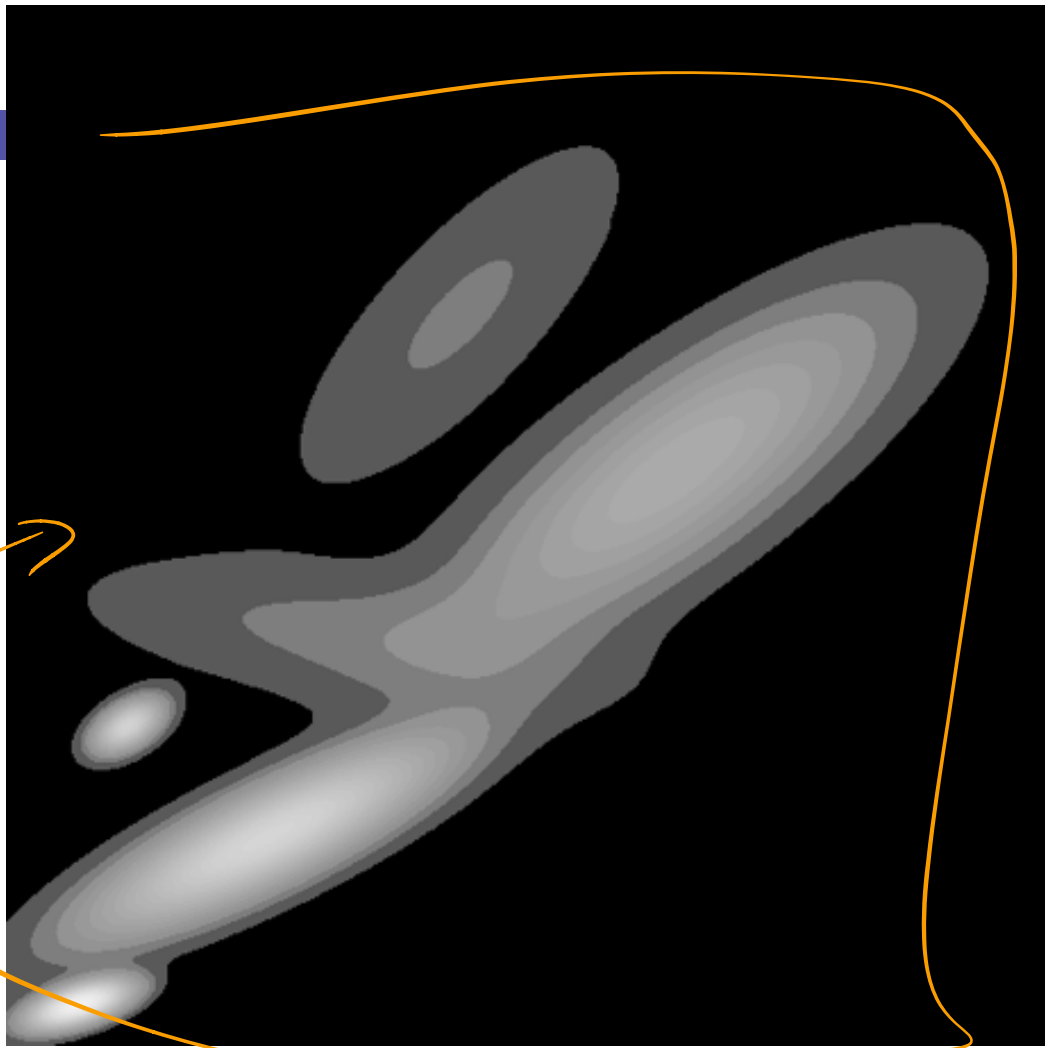
$$\Sigma = \text{diag}(\sigma^2)$$



# Resulting Density Estimator

$$\sum_{i=1}^k \pi_i \phi_{\theta_i}(x)$$

$x$





# Expectation Maximization Algorithm

The iterative gaussian mixture model (GMM) fitting algorithm is special case of EM:

---

## Algorithm 8.2 *The EM Algorithm.*

---

1. Start with initial guesses for the parameters  $\hat{\theta}^{(0)}$ .
2. *Expectation Step*: at the  $j$ th step, compute

*function of  $\theta'$*   $Q(\theta', \hat{\theta}^{(j)}) = E(\ell_0(\theta'; \mathbf{T}) | \mathbf{Z}, \hat{\theta}^{(j)})$  (8.43)

as a function of the dummy argument  $\theta'$ .

3. *Maximization Step*: determine the new estimate  $\hat{\theta}^{(j+1)}$  as the maximizer of  $Q(\theta', \hat{\theta}^{(j)})$  over  $\theta'$ .
  4. Iterate steps 2 and 3 until convergence.
- 

$\mathbf{Z}$  is observed data  
 $\Delta$  is unobserved data  
 $\mathbf{T} = (\mathbf{Z}, \Delta)$

# Missing data example

$x_i \sim \mathcal{N}(\mu, \Sigma)$  but suppose some entries of  $x_i$  are *missing*

$$x_i = \begin{bmatrix} y_i \\ \Delta_i \end{bmatrix}$$

$\mathbf{Z}$  is observed data

$\Delta$  is unobserved data

$$\mathbf{T} = (\mathbf{Z}, \Delta)$$

$$\ell(\theta | \mathbf{T}, \theta) = -\frac{1}{2} \log(2\pi|\Sigma|) + (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

E Step:  $\mathbb{E}[\ell(\theta'; \mathbf{T}) | \mathbf{Z}, \hat{\theta}^{(j)}]$

Natural choice for  $\hat{\theta}^{(0)}$ ?

# Missing data example

$x_i \sim \mathcal{N}(\mu, \Sigma)$  but suppose some entries of  $x_i$  are *missing*

$$\ell(\theta | \mathbf{T}, \theta) = -\frac{1}{2} \log(2\pi|\Sigma|) + (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

$\mathbf{Z}$  is observed data

$\Delta$  is unobserved data

$$\mathbf{T} = (\mathbf{Z}, \Delta)$$

E Step:  $\mathbb{E}[\ell(\theta'; \mathbf{T}) | \mathbf{Z}, \hat{\theta}^{(j)}]$

Natural choice for  $\hat{\theta}^{(0)}$ ?

$$\mathbb{E}[Y | X = x] = \mu_Y + \Sigma_{YX} \Sigma_{XX}^{-1} (x - \mu_X)$$

M Step:  $\hat{\theta}^{(j+1)} = \arg \max_{\theta'} \mathbb{E}[\ell(\theta'; \mathbf{T}) | \mathbf{Z}, \hat{\theta}^{(j)}]$

# Missing data example

$x_i \sim \mathcal{N}(\mu, \Sigma)$  but suppose some entries of  $x_i$  are *missing*

$$\ell(\theta | \mathbf{T}, \theta) = -\frac{1}{2} \log(2\pi|\Sigma|) + (x_i - \mu)^T \Sigma^{-1} (x - \mu)$$

$\mathbf{Z}$  is observed data

$\Delta$  is unobserved data

$$\mathbf{T} = (\mathbf{Z}, \Delta)$$

E Step:  $\mathbb{E}[\ell(\theta'; \mathbf{T}) | \mathbf{Z}, \hat{\theta}^{(j)}]$

Natural choice for  $\hat{\theta}^{(0)}$ ?

$$\mathbb{E}[Y | X = x] = \mu_Y + \Sigma_{YX} \Sigma_{XX}^{-1} (x - \mu_X)$$

M Step:  $\hat{\theta}^{(j+1)} = \arg \max_{\theta'} \mathbb{E}[\ell(\theta'; \mathbf{T}) | \mathbf{Z}, \hat{\theta}^{(j)}]$

Connection to matrix factorization?



# Density Estimation

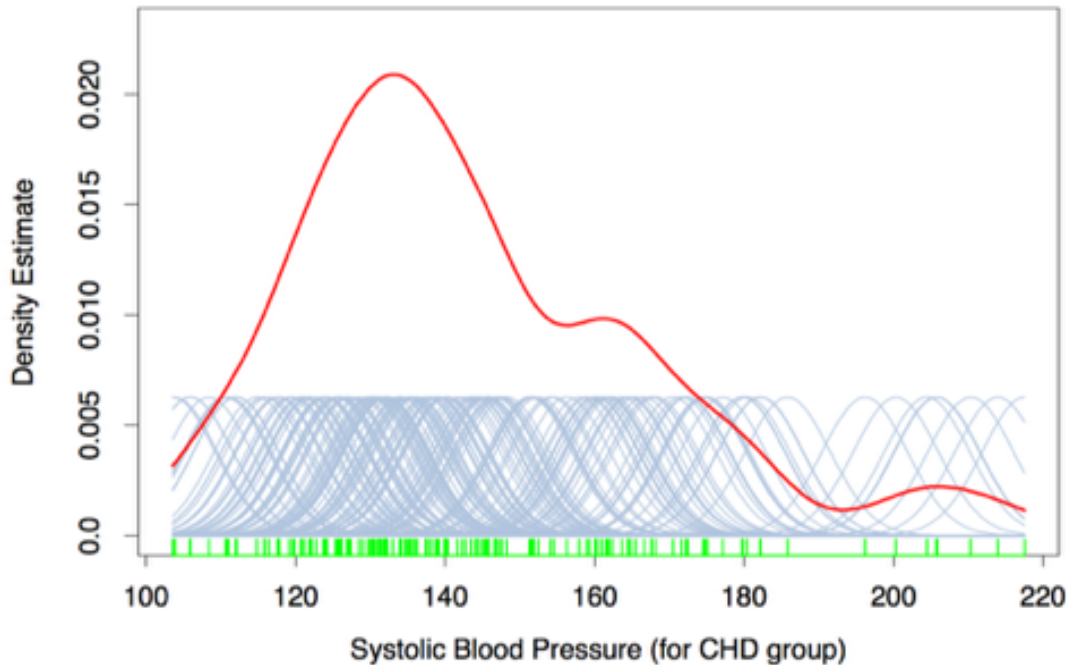
Machine Learning – CSE546

Kevin Jamieson

University of Washington

November 21, 2016

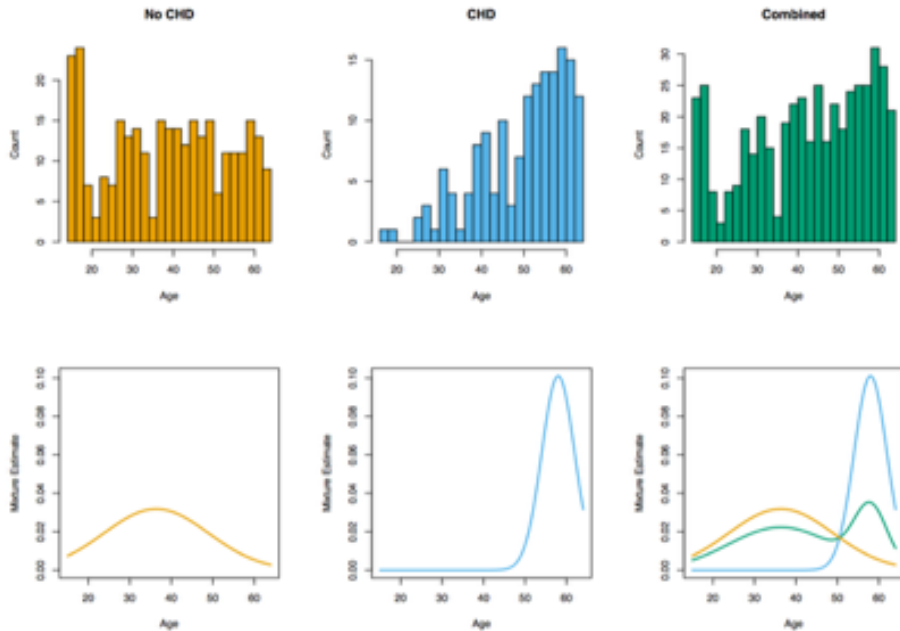
# Kernel Density Estimation



$$f(x) = \sum_{m=1}^M \alpha_m \phi(x; \mu_m, \Sigma_m)$$

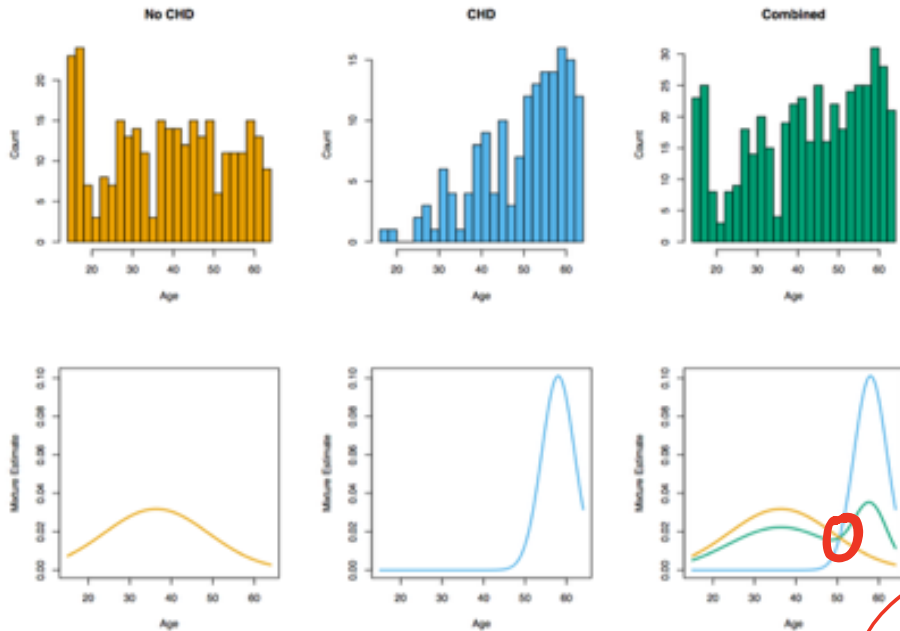
A very “lazy” GMM

# Kernel Density Estimation



$$f(x) = \sum_{m=1}^M \alpha_m \phi(x; \mu_m, \Sigma_m)$$

# Kernel Density Estimation



What is the Bayes optimal classification rule?

$$f(x) = \sum_{m=1}^M \alpha_m \phi(x; \mu_m, \Sigma_m)$$

$$\hat{r}_{im} = \frac{\hat{\alpha}_m \phi(x_i; \hat{\mu}_m, \hat{\Sigma}_m)}{\sum_{k=1}^M \hat{\alpha}_k \phi(x_i; \hat{\mu}_k, \hat{\Sigma}_k)}$$

Predict  $\arg \max_m \hat{r}_{im}$



# Generative vs Discriminative

$X \sim \text{class 1}$

$Y \sim \text{class 2}$

Generative rule: Fits densities to  $X$  and  $Y$   
and then treats fitted densities  
as truth, and applies optimal classification

Discriminative: Makes no effort to fit densities,  
only fits function to decision boundary  
 $\{z: P(X=z) = P(Y=z)\}$