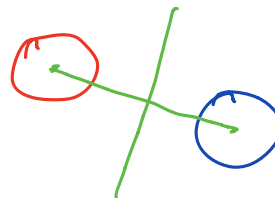


Practice

Milestone
- Appendix

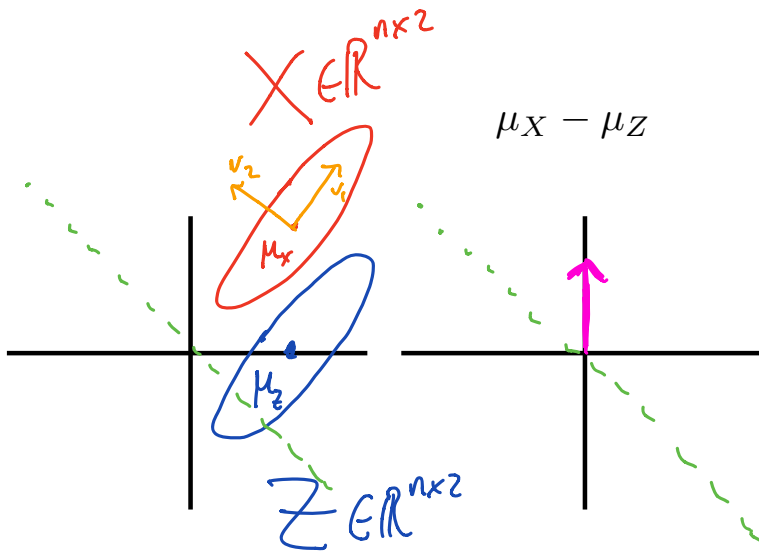


- Fill in the missing plots:

$$\Sigma = \mathbf{X}^T \mathbf{J} \mathbf{J} \mathbf{X} = \mathbf{Z}^T \mathbf{J} \mathbf{J} \mathbf{Z}$$

$$\mathbf{V} \mathbf{S} \mathbf{V}^T = \text{eig}(\Sigma) \quad \mathbf{J} = \mathbf{I} - \mathbf{1} \mathbf{1}^T / n$$

$$\mu_X = \mathbf{X}^T \mathbf{1} / n \quad \mu_Z = \mathbf{Z}^T \mathbf{1} / n$$



$$\begin{aligned} \bar{\mathbf{J}} \mathbf{X} &= \mathbf{U} \mathbf{S}^{1/2} \mathbf{V}^T \\ (\bar{\mathbf{J}} \mathbf{X})^T (\bar{\mathbf{J}} \mathbf{X}) &= \mathbf{X}^T \bar{\mathbf{J}} \bar{\mathbf{J}}^T \mathbf{X} \\ &= \mathbf{V} \mathbf{S}^{1/2} \mathbf{S}^{1/2} \mathbf{V}^T \\ &= \mathbf{V} \mathbf{S} \mathbf{V}^T \end{aligned}$$

$$\underbrace{\mathbf{V} \mathbf{S}^{-1/2} \mathbf{V}^T}_{\mathbf{W} \text{ } d \times n} (\bar{\mathbf{J}} \mathbf{X})^T = \mathbf{V} \mathbf{S}^{-1/2} \mathbf{V}^T \mathbf{V} \mathbf{S}^{1/2} \mathbf{U}^T = \mathbf{V} \mathbf{U}^T$$

$$WW^T = VU^TUV^T = VV^T$$

Matrix Completion

Machine Learning – CSE546

Kevin Jamieson

University of Washington

November 15, 2016

Singular Value Decomposition (SVD)

Theorem (SVD): Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $r \leq \min\{m, n\}$. Then $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where $\mathbf{S} \in \mathbb{R}^{r \times r}$ is diagonal with positive entries, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

$$\mathbf{U} = [u_1, \dots, u_r] \quad \mathbf{V} = [v_1, \dots, v_r]$$

$$\mathbf{A}^T \mathbf{A} v_i = \mathbf{S}_{i,i}^2 v_i$$

$$\mathbf{A} \mathbf{A}^T u_i = \mathbf{S}_{i,i}^2 u_i$$

\mathbf{V} are the first r eigenvectors of $\mathbf{A}^T \mathbf{A}$ with eigenvalues $\text{diag}(\mathbf{S})$
 \mathbf{U} are the first r eigenvectors of $\mathbf{A} \mathbf{A}^T$ with eigenvalues $\text{diag}(\mathbf{S})$

Singular Value Decomposition (SVD)

Theorem (SVD): Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $r \leq \min\{m, n\}$. Then $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where $\mathbf{S} \in \mathbb{R}^{r \times r}$ is diagonal with positive entries, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

$$\mathbf{U} = [u_1, \dots, u_r] \quad \mathbf{V} = [v_1, \dots, v_r] \quad \mathbf{S} = \text{diag}(s_1, \dots, s_r)$$

$$\mathbf{A} = \sum_{k=1}^r u_k v_k^T s_k$$

$$s_1 \geq s_2 \geq \dots \geq s_r$$

Best rank-1 approximation $\sigma > 0$ and unit vectors $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ minimizes:
 $\|\sigma xy^T - \mathbf{A}\|_F^2 =$

Singular Value Decomposition (SVD)

Theorem (SVD): Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $r \leq \min\{m, n\}$. Then $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where $\mathbf{S} \in \mathbb{R}^{r \times r}$ is diagonal with positive entries, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

$$\mathbf{U} = [u_1, \dots, u_r] \quad \mathbf{V} = [v_1, \dots, v_r] \quad \mathbf{S} = \text{diag}(s_1, \dots, s_r)$$

$$\mathbf{A} = \sum_{k=1}^r u_k v_k^T s_k \quad s_1 \geq s_2 \geq \dots \geq s_r$$

Best rank-1 approximation $\sigma > 0$ and unit vectors $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ minimizes:

$$\begin{aligned} \|\sigma xy^T - \mathbf{A}\|_F^2 &= \sigma^2 + \text{Tr}(\mathbf{A}^T \mathbf{A}) - 2\sigma x^T \mathbf{A} y \\ &= \sigma^2 + \left(\sum_{k=1}^r s_k^2 \right) - 2\sigma \left(\sum_{k=1}^r x^T u_k v_k^T y s_k \right) \end{aligned}$$

Singular Value Decomposition (SVD)

Theorem (SVD): Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $r \leq \min\{m, n\}$. Then $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where $\mathbf{S} \in \mathbb{R}^{r \times r}$ is diagonal with positive entries, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

$$\mathbf{U} = [u_1, \dots, u_r] \quad \mathbf{V} = [v_1, \dots, v_r] \quad \mathbf{S} = \text{diag}(s_1, \dots, s_r)$$

$$\mathbf{A} = \sum_{k=1}^r u_k v_k^T s_k$$

$$s_1 \geq s_2 \geq \dots \geq s_r$$

In general:

$$\sum_{k=1}^p u_k v_k^T s_k = \arg \min_{\mathbf{Z}: \text{rank}(\mathbf{Z})=p} \|\mathbf{Z} - \mathbf{A}\|_F^2$$

Matrix completion

Given historical data on how users rated movies in past:

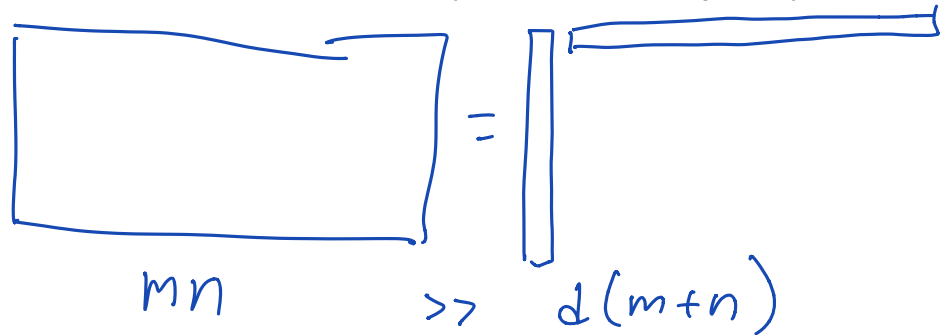
17,700 movies, 480,189 users, 99,072,112 ratings

NETFLIX

(Sparsity: 1.2%)

Predict how the same users will rate movies in the future (for \$1 million prize)

						...
Alice	1	?	?	4	?	
Bob	?	2	5	?	?	
Carol	?	?	4	5	?	
Dave	5	?	?	?	4	
⋮						



$.0012 mn \gg d(m+n) \Rightarrow$ learning is possible

Matrix completion

n movies, m users, |S| ratings

$$\arg \min_{U \in \mathbb{R}^{m \times d}, V \in \mathbb{R}^{n \times d}} \sum_{(i,j,s) \in \mathcal{S}} \|(UV^T)_{i,j} - s_{i,j}\|_2^2$$

How do we solve it? With full information?

Matrix completion

n movies, m users, |S| ratings

$$\arg \min_{U \in \mathbb{R}^{m \times d}, V \in \mathbb{R}^{n \times d}} \sum_{(i,j,s) \in \mathcal{S}} \|(UV^T)_{i,j} - s_{i,j}\|_2^2$$

user i gets assigned $u_i \in \mathbb{R}^d$
 movie j " " $v_j \in \mathbb{R}^d$

Predict user i will rate movie j as $u_i^T v_j = (UV^T)_{i,j}$

$$\begin{aligned} \nabla_{u_i} \sum_{j: (i,j) \in \mathcal{S}} \|(u_i^T v_j - s_{i,j})\|^2 &= \sum_{j: (i,j) \in \mathcal{S}} 2 v_j (u_i^T v_j - s_{i,j}) \\ &\Rightarrow \left(\sum v_j v_j^T \right) u_i = 2 \sum v_j s_{i,j} \end{aligned}$$

Matrix completion

n movies, m users, |S| ratings

$$\arg \min_{U \in \mathbb{R}^{m \times d}, V \in \mathbb{R}^{n \times d}} \sum_{(i,j,s) \in \mathcal{S}} \|(UV^T)_{i,j} - s_{i,j}\|_2^2$$

Practical techniques to solve:

- Alternating minimization (Fix U, minimize V. Then fix V and minimize U)
- Stochastic gradient descent on U, V
- Nuclear norm regularization (convex)



Clustering K-means

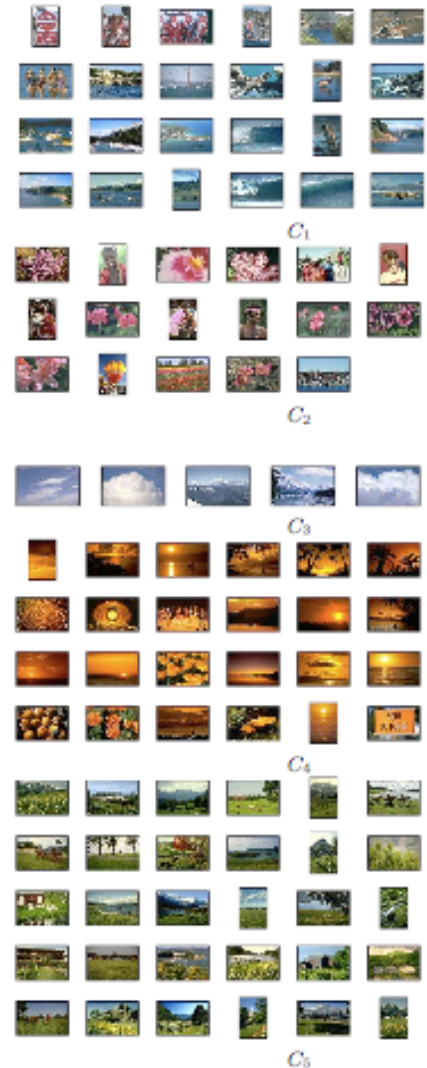
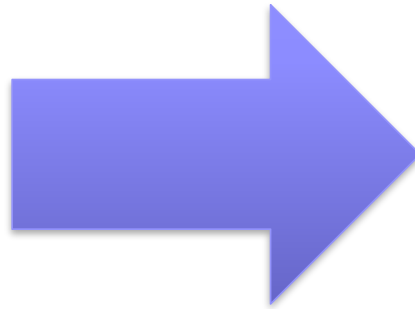
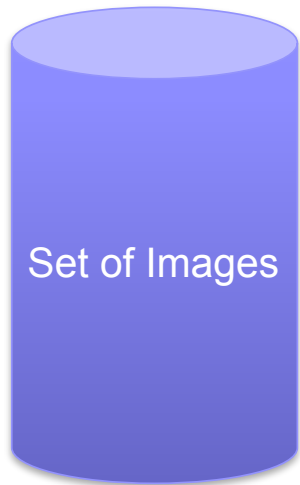
Machine Learning – CSE546

Kevin Jamieson

University of Washington

November 15, 2016

Clustering images



Clustering web search results

The screenshot shows the Clusty search interface. At the top, there are navigation links for 'web', 'news', 'images', 'wikipedia', 'blogs', 'jobs', and 'more'. A search bar contains the word 'race', and there are buttons for 'Search', 'advanced preferences', and 'remix'. Below the search bar, a sidebar on the left lists various clusters such as 'Car (28)', 'Race cars (7)', 'Photos, Races Scheduled (5)', 'Game (4)', 'Track (3)', 'Nascar (2)', 'Equipment And Safety (2)', 'Other Topics (7)', 'Photos (22)', 'Game (14)', 'Definition (13)', 'Team (18)', and 'Human (8)'. The 'Human (8)' cluster is selected, and its sub-clusters are listed: 'Classification Of Human (2)', 'Statement, Evolved (2)', 'Other Topics (4)', 'Weekend (8)', 'Ethnicity And Race (7)', 'Race for the Cure (8)', and 'Race Information (8)'. The main content area displays a list of 8 search results for the 'Human' cluster. Each result includes a title, a brief description, and a URL. The results are: 1. 'Race (classification of human beings) - Wikipedia, the free ...', 2. 'Race - Wikipedia, the free encyclopedia', 3. 'Publications | Human Rights Watch', 4. 'Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich ...', 5. 'AAPA Statement on Biological Aspects of Race', 6. 'race: Definition from Answers.com', and 7. 'Dopefish.com'. The 8th result is partially visible.

web news images wikipedia blogs jobs more »

Clusty

race Search advanced preferences

clusters sources sites remix

All Results (238)

- Car (28)
- Race cars (7)
- Photos, Races Scheduled (5)
- 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_\(classification_of_human_beings\)](http://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/backgrounder/usa/race - [cache] - Ask
- [Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich ...](#)

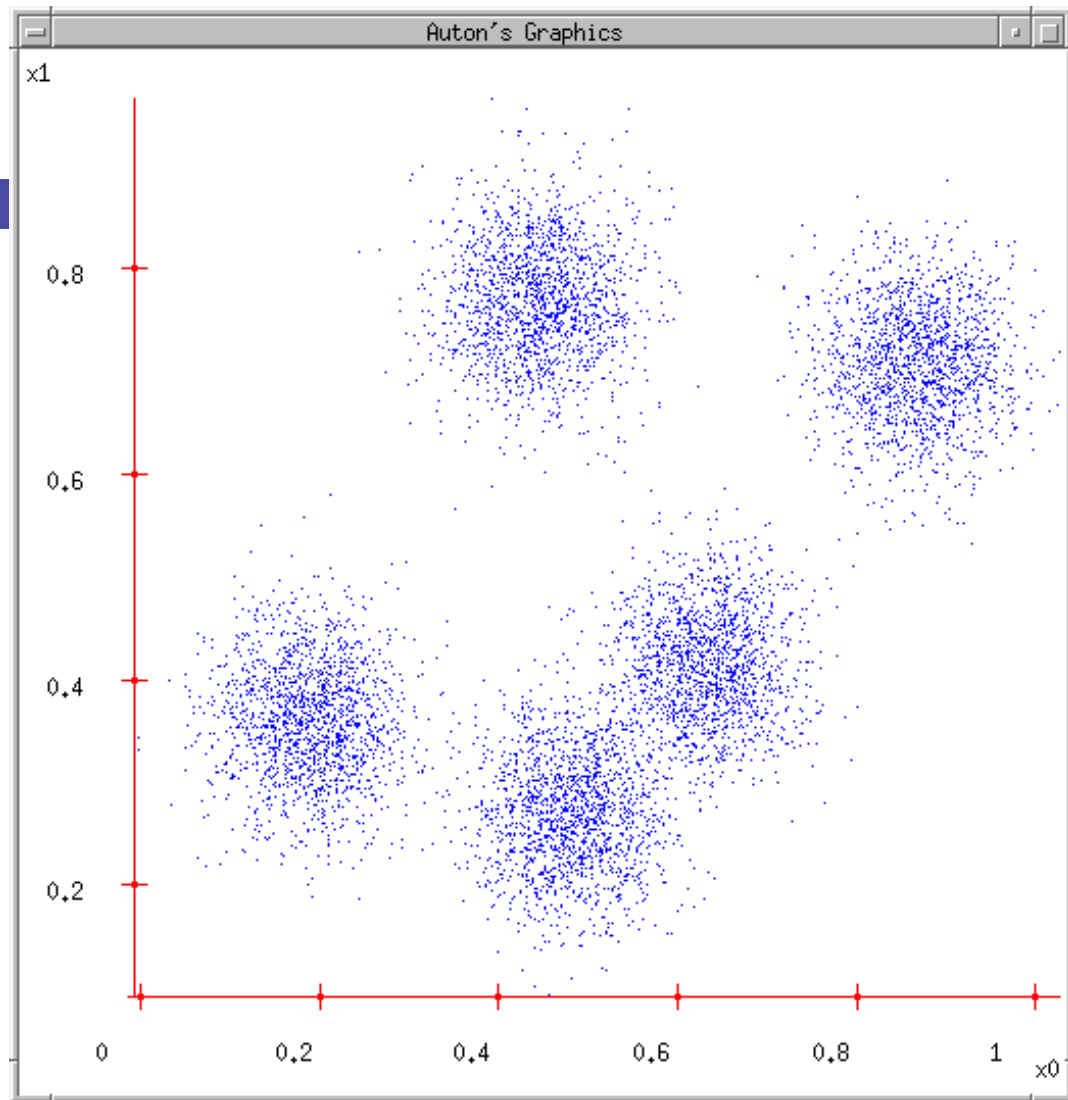
Amazon.com: **Race: The Reality Of Human Differences: Vincent Sarich, Frank Miele: Books ...** From Publishers Weekly Sarich, a Berkeley emeritus anthropologist, and Miele, an editor ...
www.amazon.com/Race-Reality-Differences-Vincent-Sarich/dp/0813340861 - [cache] - Live
- [AAPA Statement on Biological 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](#)

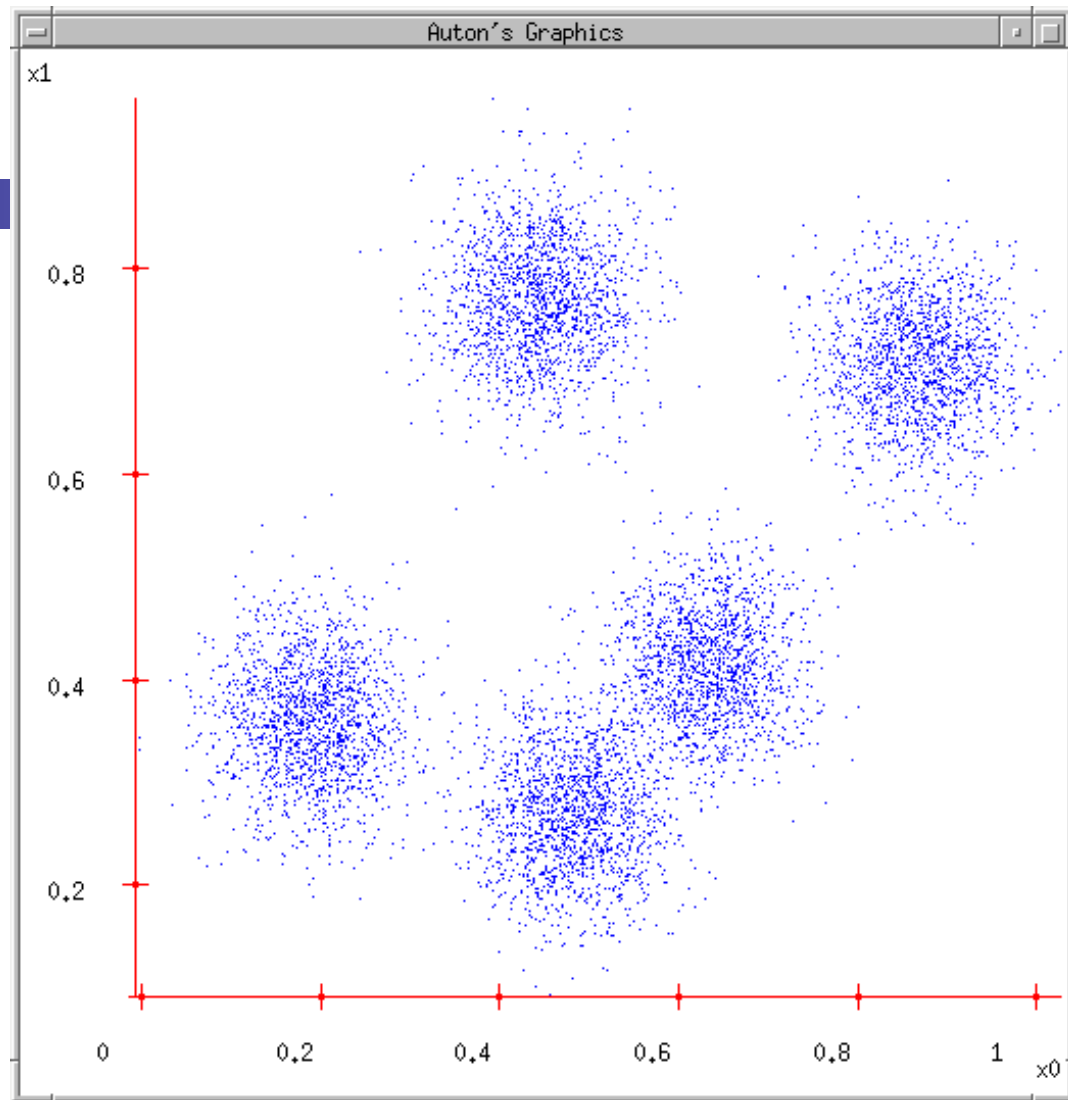
Site for newbies as well as experienced Dopefish followers, chronicling the birth of the Dopefish, its numerous appearances in several computer games, and its eventual take-over of the **human race**. Maintained by Mr. Dopefish himself, Joe Siegler of Apogee Software.
www.dopefish.com - [cache] - Open Directory

Some Data



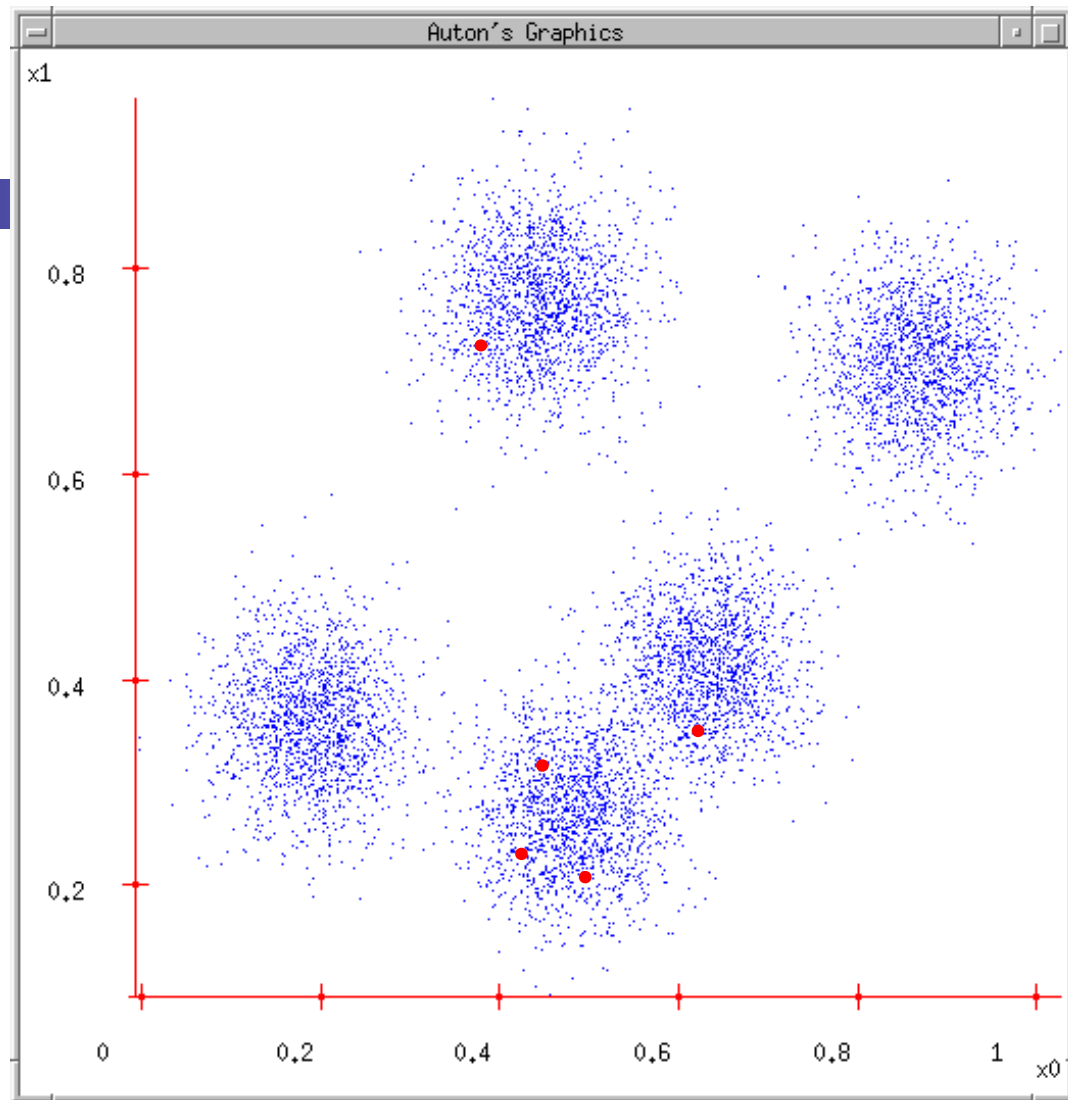
K-means

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



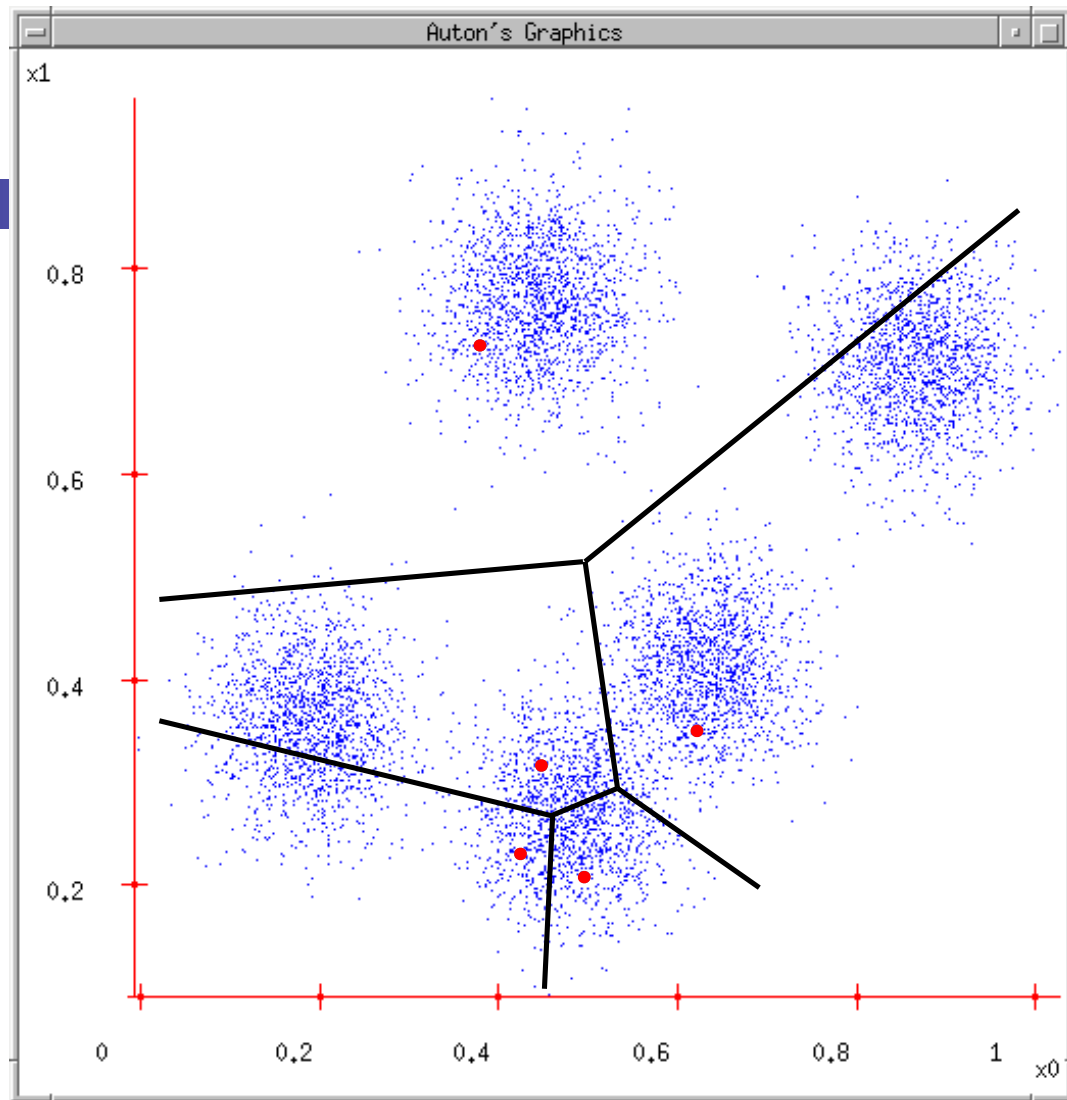
K-means

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



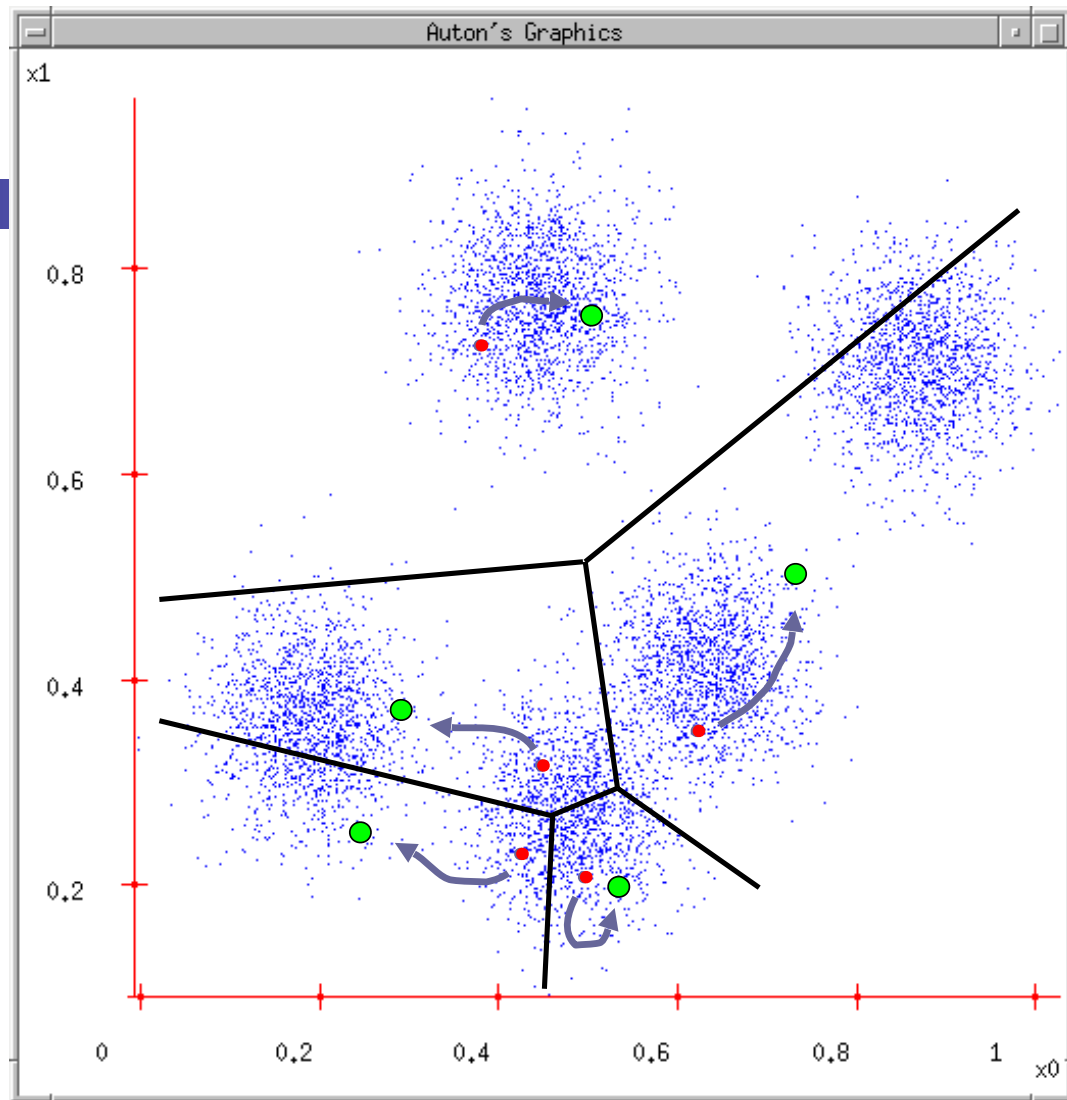
K-means

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)



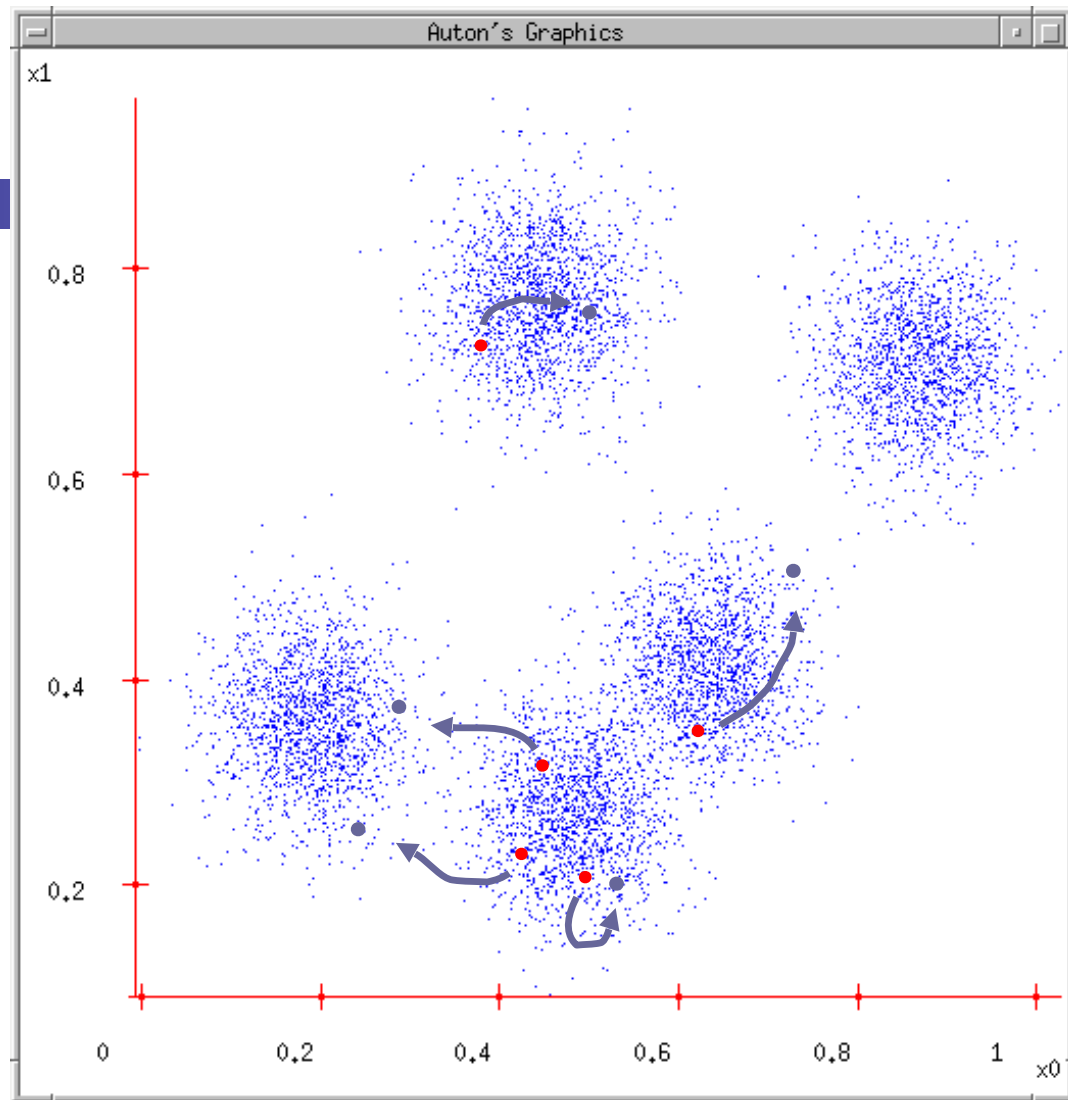
K-means

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:** μ_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!

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

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

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 1024×1024 grayscale image at 8 bits per pixel. The center image is the result of 2×2 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*

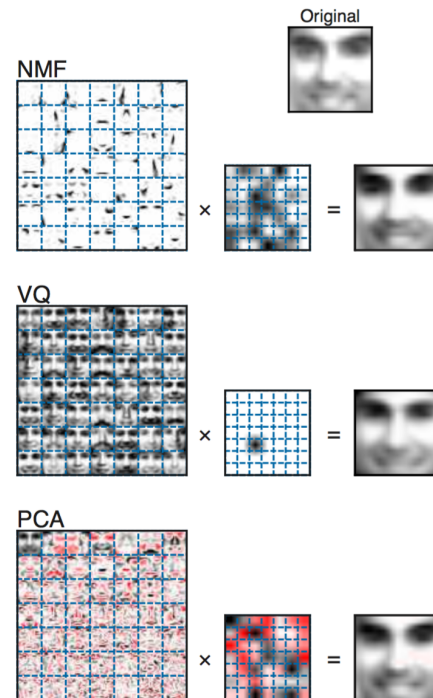
Vector Quantization, Fisher Vectors

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 1024×1024 grayscale image at 8 bits per pixel. The center image is the result of 2×2 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

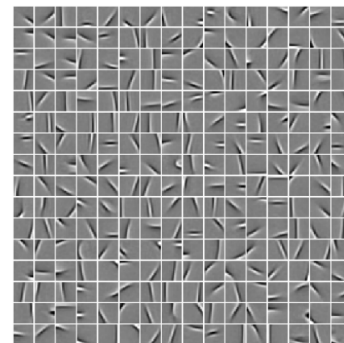
Vector Quantization (for compression)

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 1024×1024 grayscale image at 8 bits per pixel. The center image is the result of 2×2 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}

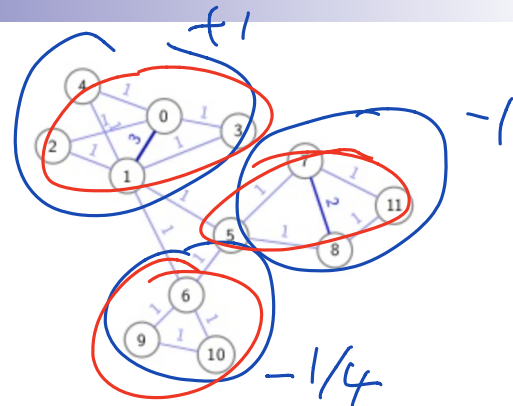
$\mathbf{W}_{i,j}$ = weight of edge (i, j)

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

$$\mathbf{L} = \mathbf{D} - \mathbf{W}$$

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

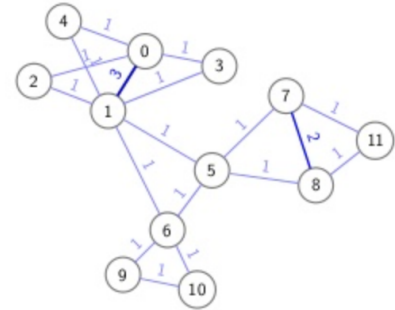
$$\begin{aligned} \mathbf{f}^T \mathbf{L} \mathbf{f} &= \sum_{i=1}^N g_i f_i^2 - \sum_{i=1}^N \sum_{i'=1}^N f_i f_{i'} w_{ii'} \\ &= \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N w_{ii'} (f_i - f_{i'})^2. \end{aligned}$$

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-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: W

$W_{i,j}$ = weight of edge (i, j)

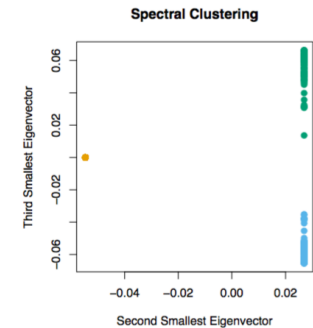
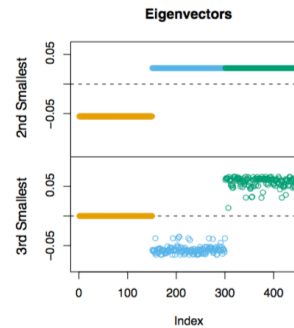
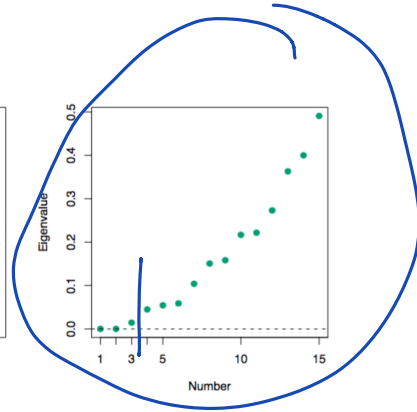
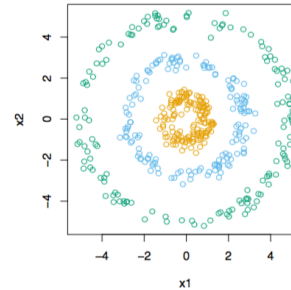
$$D_{i,i} = \sum_{j=1}^n W_{i,j} \quad L = D - W$$

$i \neq j$

$D_{i,j} = 0$

Given feature vectors, could construct:

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





Mixtures of Gaussians

Machine Learning – CSE546

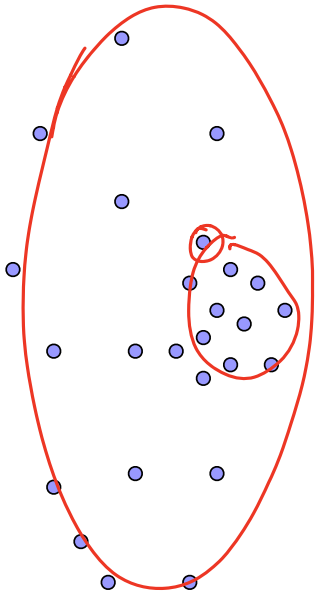
Kevin Jamieson

University of Washington

November 15, 2016

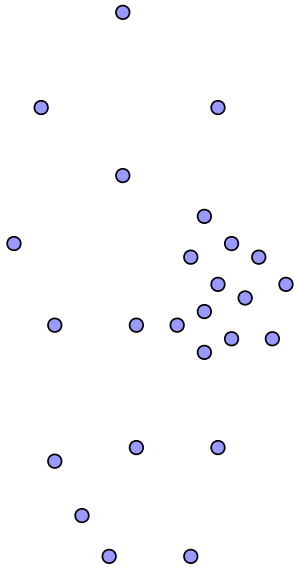
(One) bad case for k-means

- 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

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

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

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

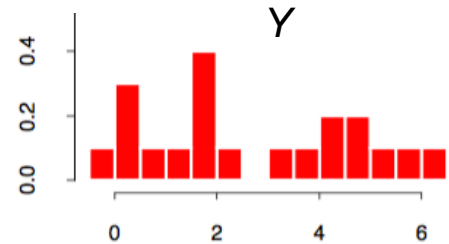
$$\Delta \in \{0, 1\} \text{ with } \Pr(\Delta = 1) = \pi$$

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

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

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

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta} l(\theta; \mathbf{z})$$



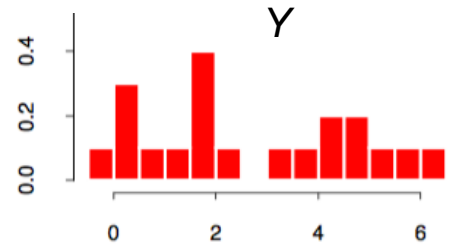
$\mathbf{Z} = \{y_i\}_{i=1}^n$ is observed data

Mixture models

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

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

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



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

$$\ell(\theta; y_i, \Delta_i = 0) = \log\left((1-\pi) \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)\right)$$

$$\ell(\theta; y_i, \Delta_i = 1) = \log\left(\pi \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)\right)$$

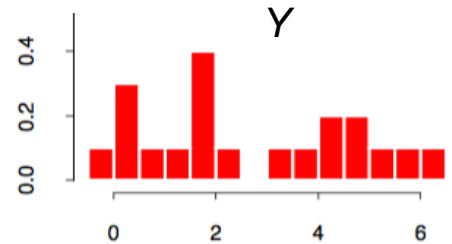
$$\sum \ell(\theta; \Delta_i, y_i)$$

Mixture models

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

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

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



$\mathbf{Z} = \{y_i\}_{i=1}^n$ 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 Δ , how would we choose θ ?

□

compute $\hat{\mu}_1$ $\hat{\mu}_2$
 $\hat{\sigma}_1$ $\hat{\sigma}_2$

Mixture models

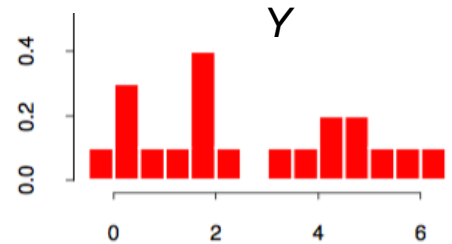
$$Y_1 \sim N(\mu_1, \sigma_1^2),$$

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

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

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

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



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

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

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

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

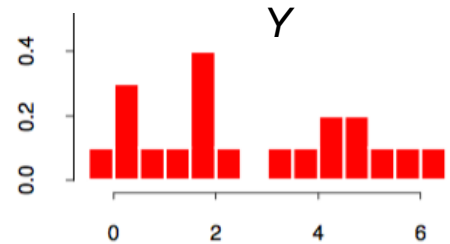
If we knew θ , how would we choose Δ ?

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

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

$$\gamma_i(\theta) = \mathbb{E}[\Delta_i | \theta, \mathbf{Z}] = \frac{\pi \phi_{\mu_2, \sigma_2^2}(y_i)}{\pi \phi_{\mu_2, \sigma_2^2}(y_i) + (1 - \pi) \phi_{\mu_1, \sigma_1^2}(y_i)}$$

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

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1 - \hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)}, \quad i = 1, 2, \dots, N. \quad (8.42)$$

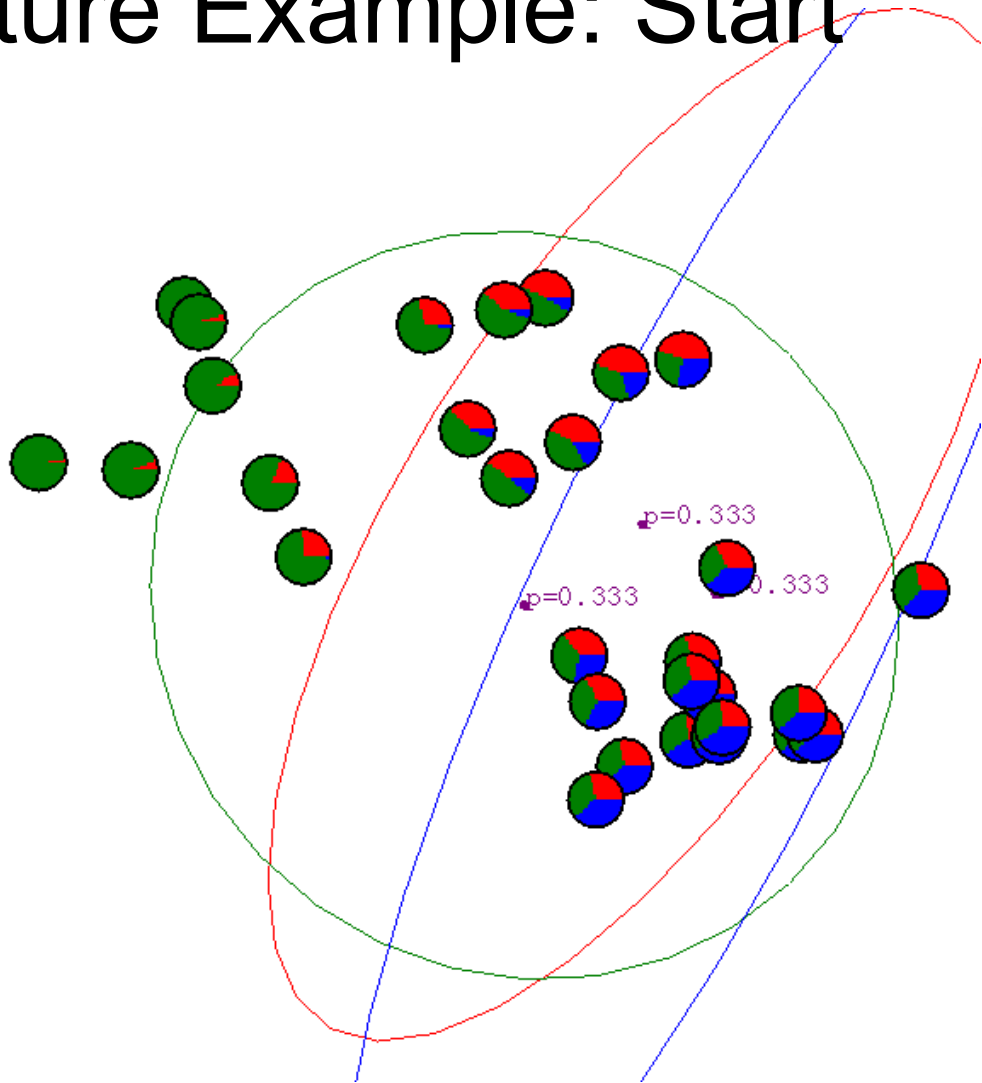
3. *Maximization Step*: compute the weighted means and variances:

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

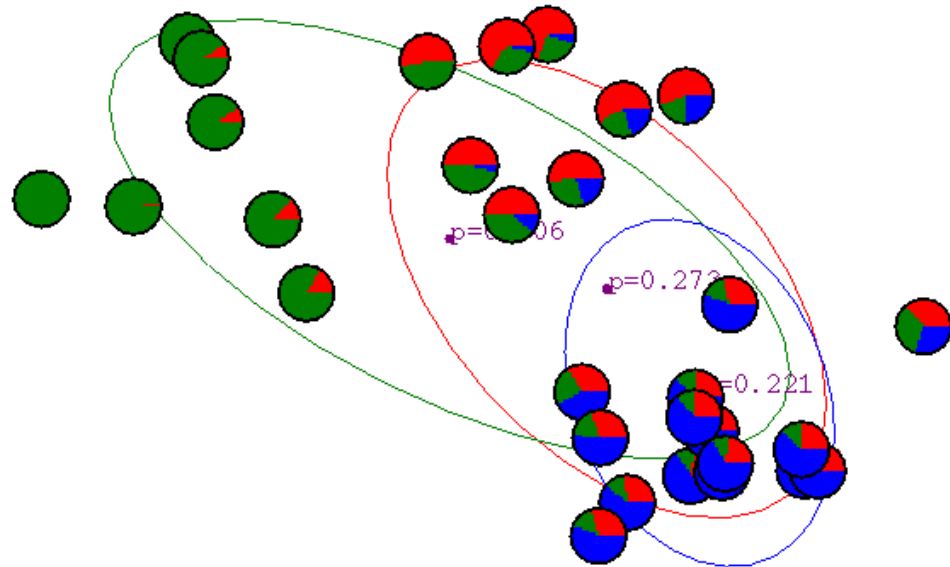
and the mixing probability $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$.

4. Iterate steps 2 and 3 until convergence.
-

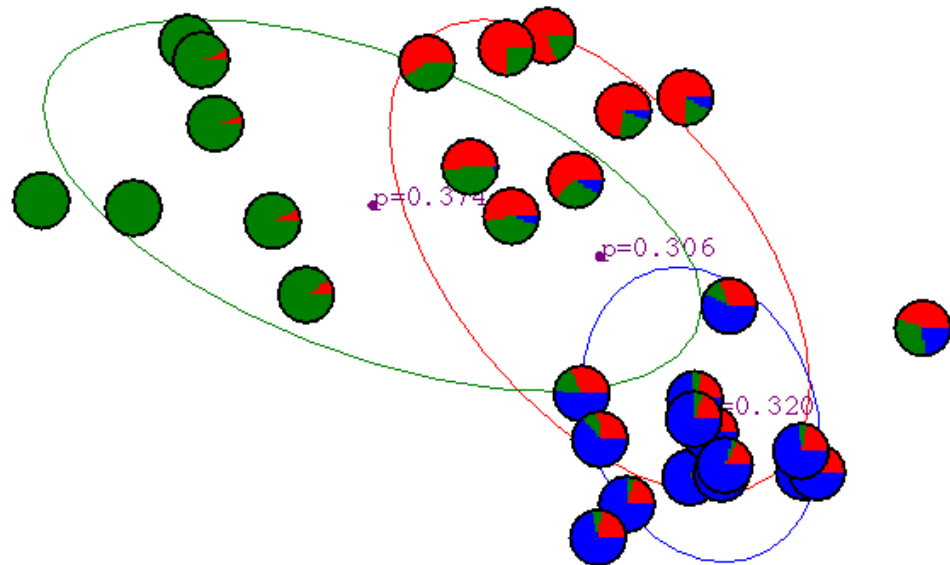
Gaussian Mixture Example: Start



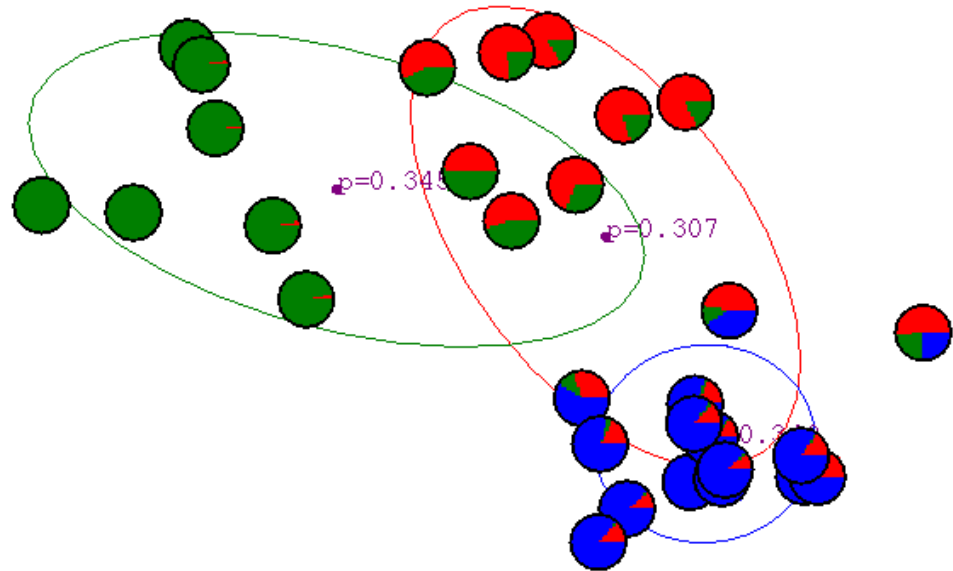
After first iteration



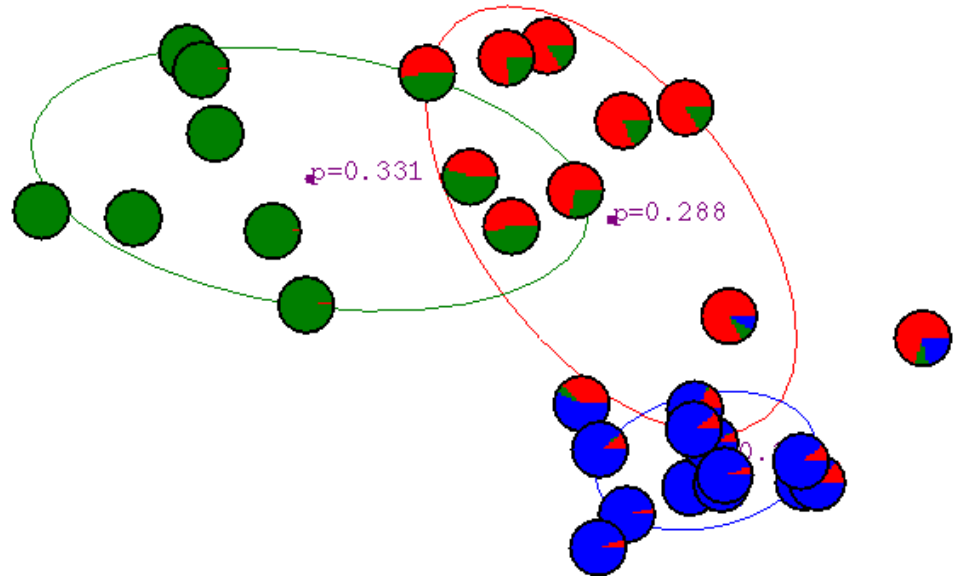
After 2nd iteration



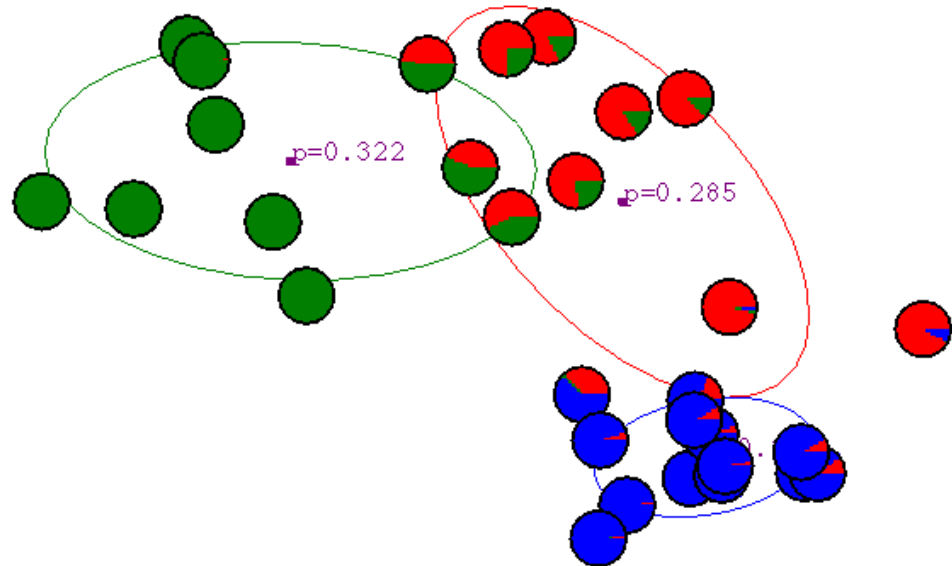
After 3rd iteration



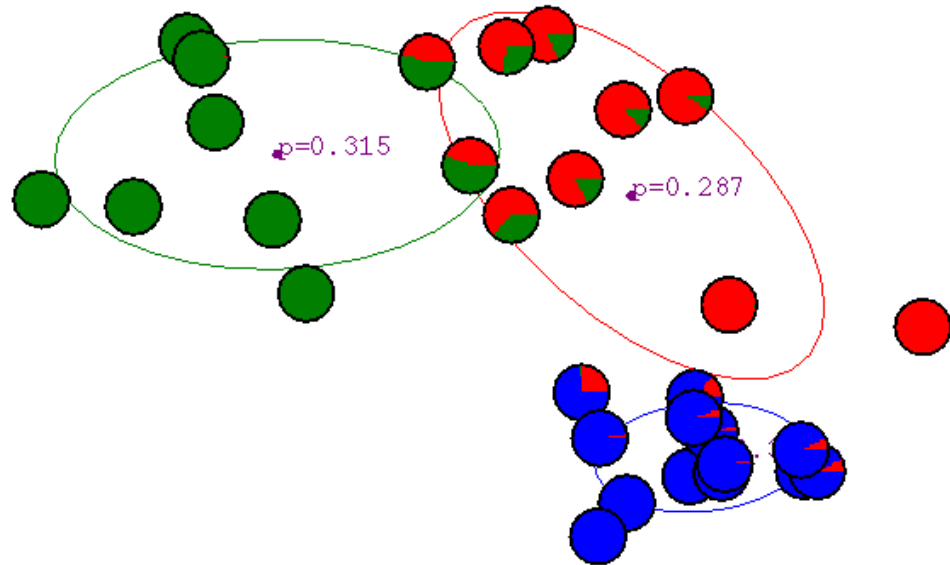
After 4th iteration



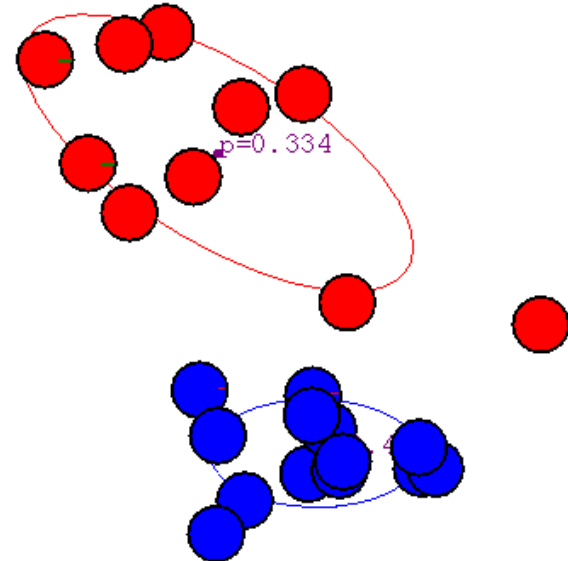
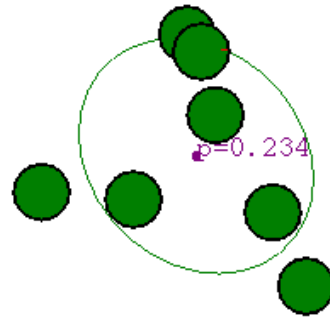
After 5th iteration



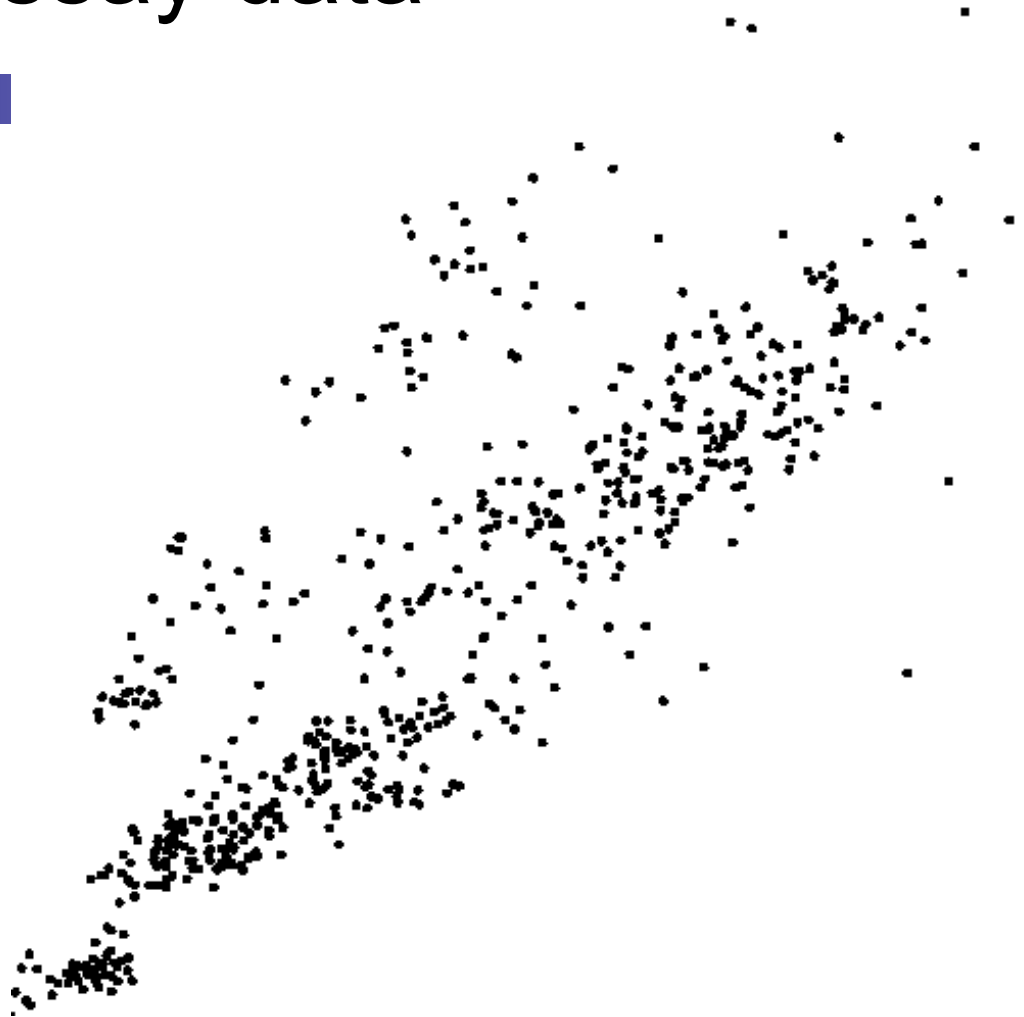
After 6th iteration



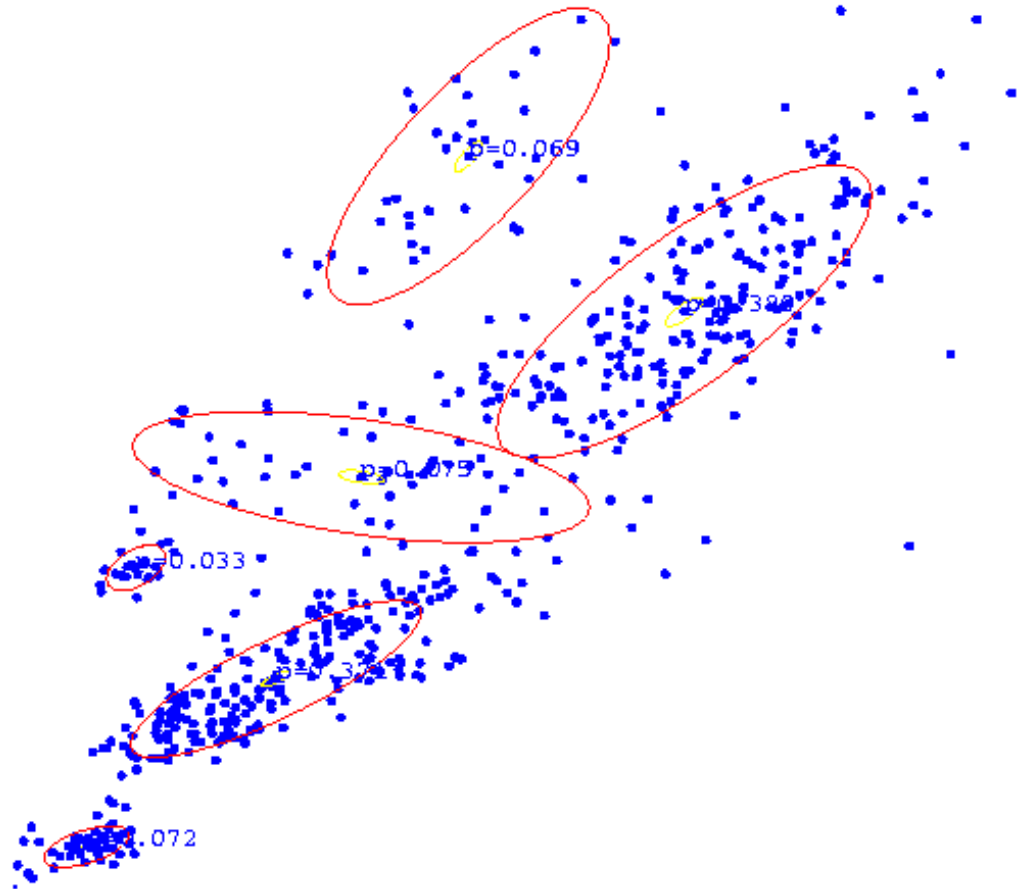
After 20th iteration




Some Bio Assay data



GMM clustering of the assay data





Resulting Density Estimator

