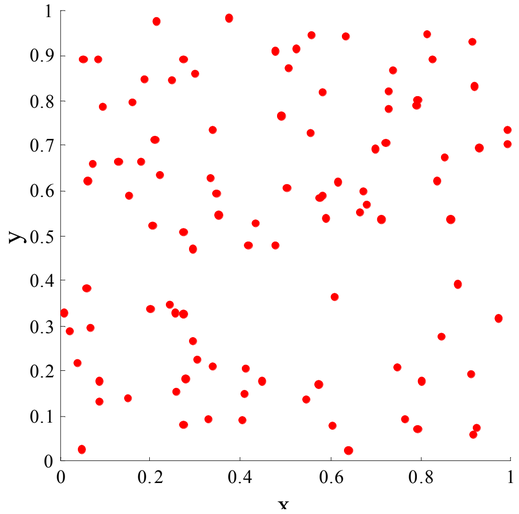

# Clustering

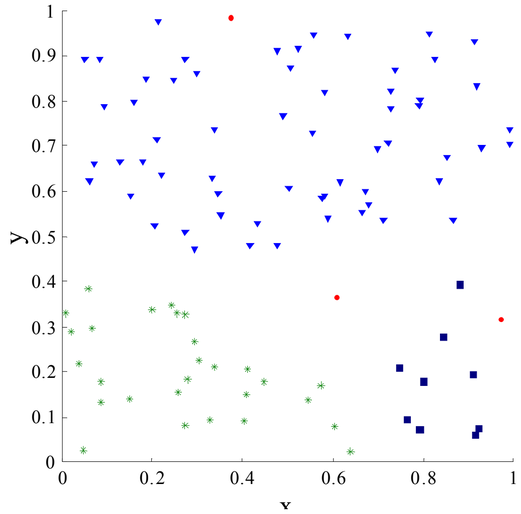
Shantanu Jain


# Evaluation of Clustering

# Clusters in Random Data

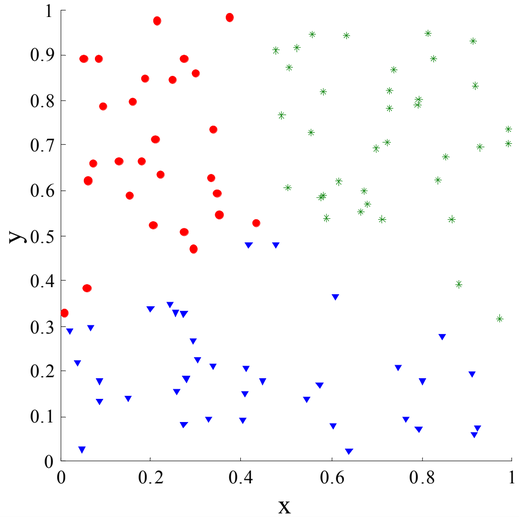
**Random Points**

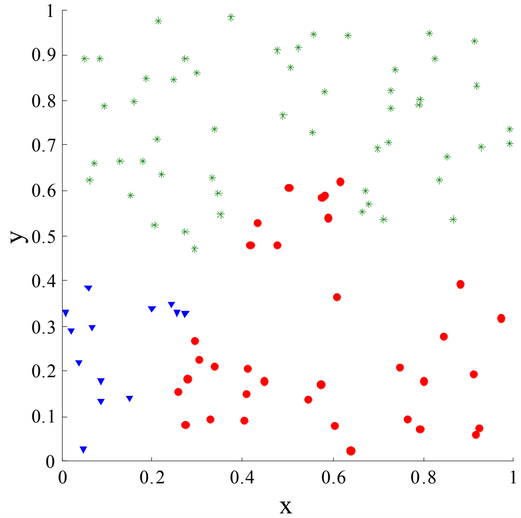



**DBSCAN**

**K-means**




**Complete Link**

# Clustering Criteria

***Internal Quality Criteria***

Measure compactness of clusters

- Sum of Squared Error (SSE)
- Scatter Criteria

***External Quality Criteria***

Measure correspondence to true labels

- Precision-Recall Measure
- Mutual Information

# Scatter Criteria (Internal)

$\{x_1, \ldots, x_N\}$ Data

$\{z_1, \ldots, z_N\}$ Cluster Assignments

$$N_k = \sum_{n=1}^{N} I[z_n = k]$$ Number of points in cluster k

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} I[z_n = k] x_n$$ Center of cluster k

$$\bar{\mu} = \frac{1}{N} \sum_{n=1} x_n = \frac{1}{N} \sum_{k=1}^{K} N_k \mu_k$$ Mean of all data (total mean)

# Scatter Criteria (Internal)

Per-cluster mean and data mean (from last slide)

$$\bar{\mu} = \frac{1}{N}\sum_{k=1}^{K} N_k \mu_k \qquad \mu_k = \frac{1}{N_k}\sum_{n=1}^{N} I[z_n = k] x_n$$

Within-cluster Scatter Matrix

$$S^{\mathrm{W}} = \sum_{k=1}^{K} S_k \qquad S_k = \sum_{n=1}^{N} I[z_n = k](x_n - \mu_k)(x_n - \mu_k)^T$$

D x 1 1 x D

D x D

Between-cluster Scatter Matrix

$$S^{\mathrm{B}} = \sum_{k=1}^{K} N_k (\mu_k - \bar{\mu})(\mu_k - \bar{\mu})^\top$$

# Within cluster scatter matrix

$$S^{\mathrm{W}} = \sum_{k=1}^{K} S_k \qquad S_k = \sum_{n=1}^{N} I[z_n = k](x_n - \mu_k)(x_n - \mu_k)^T$$

# Scatter Criteria (Internal)

Per-cluster mean and data mean (from last slide)

$$\bar{\mu} = \frac{1}{N}\sum_{k=1}^{K} N_k \mu_k \qquad \mu_k = \frac{1}{N_k}\sum_{n=1}^{N} I[z_n = k] x_n$$

Within-cluster Scatter Matrix

$$S^{\mathrm{W}} = \sum_{k=1}^{K} S_k \qquad S_k = \sum_{n=1}^{N} I[z_n = k](x_n - \mu_k)(x_n - \mu_k)^T$$

D x 1 1 x D

D x D

Between-cluster Scatter Matrix

$$S^{\mathrm{B}} = \sum_{k=1}^{K} N_k (\mu_k - \bar{\mu})(\mu_k - \bar{\mu})^\top$$

# Scatter Criteria (Internal)

Total Scatter
(*covariance of data*)

Within-Cluster Scatter

Between-Cluster Scatter

$$S^{\mathrm{T}} = \sum_{n=1}^{N} (x_n - \bar{\mu})(x_n - \bar{\mu})^{\top} = S^{\mathrm{W}} + S^{\mathrm{B}}$$

Depends only on data,
(*independent of clusters*)

Want this
to be small

Want this
to be large

Calinski-Harabaz (CH) index

$$\frac{\mathrm{Tr}(S^{\mathrm{B}})}{\mathrm{Tr}(S^{\mathrm{W}})}$$

Trace
(from matrix cookbook)

$\mathrm{Tr}(S)$ = sum of diagonal elements of $S$

# Silhouette Coefficient (Internal)

Silhouette Coefficient

$$\mathrm{SC} = \frac{1}{N}\sum_{n=1}^{N}\frac{b_n - a_n}{\max(a_n, b_n)}$$

$-1 \leq \mathrm{SC} \leq 1$

For good clustering above 0

Distance to center of cluster

$$a_n = ||x_n - \mu_{z_n}||$$

Distance to closest other clusters

$$b_n = \min_{l \neq z_n} ||x_n - \mu_l||$$

# Mutual Information (External)

Random variables $A$ and $B$

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)} \quad = \mathbf{E}\left[\log \frac{p(A,B)}{p(A)p(B)}\right]$$

**Abuse of notation:**

- Let $\Omega_A$ be the sample space of r.v. $A$.
- $a \in A$ is short for $a \in \Omega_A$

$$= \mathbf{KL}\big(p(a,b) \,||\, p(a)p(b)\big)$$

# Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Independent Variables

$$p(a,b) = p(a)p(b)$$

# Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Independent Variables

$$p(a,b) = p(a)p(b)$$

$$I(A;B) = \sum_{a \in A, b \in B} p(a)p(b) \log \frac{p(a)p(b)}{p(a)p(b)} = 0$$

# Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Perfectly Correlated Variables

$$A = B$$

$$p(A = a, B = b) = p(B = a, B = b)$$
$$= I[a = b]p(B = b)$$

$$p(A = a) = p(B = a)$$

# Mutual Information (External)

$$I(A;B) = \sum_{a \in A, b \in B} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

Perfectly Correlated Variables

$$I(A;B) = \sum_{a \in A, b \in B} p(A=a, B=b) \log \frac{p(A=a, B=b)}{p(A=a)p(B=b)}$$

$$= \sum_{b \in B} p(A=b, B=b) \log \frac{p(A=b, B=b)}{p(A=b)p(B=b)}$$

$$= \sum_{b \in B} p(B=b) \log \frac{p(B=b)}{p(B=b)p(B=b)}$$

$$= \sum_{b \in B} p(B=b) \log \frac{1}{p(B=b)} = H(B)$$

Entropy of r.v. $B$: a measure of uncertainty.

# Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

$Y$ and $Z$ are r.v.'s representing true and predicted cluster label of a random point in the dataset. The do not contain labels for all points in the dataset.

$$Z \in \Omega_Z = \{1,2 \ldots K\}$$
$$Y \in \Omega_Y = \{1,2 \ldots L\}$$

The summation is not over the data points. The summation is over pairs of values $Y$ and $Z$ can take;
i.e., (1,1), (1,2) … (1,L),
(2,1), (2,2), … (2,L),
…
(K,1), (K,2), … (K,L)

# Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

Here $Y$ and $Z$ are r.v.'s are representing the true and predicted labels of a single point, not the entire dataset.

$y_n$: Is a realization of $Y$ for the $n^{th}$ *data point*
$z_n$: Is a realization of $Z$ for the $n^{th}$ *data point*

$$p(Y = k) = \frac{1}{N} \sum_n I(y_n = k) \qquad p(Z = l) = \frac{1}{N} \sum_n I(z_n = l)$$

$$p(Y = k, Z = l) = \frac{1}{N} \sum_n I(y_n = k \wedge z_n = l)$$

# Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

| $p(y,z)$ | 1 | 2 | 3 | $p(y)$ |
|---|---|---|---|---|
| cat | 0.39 | 0.08 | 0.02 | 0.49 |
| dog | 0.06 | 0.31 | 0.01 | 0.38 |
| parrot | 0.01 | 0.01 | 0.11 | 0.13 |
| $p(z)$ | 0.46 | 0.40 | 0.14 | |

# Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

| $p(y,z)$ | 1 | 2 | 3 | $p(y)$ |
|---|---|---|---|---|
| cat | 0.39 | 0.08 | 0.02 | 0.49 |
| dog | 0.06 | 0.31 | 0.01 | 0.38 |
| parrot | 0.01 | 0.01 | 0.11 | 0.13 |
| $p(z)$ | 0.46 | 0.40 | 0.14 | |

What happens to I(Y;Z) if we swap cluster labels?

# Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

| $p(y,z)$ | 1 | 2 | 3 | $p(y)$ |
|---|---|---|---|---|
| cat | 0.08 | 0.39 | 0.02 | 0.49 |
| dog | 0.31 | 0.06 | 0.01 | 0.38 |
| parrot | 0.01 | 0.01 | 0.11 | 0.13 |
| $p(z)$ | 0.40 | 0.46 | 0.04 | |

What happens to I(Y;Z) if we swap cluster labels?

# Mutual Information (External)

$$I(Y;Z) = \sum_{y,z} p(y,z) \log \frac{p(y,z)}{p(y)p(z)}$$

| $p(y,z)$ | 1 | 2 | 3 | $p(y)$ |
|---|---|---|---|---|
| cat | 0.08 | 0.39 | 0.02 | 0.49 |
| dog | 0.31 | 0.06 | 0.01 | 0.38 |
| parrot | 0.01 | 0.01 | 0.11 | 0.13 |
| $p(z)$ | 0.40 | 0.46 | 0.04 | |

Mutual information can also be defined if the number of true and predicted clusters are different.

Mutual Information is ***invariant*** under label permutations

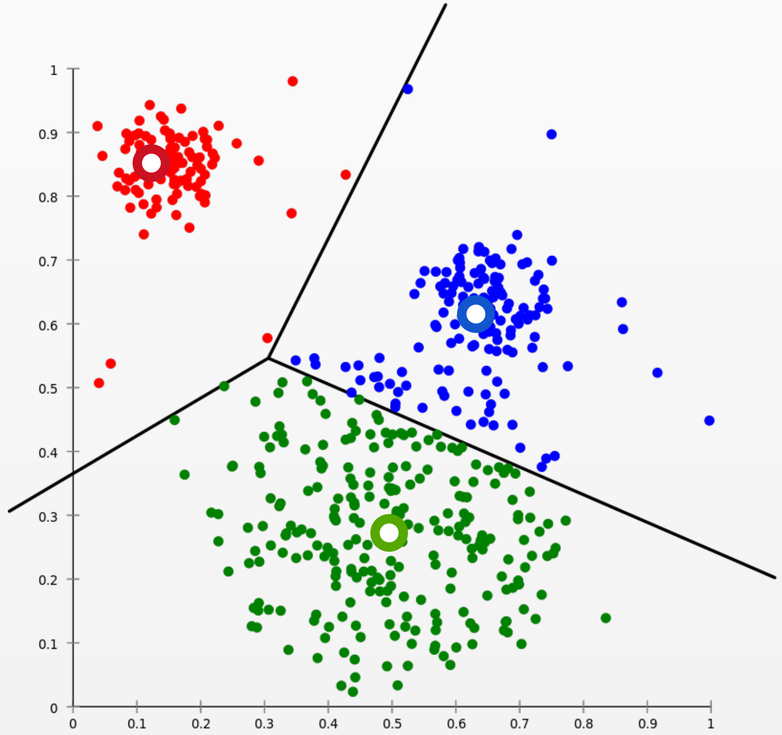

# Clustering

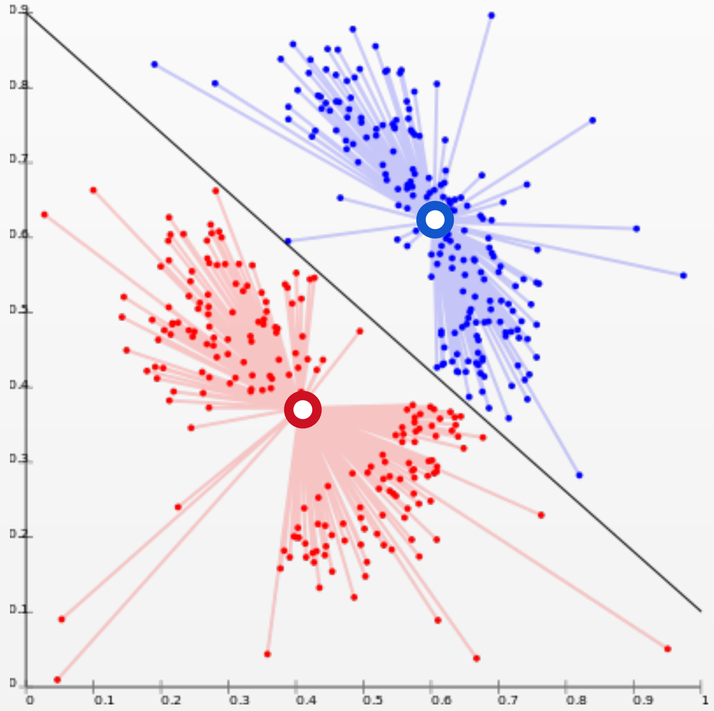
Shantanu Jain


# Gaussian Mixture Models

# Four Types of Clustering

## 1. *Centroid-based (K-means, K-medoids)*

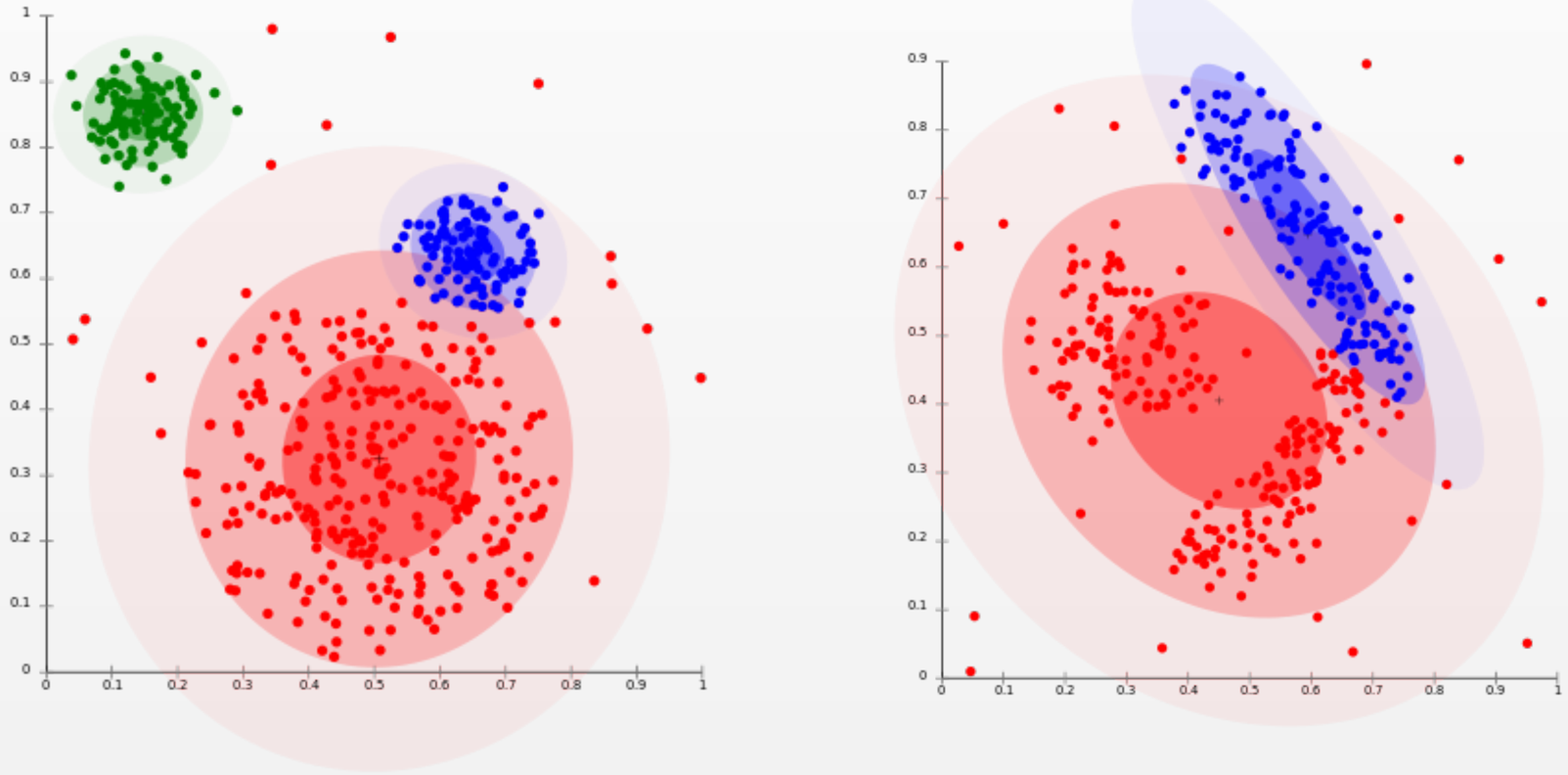



Notion of Clusters: Voronoi tesselation

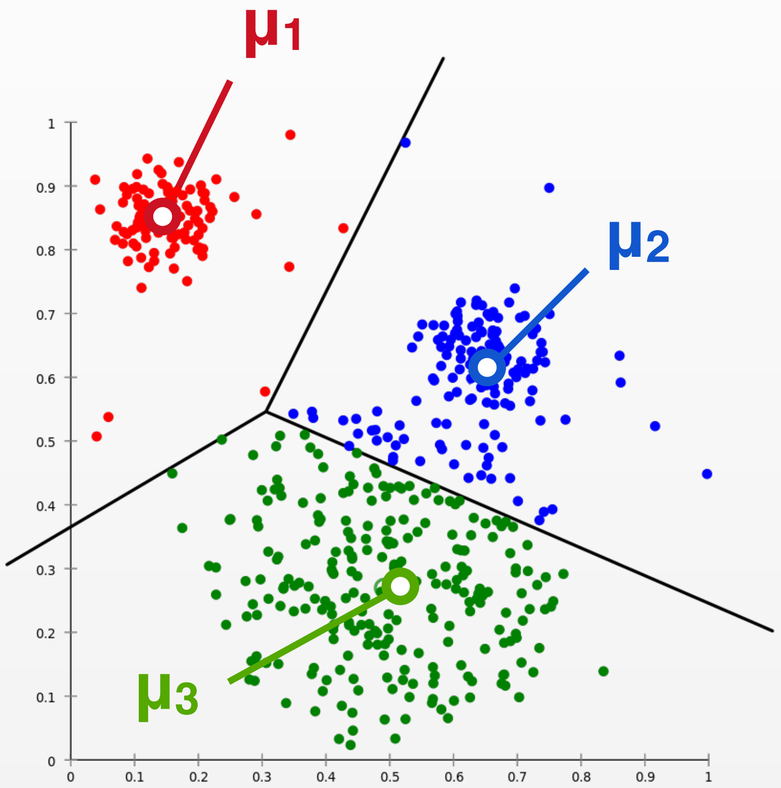
# Four Types of Clustering

## 4. *Distribution-based (Mixture Models)*




Notion of Clusters: Distributions on features

# Review K-means Clustering

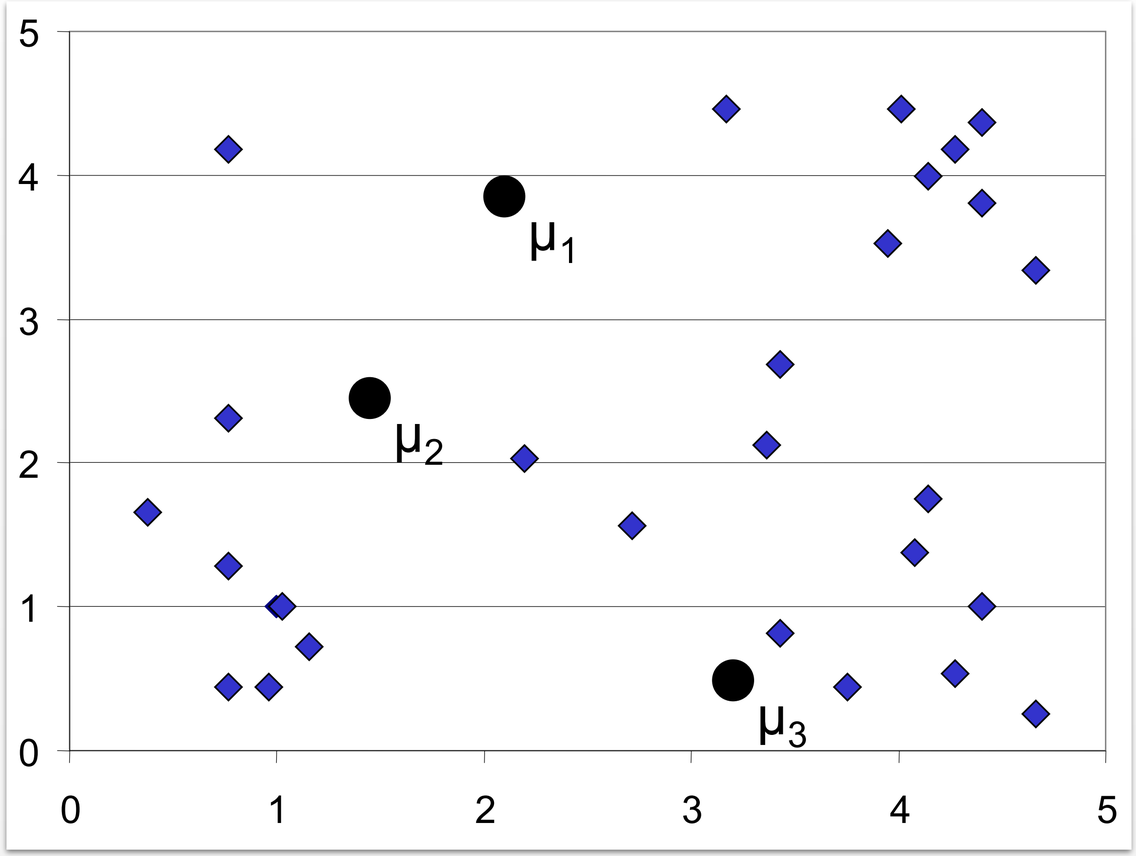



Loss: Sum of Squared Distances

$$L(\boldsymbol{\mu}, \boldsymbol{z}) = \sum_{k=1}^{K} \sum_{n=1}^{N} I[z_n = k] (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^2$$

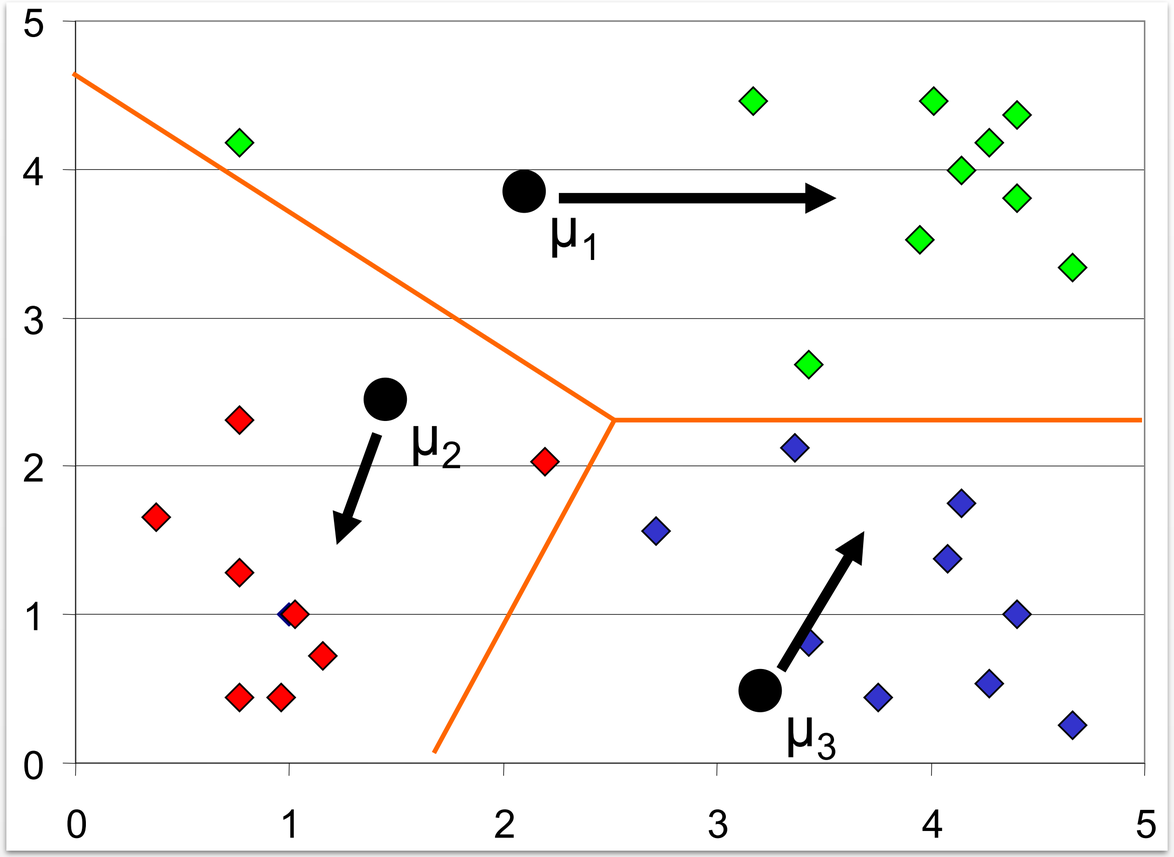
- Randomly initialize $\mu$
- Alternate between two steps
  1. Minimize L(μ, z) with respect to $z$ *(assign points to closest cluster)*
  2. Minimize L(μ, z) with respect to $\mu$ *(place clusters close to points)*

# K-means Clustering



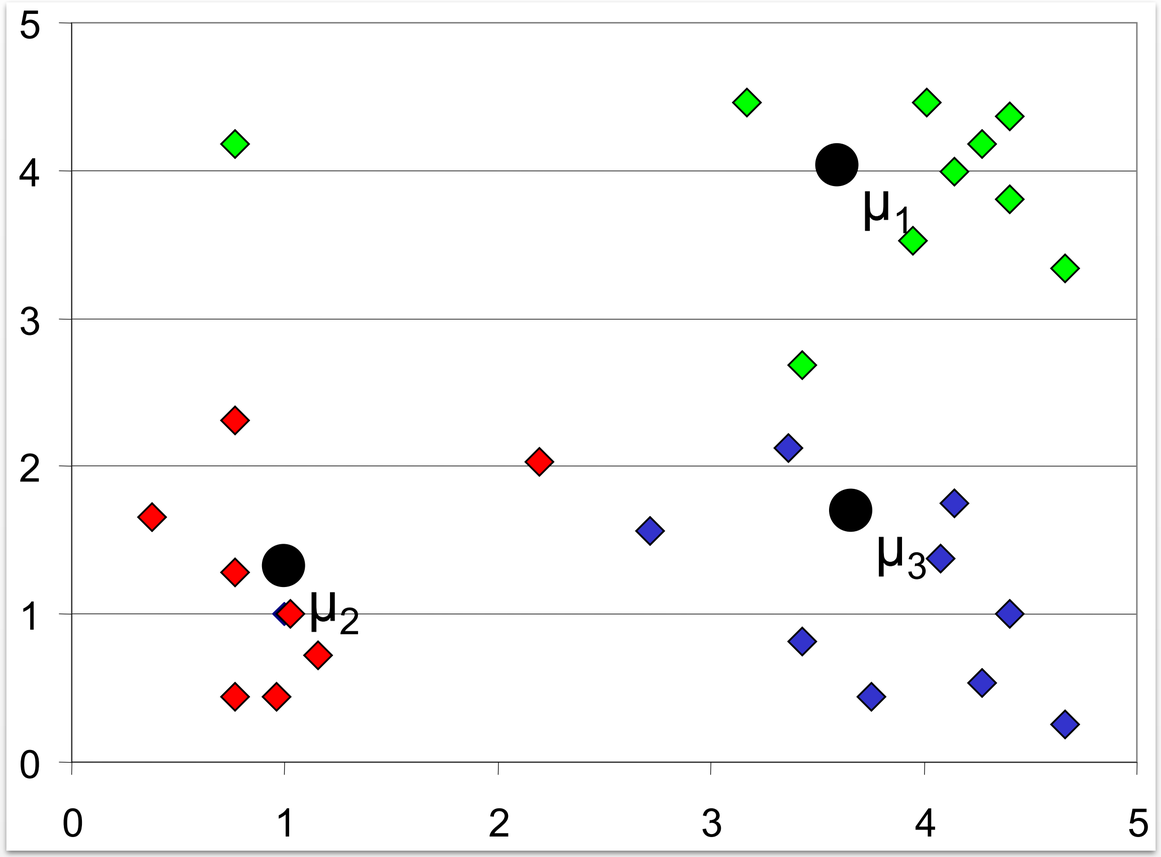

Randomly initialize $K$ means $\mu_k$

# K-means Clustering



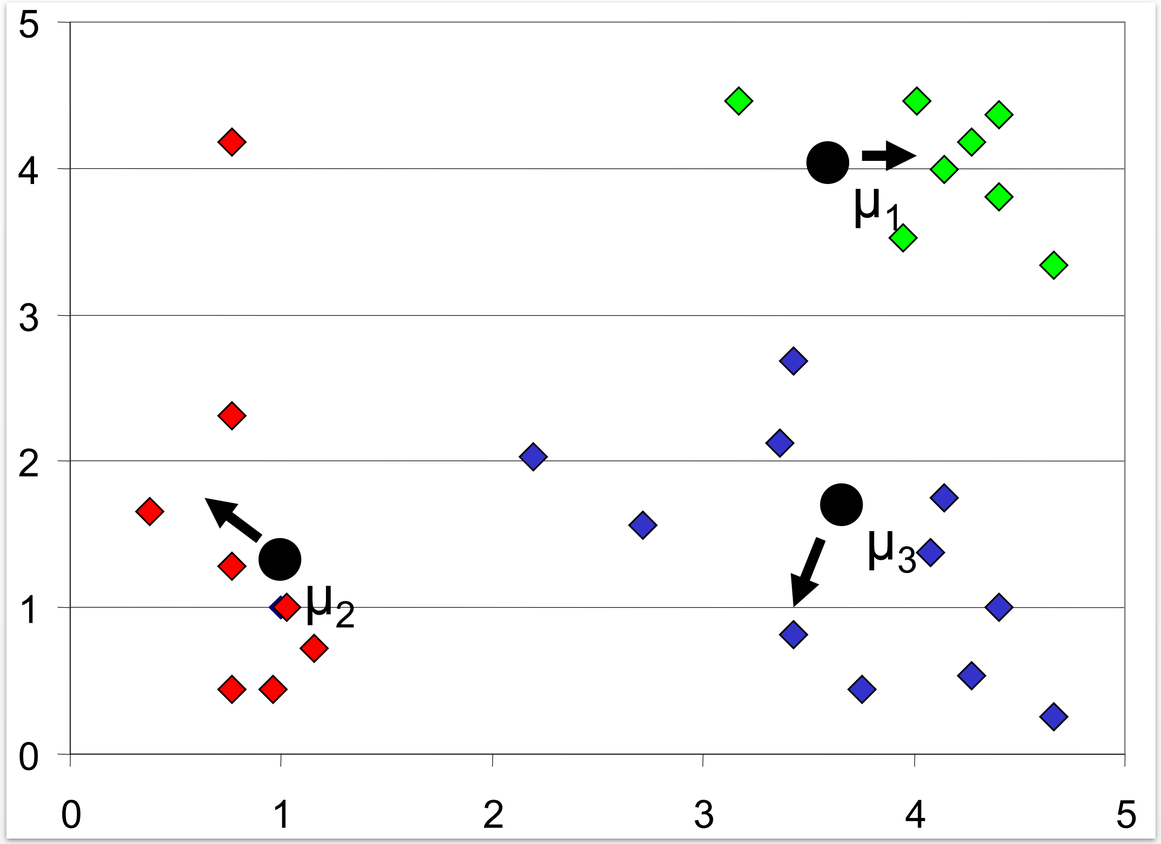

Assign each point to closest cluster,
then update means to average of points

# K-means Clustering




Assign each point to closest cluster,
then update means to average of points

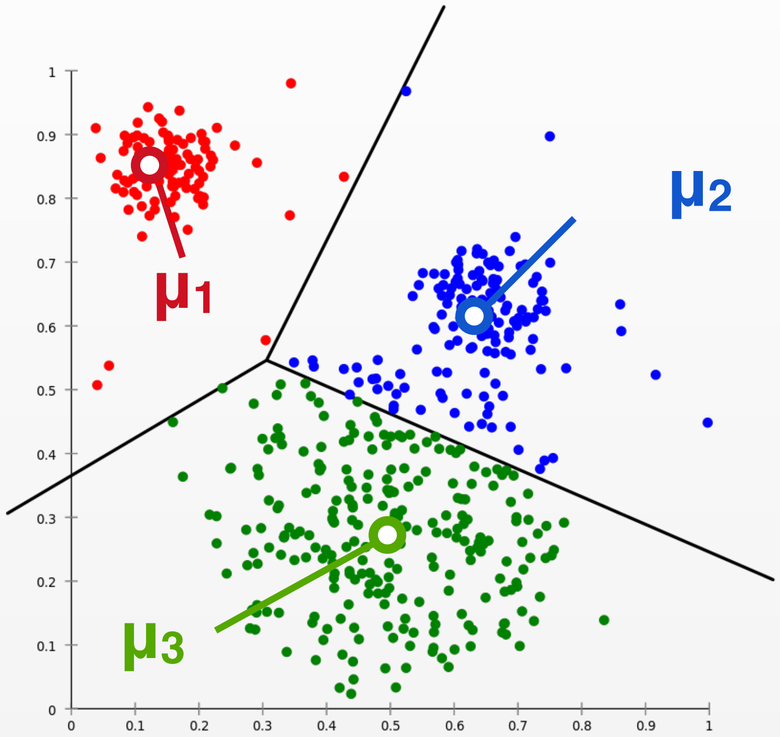
# K-means Clustering

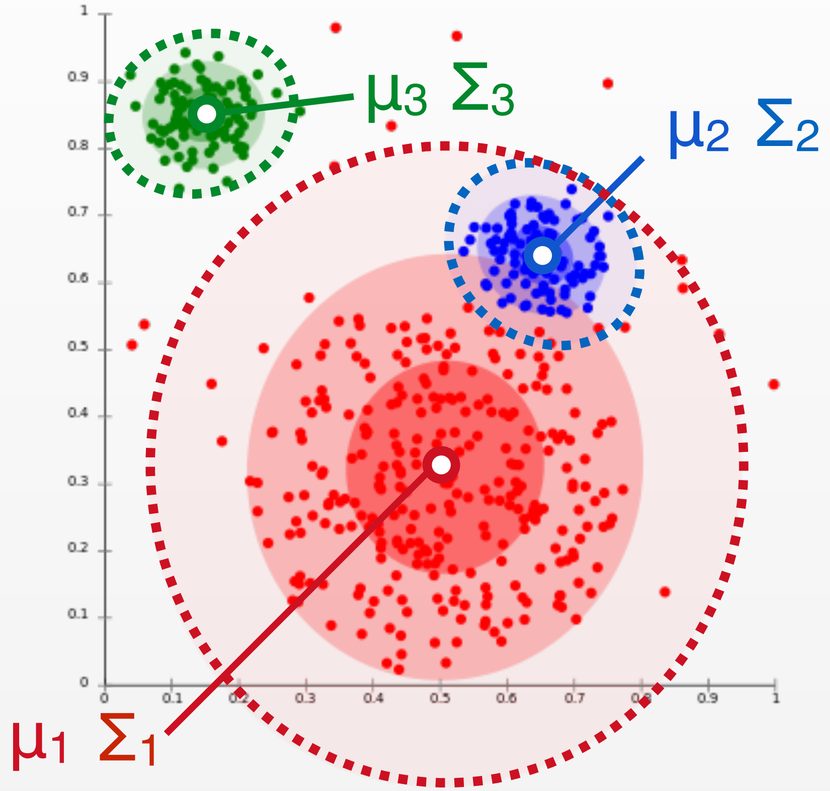



Repeat until convergence
(no points reassigned, means unchanged)

# K-Means vs Gaussian Mixture Models

Idea1: Learn both means $\mu_k$ and covariances $\Sigma_k$

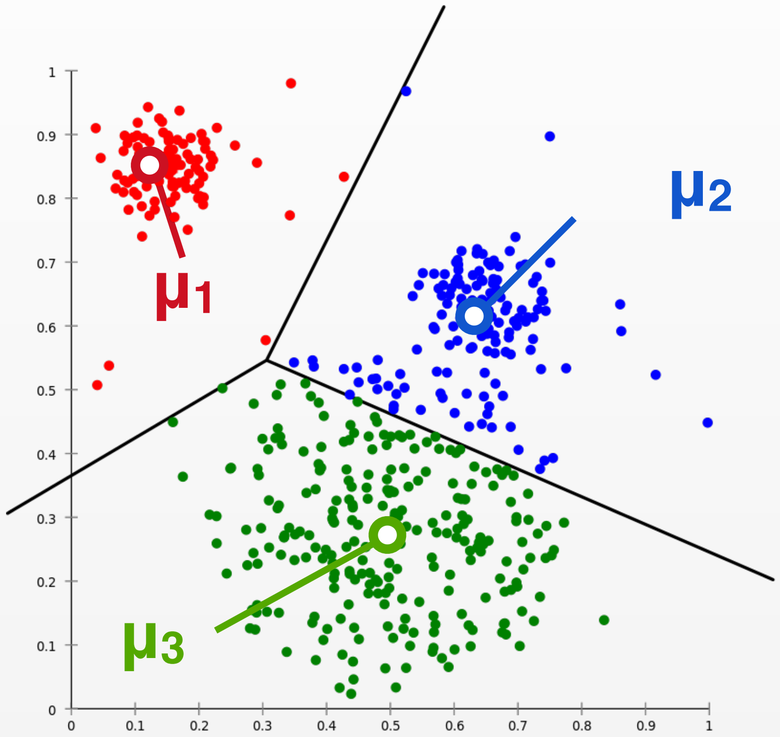


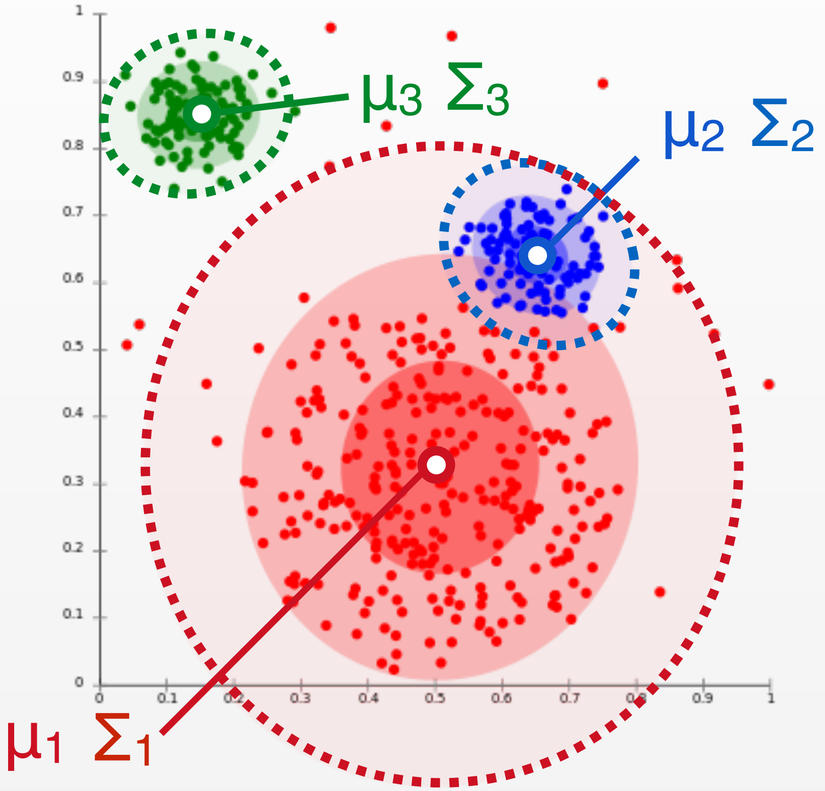




Don't just learn *where* the center of the cluster is,
but also *how big it is*, and *what shape it has*.

# K-Means vs Gaussian Mixture Models

Idea 2: Replace *hard* assignments with *soft* assignments




**Hard** assignments to clusters

$$\gamma_{nk} = I[z_n = k]$$

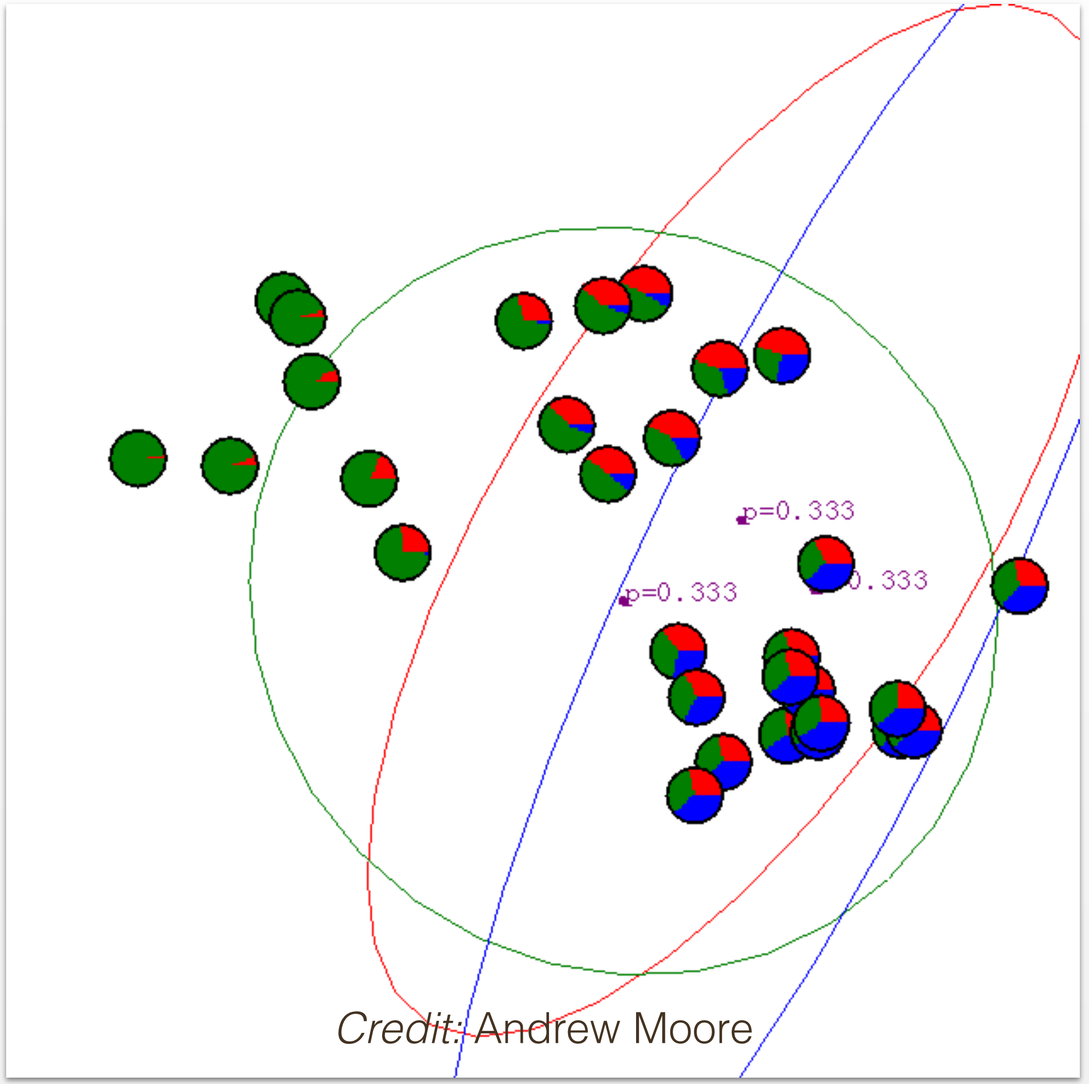
(one-hot vector)

**Soft** assignments to clusters

$$\gamma_{nk} = p(z_n = k \mid \boldsymbol{x}_n)$$

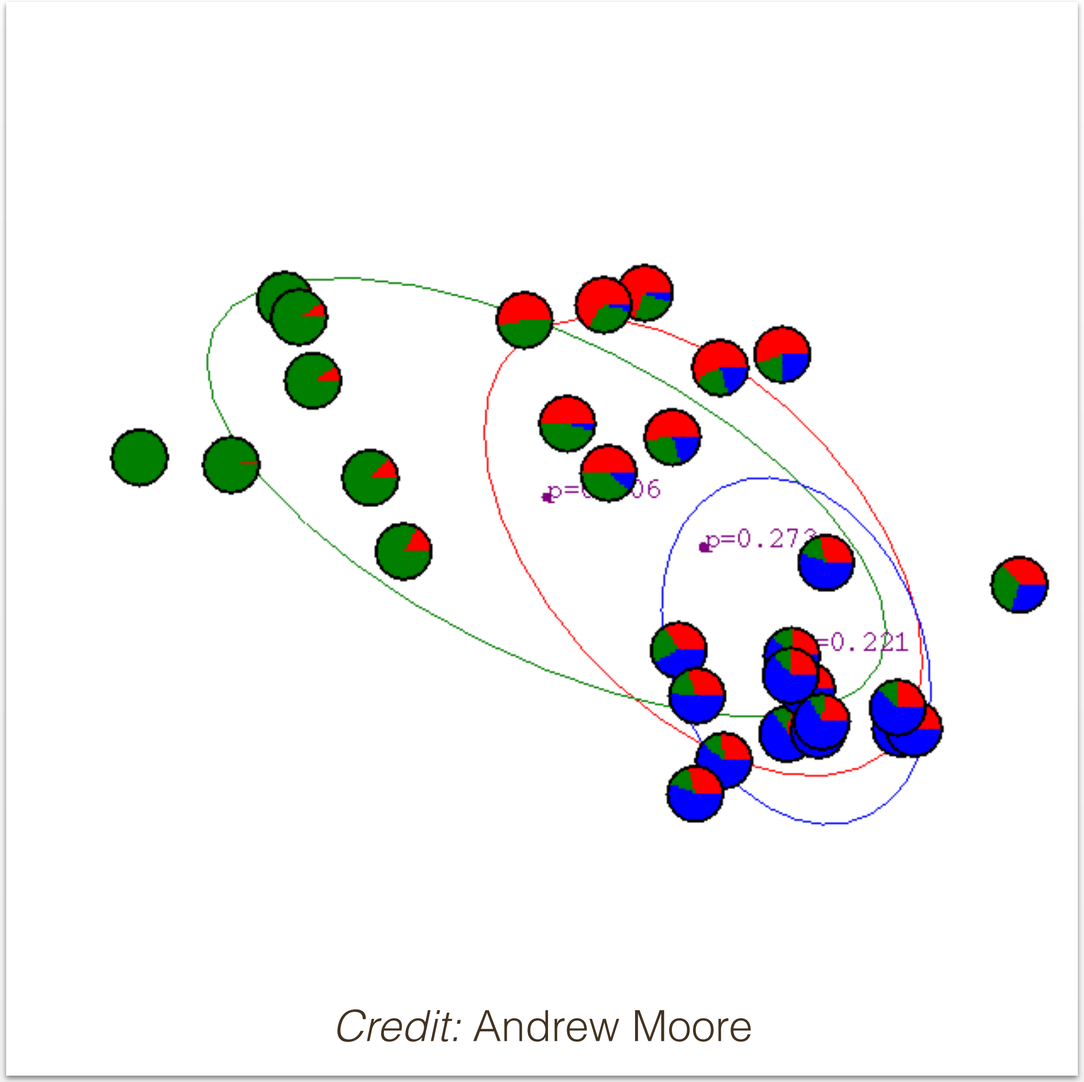
(posterior probability)

# Expectation Maximization



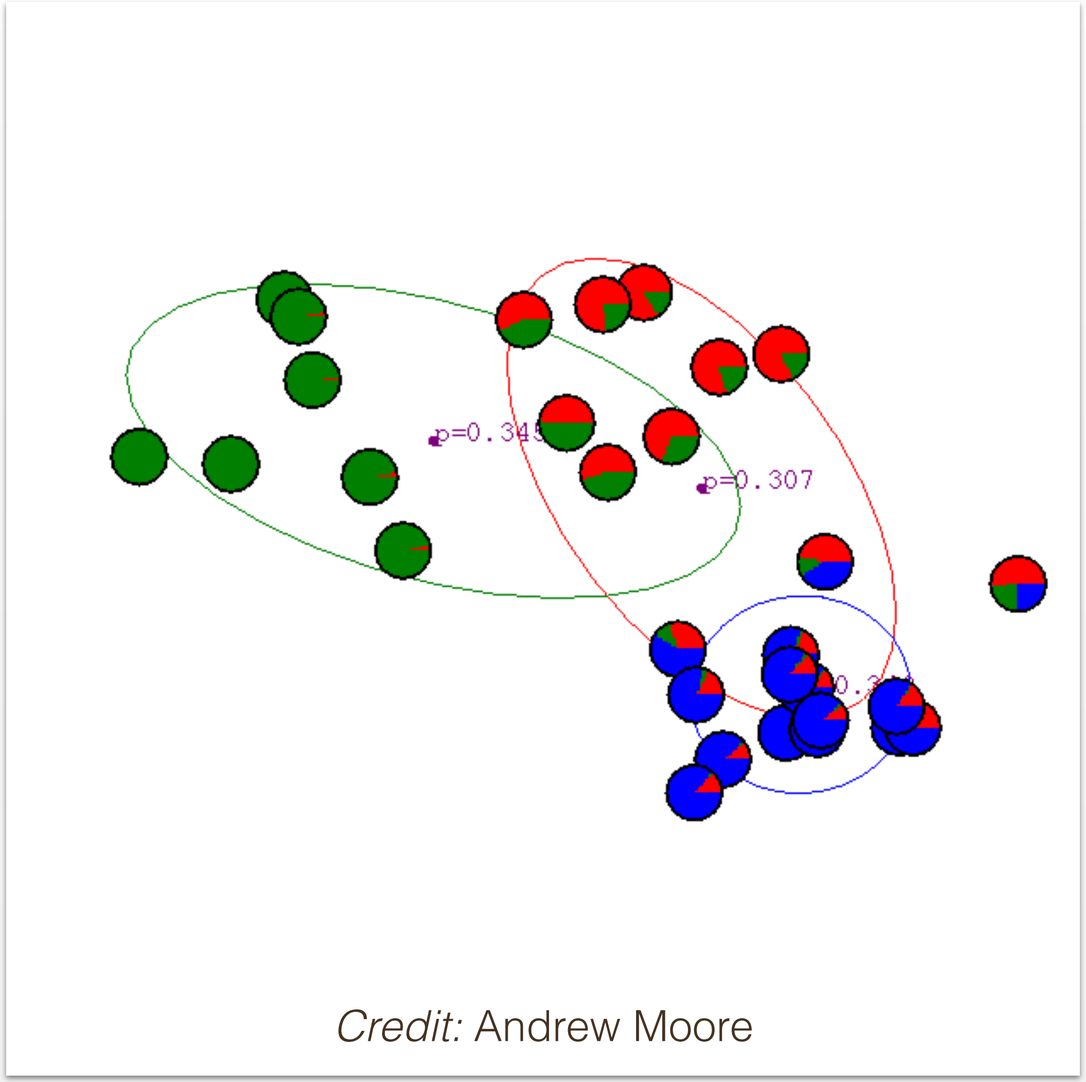

*Credit:* Andrew Moore

# Expectation Maximization



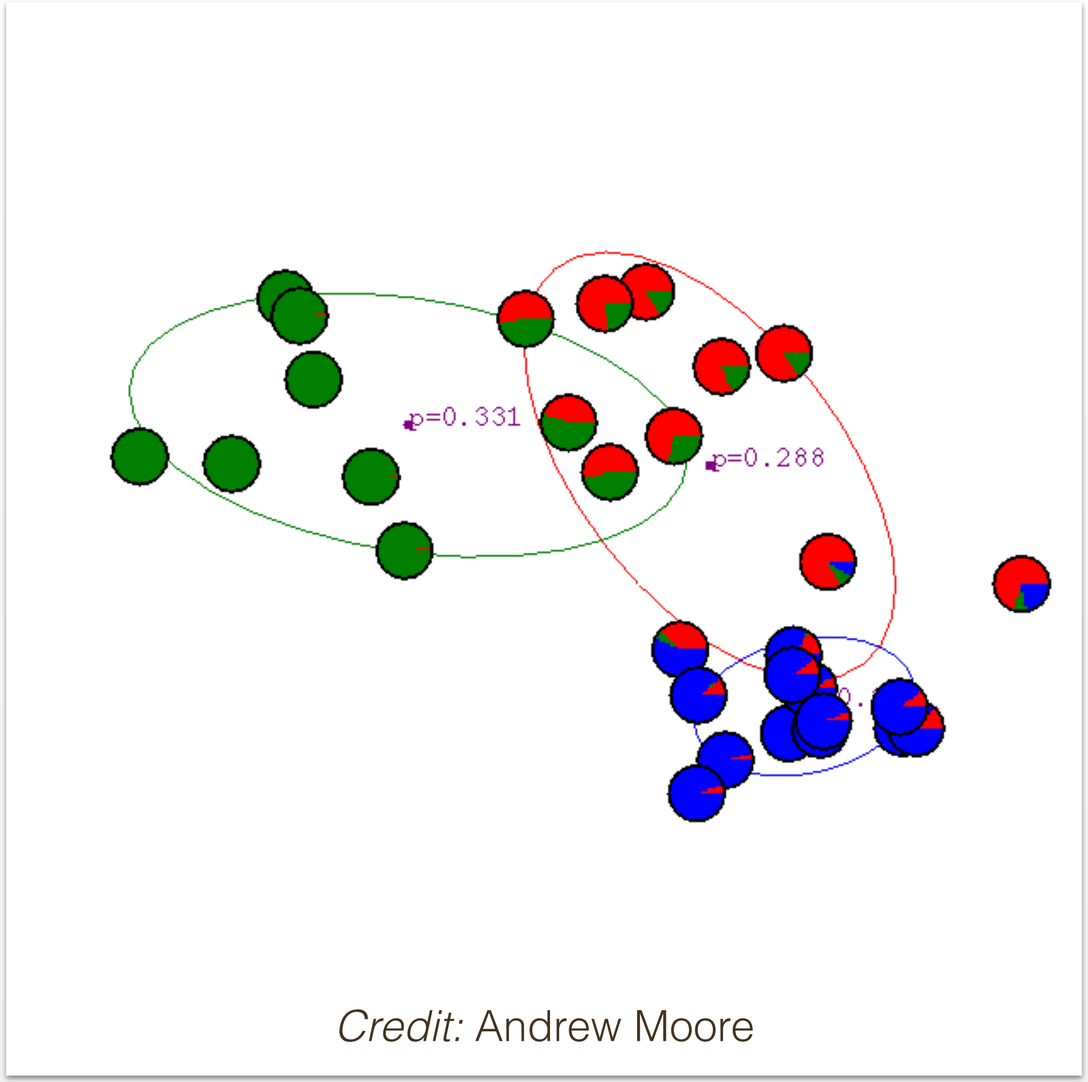

*Credit:* Andrew Moore

# Expectation Maximization



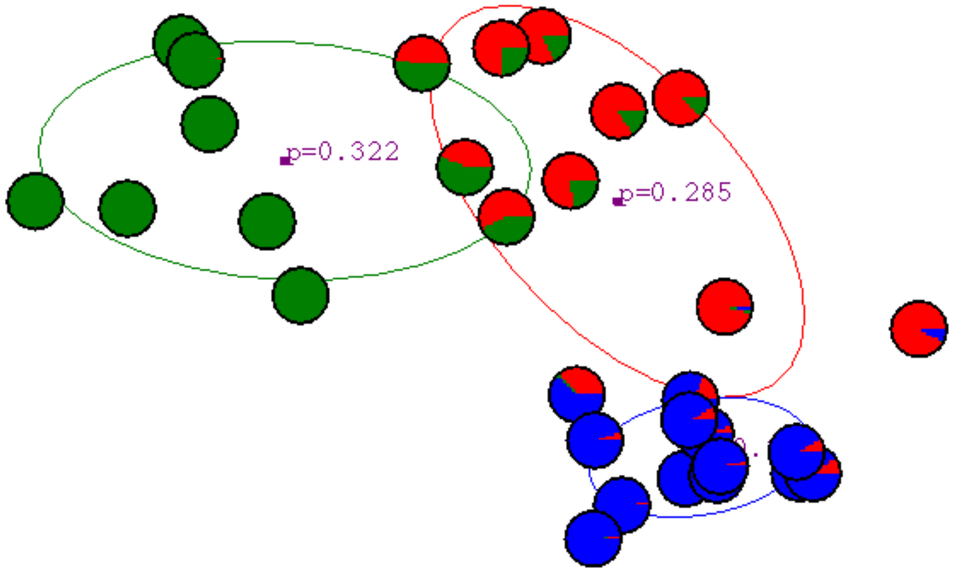

*Credit:* Andrew Moore

# Expectation Maximization



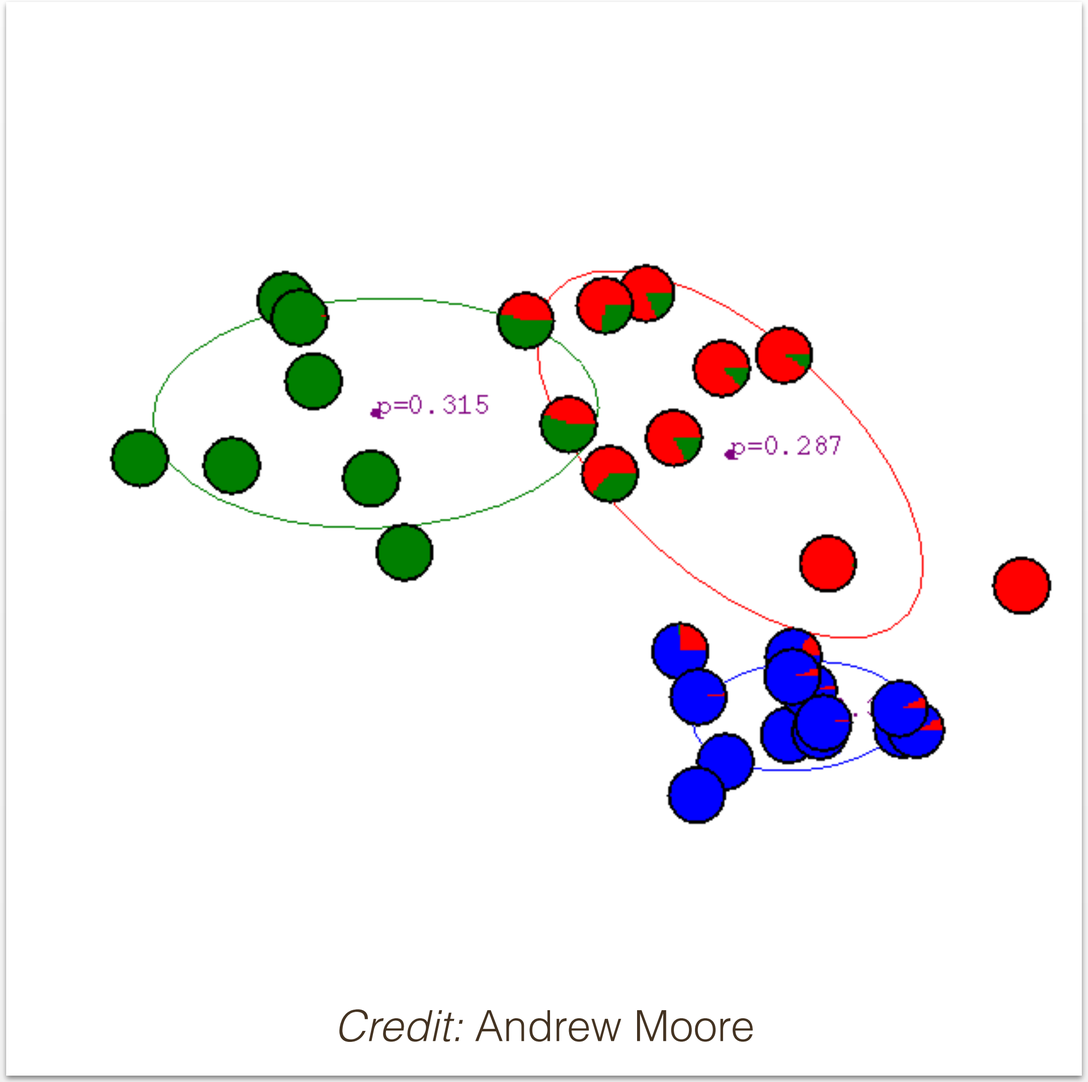

*Credit:* Andrew Moore

# Expectation Maximization



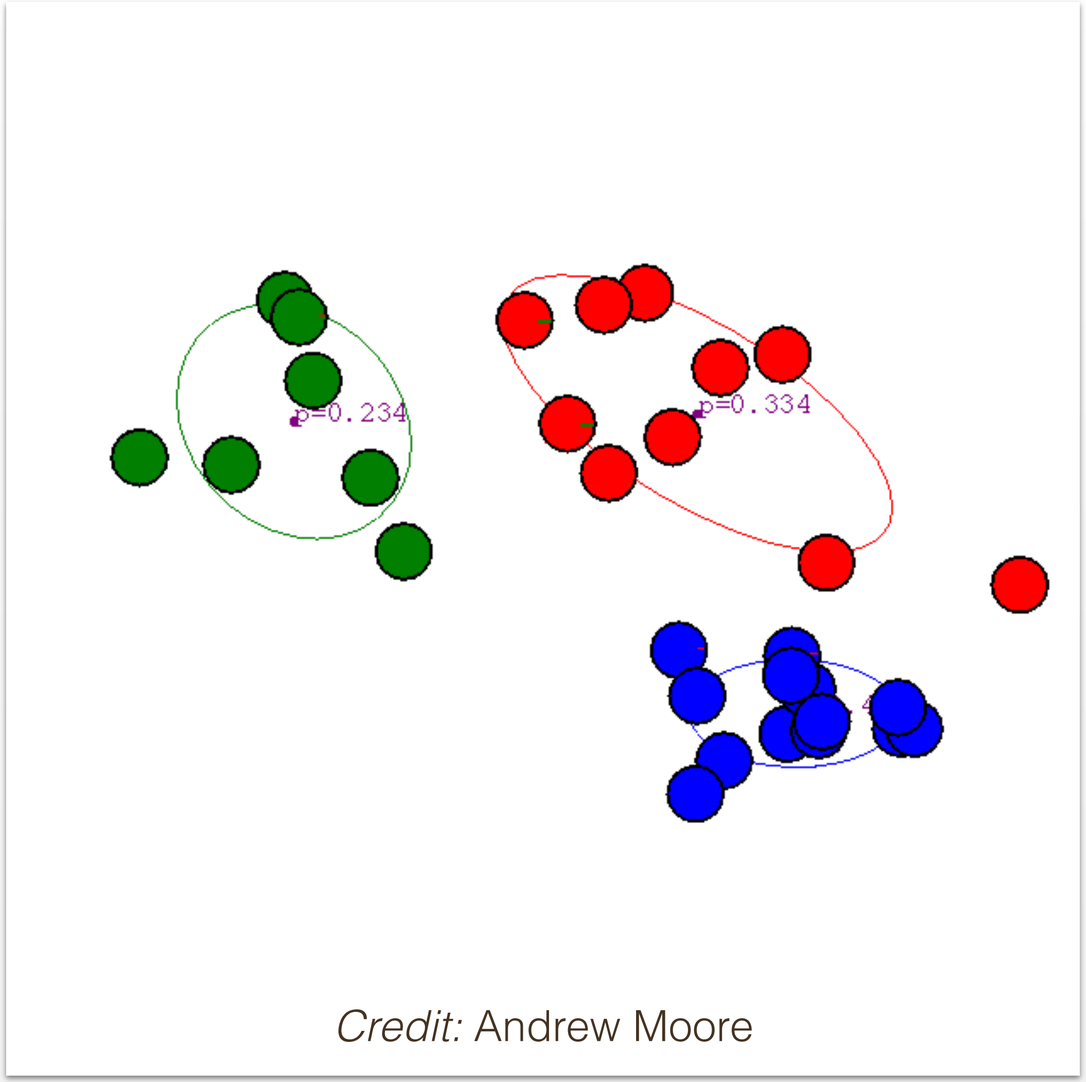

*Credit:* Andrew Moore

# Expectation Maximization



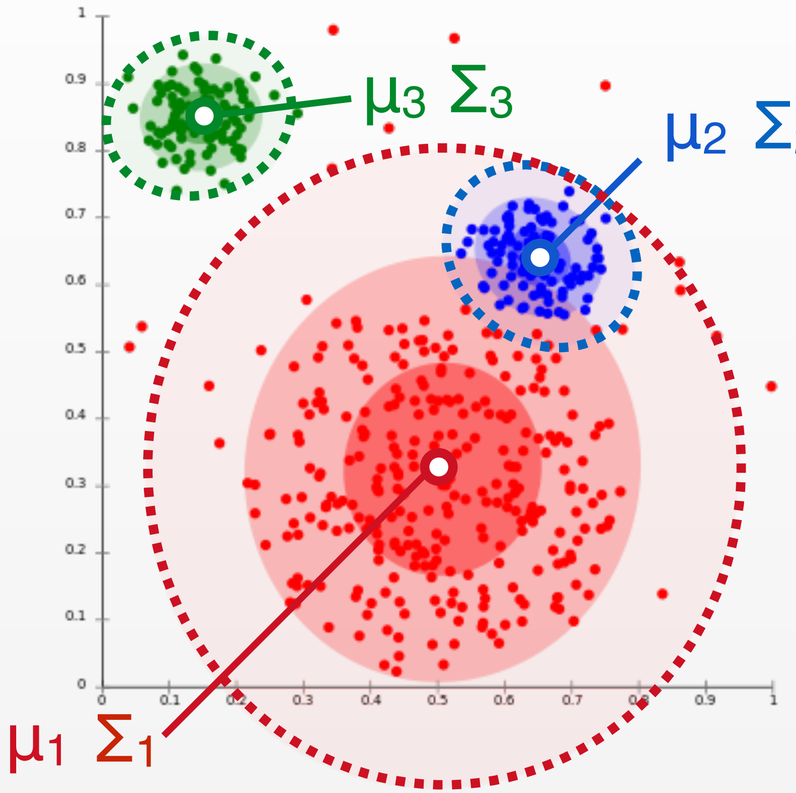

*Credit:* Andrew Moore

# Expectation Maximization




*Credit:* Andrew Moore

# Gaussian Mixture Models




$$x_n \mid z_n = k \sim \text{Norm}(\mu_k, \Sigma_k)$$

Each cluster is a multivariate Gaussian with unknown $\mu_k, \Sigma_K$

$$z_n \sim \text{Categorical}(\alpha_1, \alpha_2 \ldots \alpha_K)$$

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

Each cluster has approximately $\pi_k N$ points for some unknown $\pi_K$

# Gaussian Mixture Models

$$x_n \mid z_n = k \sim \text{Norm}(\mu_k, \Sigma_k)$$

$$p(\boldsymbol{x}_n \mid z_n = k) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_k|}} e^{-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)}$$

$$z_n \sim \text{Categorical}(\alpha_1, \alpha_2 \ldots \alpha_K)$$

$$p(z_n = k) = \alpha_k$$

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

# Idea 2: Learn *Soft* Assignments to Clusters

## Posterior on Cluster Assignments (from Bayes' Rule)

*Likelihood* *Prior*

$$\gamma_{nk} = p(z_n = k \mid \boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_n \mid z_n = k)p(z_n = k)}{p(\boldsymbol{x}_n)}$$

*Posterior* *Marginal Likelihood*

*Prior*

$$p(z_n = k) = \alpha_k$$

*Likelihood*

$$p(\boldsymbol{x}_n \mid z_n = k) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_k|}} e^{-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)}$$

*Marginal Likelihood*

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} p(\boldsymbol{x}_n \mid z_n = k)p(z_n = k)$$

# Idea 1: Learn *Gaussian* for Each Cluster

## Maximum Likelihood Estimation

$$\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\pi}^* = \underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}}{\operatorname{argmax}} \quad \log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$$

Idea: Use weights $\gamma_{nk} = p(z_n{=}k \mid x_n)$ to compute estimates

Notice similarity to K-Means →

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_n \gamma_{nk} \boldsymbol{x}_n \qquad N_k = \sum_n \gamma_{nk}$$

Cluster Mean

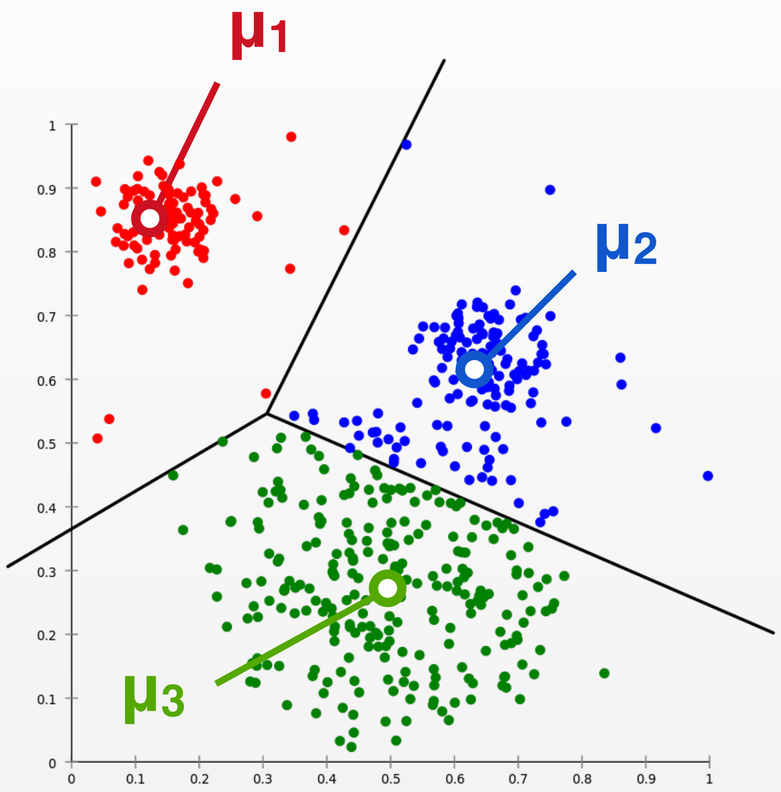
$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_n \gamma_{nk} (\boldsymbol{x}_n - \mu_k)(\boldsymbol{x}_n - \mu_k)^\top$$

Cluster Covariance

$$\pi_k = \frac{N_k}{N}$$

Fraction of points in each cluster

# Comparison: K-means Clustering




*Objective:* Sum of Squares

$$L(\boldsymbol{\mu}, \boldsymbol{z}) = \sum_{k=1}^{K} \sum_{n=1}^{N} I[z_n = k] (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^2$$

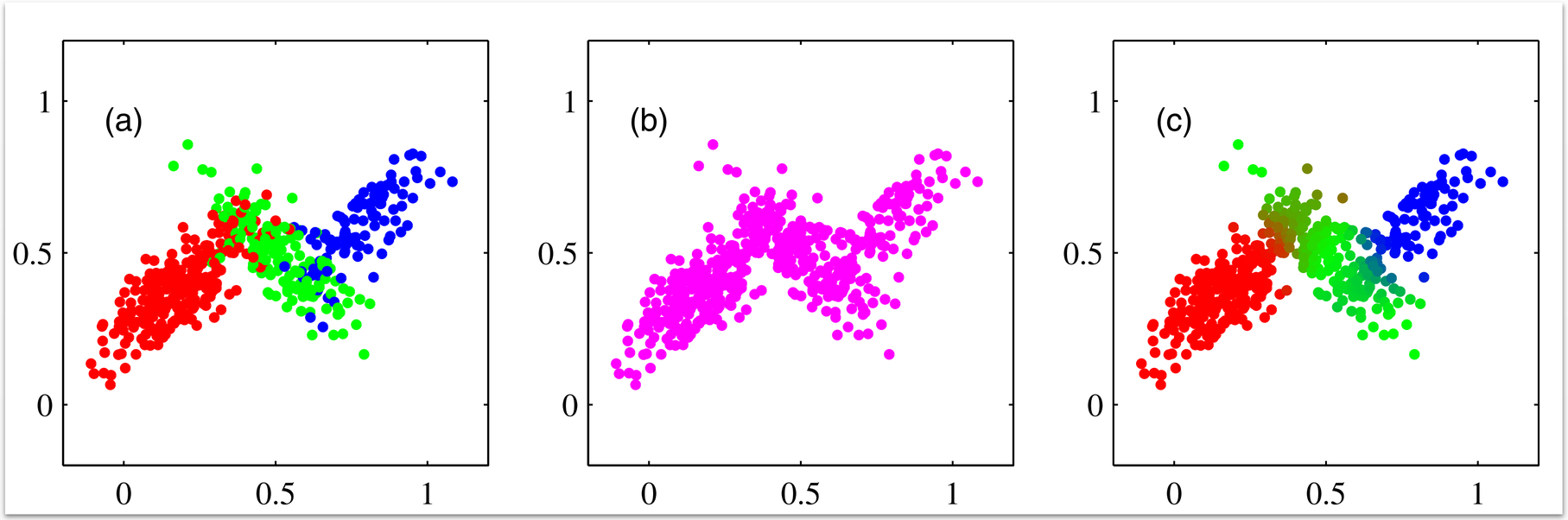
$I[z_n = k]$ One-hot assignment

$\boldsymbol{\mu}_k$ Center for cluster *k*

*Alternate between two steps*

1. Minimize loss w.r.t. $\boldsymbol{z}_n$
2. Minimize loss w.r.t. $\boldsymbol{\mu}_k$

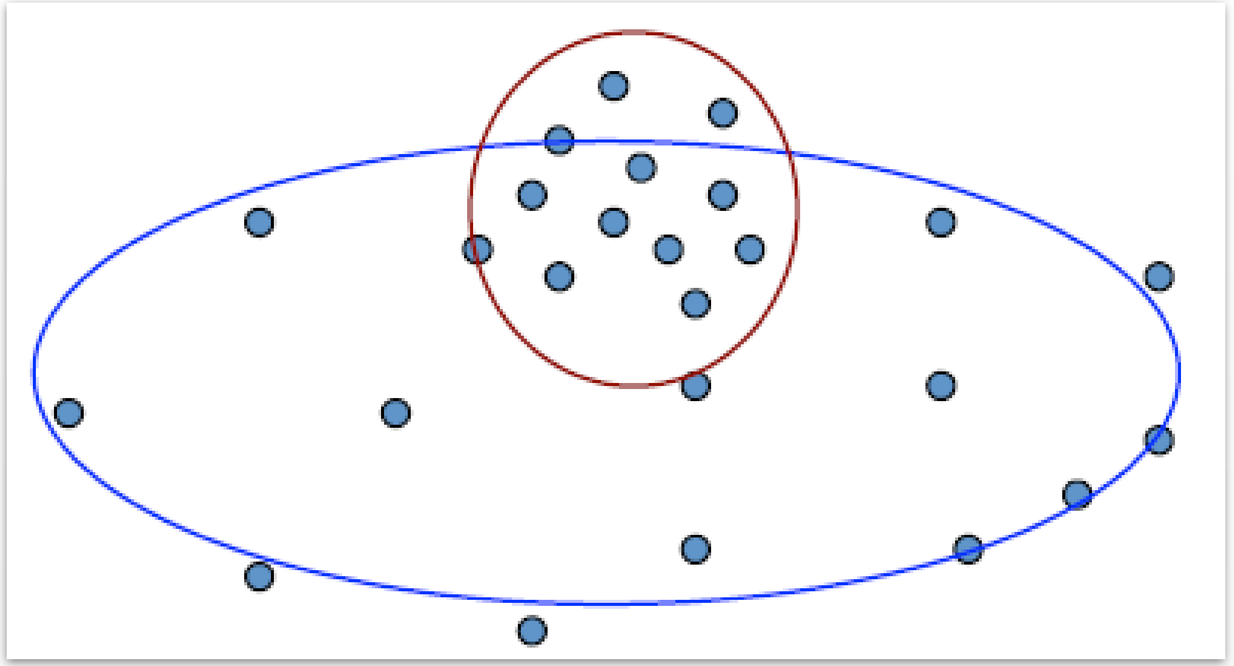
# GMM Advantages / Disadvantages




+ Works with overlapping clusters
+ Works with clusters of different densities
+ Same complexity as K-means

- Can get stuck in local maximum
- Need to set number of components

# GMM Advantages / Disadvantages



+ Works with overlapping clusters
+ Works with clusters of different densities
+ Same complexity as K-means

- Can get stuck in local maximum
- Need to set number of components


# Clustering

Shantanu


# Expectation Maximization

## Maximum Likelihood for Mixtures

# Maximum likelihood estimation

$X$: a dataset with $N$ points:

$$X = \{x_n\}_{n=1}^{N}$$

$x_n \in \mathcal{X}$: the space where $x_n$ takes values. Typically $\mathcal{X} = \mathbb{R}^D$

Model Assumption:

$$x_n \sim P_\theta$$

A distribution parametrized by $\theta$ (unknown) defined on the sample space the space $\mathcal{X}$

How do we estimate $\theta$ ?

# Maximum likelihood estimation

Idea: Find a $\theta$ that maximizes the probability of the observed data

$$\hat{\theta} = \text{argmax}_\theta \, L(\theta)$$

$$X = \{x_n\}_{n=1}^N$$

$$x_n \sim P_\theta$$

$$L(\theta) = p(X|\theta)$$

$$= \prod_{i=1}^N p(x_i|\theta)$$

Because $x_n \overset{\text{iid}}{\sim} P_\theta$

Independently and identically distributed

$p(x|\theta)$ is the density function of the distribution $P_\theta$

$p(X|\theta)$ is the density function of the distribution on $\mathcal{X}^N$, the sample space containing all datasets of size $N$. The probability of a dataset is derived from the probability of the points in the dataset.

# Maximum likelihood estimation

$$\hat{\theta} = \text{argmax}_{\theta}\ L(\theta)$$

$$X = \{x_n\}_{n=1}^{N}$$
$$x_n \sim P_{\theta}$$

## How do we optimize $L(\theta)$ w.r.t. $\theta$?

Idea1: Take derivative of $L(\theta)$ w.r.t. $\theta$ and equate it to $0$ to find the local maximum.

Idea 2: Use iterative gradient ascent or any other optimization method.

$$\theta^{k+1} \leftarrow \theta^{k} + \eta \nabla_{\theta} L(\theta)$$

Need to compute the gradients

$$\frac{d}{d\theta} L(\theta) = \frac{d}{d\theta} \prod_{n=1}^{N} p(x_n | \theta)$$
$$= \sum_{k=1}^{N} \prod_{n \neq k} p(x_n | \theta) \frac{d}{d\theta} p(x_k | \theta)$$

The derivative expression is complex. Which makes it difficult to get a closed form expression when equated to 0.
The product of N-1 probabilities would be a very small number which would lead to numerical issues in Gradient Ascent.

# Maximum likelihood estimation

Solution: Work with the log-likelihood

$$l(\theta) = \log L(\theta)$$

$$\hat{\theta} = \text{argmax}_\theta \, L(\theta)$$

$$= \text{argmax}_\theta \, l(\theta)$$

Because $\log$ is a monotonically increasing function

$$X = \{x_n\}_{n=1}^N$$

$$x_n \sim P_\theta$$

$$L(\theta) = \prod_{i=1}^N p(x_i \mid \theta)$$

Simpler and stable gradient computation

$$\frac{d}{d\theta} l(\theta) = \frac{d}{d\theta} \log \left[ \prod_{n=1}^N p(x_n \mid \theta) \right]$$

$$= \frac{d}{d\theta} \sum_{n=1}^N \log p(x_n \mid \theta)$$

$$= \sum_{n=1}^N \frac{d}{d\theta} \log p(x_n \mid \theta)$$

# Maximum likelihood estimation

$$x_n \sim \text{Norm}(\mu, \sigma) \qquad \theta = (\mu, \sigma)$$

How to find the maximum likelihood estimate of $\mu$?

$$\begin{aligned} l(\mu, \sigma) &= \sum_{n=1}^{N} \log p(x_n \mid \mu, \sigma) \\ &= \sum_{n=1}^{N} \log \left[ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n - \mu)^2}{\sigma^2}} \right] \\ &= \sum_{n=1}^{N} \left[ -\frac{1}{2} \log 2\pi\sigma^2 - \frac{(x_n - \mu)^2}{\sigma^2} \right] \end{aligned}$$

Compute gradient w.r.t. $\mu$

$$\frac{d}{d\mu} l(\mu, \sigma) = \sum_{n=1}^{N} 2\frac{(x_n - \mu)}{\sigma^2}$$

Solving $\frac{d}{d\mu} l(\mu, \sigma) = 0$ gives

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

# GMM model

$$x_n \mid z_n = k \sim \text{Norm}(\mu_k, \Sigma_k)$$

$$z_n \sim \text{Categorical}(\alpha_1, \alpha_2, \ldots \alpha_K)$$

$$p(x_n \mid z_n = k, \theta) = f_k(x_n)$$

$$p(z_n = k \mid \theta) = \alpha_k$$

$$p(x_n \mid \theta) = \sum_{i=1}^{K} p(x_n \mid z_n = k, \theta) p(z_n = k \mid \theta)$$

$$= \sum_{i=1}^{K} \alpha_k f_k(x_n)$$

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

$$\theta = [\alpha_1, \mu_1, \Sigma_1, \ldots \alpha_K, \mu_K, \Sigma_K]$$

$$f(x \mid \mu, \Sigma) = \frac{\exp\left( -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right)}{(2\pi)^{-\frac{D}{2}} \sqrt{|\Sigma|}}$$

# GMM model

$$x_n \mid z_n = k \sim \text{Norm}(\mu_k, \Sigma_k)$$

$$z_n \sim \text{Categorical}(\alpha_1, \alpha_2, \ldots \alpha_K)$$

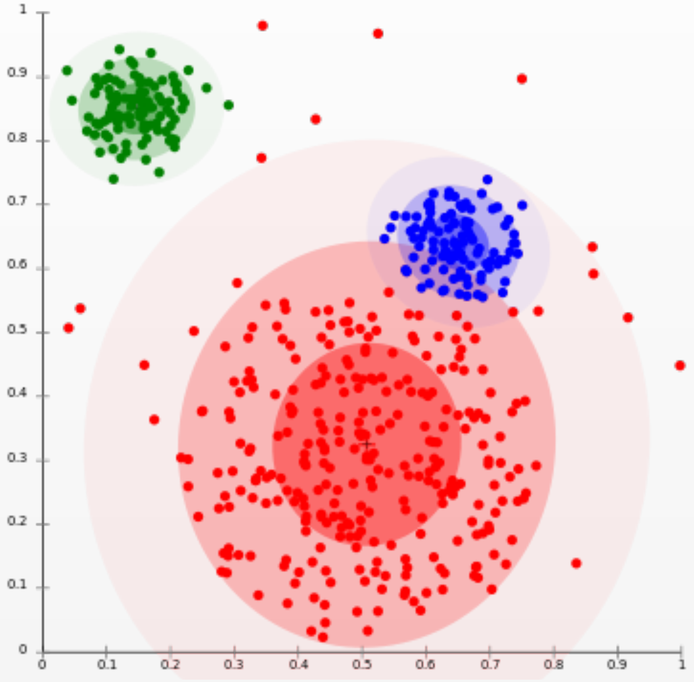
$$\theta = [\alpha_1, \mu_1, \Sigma_1, \ldots \alpha_K, \mu_K, \Sigma_K]$$

$$\begin{aligned} l(\theta) &= \sum_{n=1}^{N} \log p(x_n \mid \theta) \\ &= \sum_{n=1}^{N} \log\Big( \sum_{k=1}^{K} \alpha_k f(x_n \mid \mu_k, \Sigma_k)\Big) \end{aligned}$$

Gradient relatively hard to compute.

$$f(x \mid \mu, \Sigma) = \frac{\exp\left( -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)}{(2\pi)^{-\frac{D}{2}} \sqrt{|\Sigma|}}$$

# Review: Gaussian Mixture Models




*Soft* Assignment Update

$$\gamma_{nk} := p(z_n = k \mid \boldsymbol{x}_n, \boldsymbol{\theta})$$

Parameter Updates (*Mean and Covariance*)

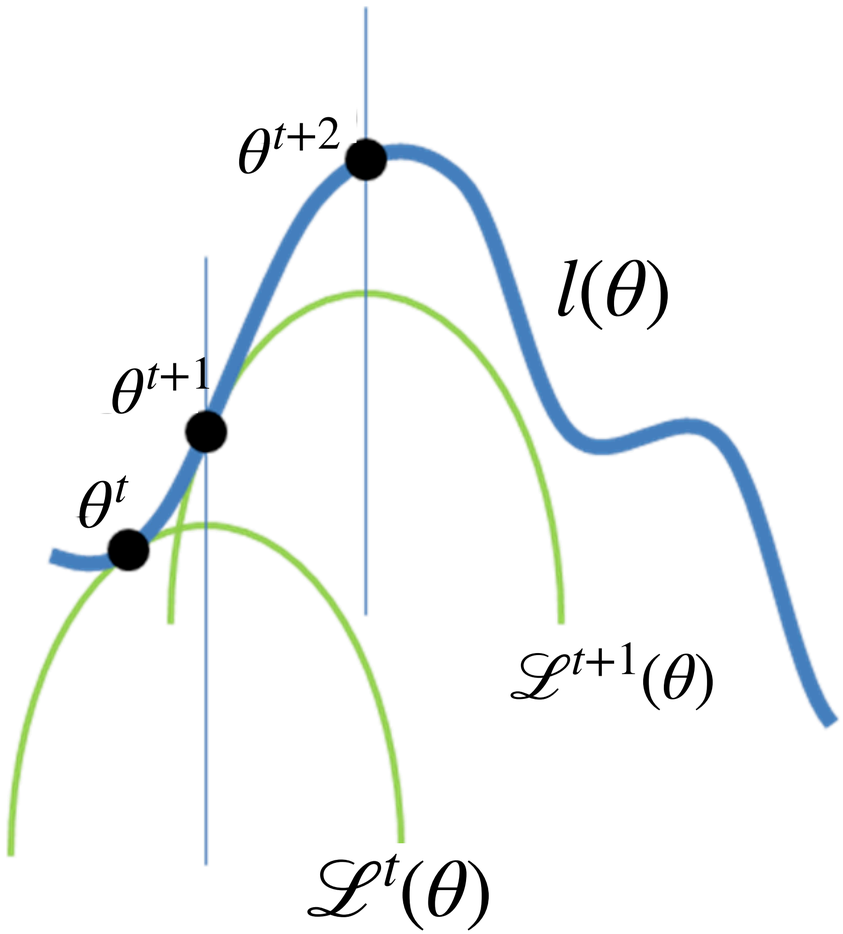
$$N_k := \sum_{n=1}^{N} \gamma_{nk}$$

$$\boldsymbol{\pi} = (N_1/N, \ldots, N_K/N)$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk}\, \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk}\, (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top$$

# EM-Algorithm




$$\mathscr{L}^{t}(\theta) \leq l(\theta)$$

$$\mathscr{L}^{t}(\theta^{t}) = l(\theta^{t})$$

# Objective: Lower Bound on Log Likelihood

$$
\begin{aligned}
l(\theta) &= \log p(X|\theta) \\
&= \mathbf{E}_{q^t(Z)} \log p(X|\theta) \\
&= \mathbf{E}_{q^t(Z)} \log \left[ \frac{p(X,Z|\theta)}{p(Z|X,\theta)} \right] \\
&= \mathbf{E}_{q^t(Z)} \log \left[ \frac{p(X,Z|\theta)}{p(Z|X,\theta)} \frac{q^t(Z)}{q^t(Z)} \right] \\
&= \mathbf{E}_{q^t(Z)} \log \left[ \frac{p(X,Z|\theta)}{q^t(Z)} \right] + \mathbf{E}_{q^t(Z)} \log \left[ \frac{q^t(Z)}{p(Z|X,\theta)} \right]
\end{aligned}
$$

$$X = [x_1, x_2, \ldots, x_N]$$

$$Z = [z_1, z_2, \ldots, z_N]$$

$$q^t(Z) = p(Z|X,\theta^t)$$

$\theta^t$ is the estimate of $\theta^*$ (the true parameter) in the $t^{th}$ iteration

Evidence lower bound: $\mathscr{L}^t(\theta)$

$\mathrm{KL}\left(q^t(Z) \,\|\, p(Z|X,\theta)\right) \geq 0$

$$l(\theta) \geq \mathscr{L}^t(\theta)$$

# Objective: Lower Bound on Log Likelihood

$$l(\theta) = \mathscr{L}^t(\theta) + \mathrm{KL}\left(q^t(Z) \parallel p(Z|X,\theta)\right)$$

$$l(\theta^t) = \mathscr{L}^t(\theta^t) + \mathrm{KL}\left(q^t(Z) \parallel p(Z|X,\theta^t)\right)$$

$$= \mathscr{L}^t(\theta^t)$$

$$\mathscr{L}^t(\theta) = \mathbf{E}_{q^t(Z)} \log\left[\frac{p(X,Z|\theta)}{q^t(Z)}\right]$$

$$q^t(Z) = p(Z|X,\theta^t)$$

$$\mathrm{KL}\left(q^t(Z) \parallel p(Z|X,\theta)\right)$$

$$= \mathbf{E}_{q^t(Z)} \log\left[\frac{q^t(Z)}{p(Z|X,\theta)}\right]$$

$$l(\theta) - l(\theta^t) = \left(\mathscr{L}^t(\theta) - \mathscr{L}^t(\theta^t)\right) + \mathrm{KL}\left(q^t(Z) \parallel p(Z|X,\theta)\right)$$

Because KL-divergence $\geq 0$

$$\mathscr{L}^t(\theta) \geq \mathscr{L}^t(\theta^t) \Rightarrow l(\theta) \geq l(\theta^t)$$

# Objective: Lower Bound on Log-Likelihood

Complete data log-likelihood: $\log P(X, Z \mid \theta)$. The log-likelihood of observed and unobserved variables.

$$\begin{aligned}\mathscr{L}^t(\theta) &= \mathbf{E}_{q^t(Z)} \log p(X, Z \mid \theta) - \mathbf{E}_{q^t(Z)} \log q^t(Z) \\ &= \mathscr{Q}^t(\theta) - \mathbf{E}_{q^t(Z)} \log q^t(Z) \\ &\overset{\theta}{=} \mathscr{Q}^t(\theta)\end{aligned}$$

equal unto a constant w.r.t. $\theta$

$\mathscr{Q}^t(\theta)$ is the expectation of the complete log-likelihood function $p(X, Z \mid \theta)$ w.r.t. the distribution of $Z$ given $X$ under $\theta^t$.

$$\mathscr{L}^t(\theta) = \mathbf{E}_{q^t(Z)} \log \left[ \frac{p(X, Z \mid \theta)}{q^t(Z)} \right]$$

$$q^t(Z) = p(Z \mid X, \theta^t)$$

$$\mathrm{KL}\left(q^t(Z) \| \, p(Z \mid X, \theta)\right)$$

$$= \mathbf{E}_{q^t(Z)} \log \left[ \frac{q^t(Z)}{p(Z \mid X, \theta)} \right]$$

$$\mathscr{Q}^t(\theta) \geq \mathscr{Q}^t(\theta^t) \Rightarrow \mathscr{L}^t(\theta) \geq \mathscr{L}^t(\theta^t) \Rightarrow l(\theta) \geq l(\theta^t)$$

# EM-Algorithm

- Initialize $\theta^0$
- Repeat until convergence
  - E-step: Construct $\mathcal{Q}^t(\theta)$ by taking expectation of the complete log-likelihood function $p(X, Z \mid \theta)$ w.r.t. the distribution of $Z$ given $X$ under $\theta^t$.
  - M-step: $\theta^{t+1} \leftarrow \text{argmax}_\theta \mathcal{Q}^t(\theta)$

$$\mathcal{L}^t(\theta) = \mathbf{E}_{q^t(Z)} \log \left[ \frac{p(X, Z \mid \theta)}{q^t(Z)} \right]$$

$$q^t(Z) = p(Z \mid X, \theta^t)$$

$$\mathcal{Q}^t(\theta) = \mathbf{E}_{q^t(Z)} \log p(X, Z \mid \theta)$$

# The Kullback-Leibler Divergence

## KL Divergence

$$KL(q \,||\, p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

## Properties

- KL(q || p) ≥ 0
- If KL(q || p) = 0, then q = p
- KL(q || p) ≠ KL(p || q)

# The Kullback-Leibler Divergence

## KL Divergence

$$KL(q \,||\, p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

## Properties

- KL(q || p) ≥ 0
- If KL(q || p) = 0, then q = p
- KL(q || p) ≠ KL(p || q)

## Jensen's Inequality

For *convex* functions f(x)

$$\sum_x p(x) f(x) \geq f\left(\sum_x p(x) x\right)$$

For *concave* functions f(x)

$$\sum_x p(x) f(x) \leq f\left(\sum_x p(x) x\right)$$

# The Kullback-Leibler Divergence

## KL Divergence

$$KL(q \,||\, p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

## Properties

- KL(q || p) ≥ 0
- If KL(q || p) = 0, then q = p
- KL(q || p) ≠ KL(p || q)

## Proof

$$\begin{aligned}
-\mathrm{KL}(q \,||\, p) &= -\sum_x q(x) \log \frac{q(x)}{p(x)} \\
&= \sum_x q(x) \log \frac{p(x)}{q(x)} \\
&\leq \log \sum_x q(x) \frac{p(x)}{q(x)} \\
&= \log \sum_x p(x) \\
&= \log 1 = 0
\end{aligned}$$

# EM for GMM

$$
\begin{aligned}
p(X, Z \mid \theta) &= \prod_{n=1}^{N} p(x_n, z_n \mid \theta) \\
&= \prod_{n=1}^{N} p(x_n \mid z_n, \theta) p(z_n \mid \theta) \\
&= \prod_{n=1}^{N} \alpha_{z_n} f(x_n \mid \mu_{z_n}, \Sigma_{z_n}) \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} \left[ \alpha_k f(x_n \mid \mu_k, \Sigma_k) \right]^{\mathbf{I}[z_n = k]}
\end{aligned}
$$

$$
p(x_n \mid z_n = k, \theta) = f_k(x_n \mid \mu_k, \Sigma_k)
$$

$$
p(z_n = k \mid \theta) = \alpha_k
$$

$$
f(x \mid \mu, \Sigma) = \frac{\exp\left( -\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right)}{(2\pi)^{-\frac{D}{2}} \sqrt{|\Sigma|}}
$$

$$
\begin{aligned}
\log p(X, Z \mid \theta) &= \sum_{n=1}^{N} \sum_{k=1}^{K} \log \left[ \alpha_k f(x_n \mid \mu_k, \Sigma_k) \right]^{\mathbf{I}[z_n = k]} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbf{I}[z_n = k] \left[ \log \alpha_k + \log f(x_n \mid \mu_k, \Sigma_k) \right]
\end{aligned}
$$

# EM for GMM

$$\log p(X, Z \mid \theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbf{I}[z_n = k] \left[ \log \alpha_k + \log f(x_n \mid \mu_k, \Sigma_k) \right]$$

$$\begin{aligned}
\mathcal{Q}^t(\theta) &= \mathbf{E}_{q^t(Z)} \left[ \log p(X, Z \mid \theta) \right] \\
&= \mathbf{E}_{q^t(Z)} \left[ \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbf{I}[z_n = k] \left[ \log \alpha_k + \log f(x_n \mid \mu_k, \Sigma_k) \right] \right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbf{E}_{(z_n \mid x_n, \theta^t)} \left[ \mathbf{I}[z_n = k] \right] \left[ \log \alpha_k + \log f(x_n \mid \mu_k, \Sigma_k) \right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_n = k \mid x_n, \theta^t) \left[ \log \alpha_k + \log f(x_n \mid \mu_k, \Sigma_k) \right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk}^t \left[ \log \alpha_k + \log f(x_n \mid \mu_k, \Sigma_k) \right]
\end{aligned}$$

E-step requires computing $\gamma_{nk}^t$, the soft cluster assignments with $\theta^t$.

$$f(x \mid \mu, \Sigma) = \frac{\exp\left( -\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) \right)}{(2\pi)^{-\frac{D}{2}} \sqrt{|\Sigma|}}$$

# GMM model

$$\mathcal{Q}^t(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk}^t \left[ \log \alpha_k + \log f(x_n \mid \mu_k, \Sigma_k) \right]$$

M-Step 1: maximize $\mathcal{Q}^t(\theta)$ w.r.t. $\alpha_k$

- Solve $\nabla_{\alpha_k} \mathcal{Q}^t(\theta) = 0$ under the constraint $\sum_{k=1}^{K} \alpha_k = 1$

$$\alpha_k^{t+1} \leftarrow \frac{\sum_{n=1}^{N} \gamma_{nk}^t}{N}$$

M-Step 2: maximize $\mathcal{Q}^t(\theta)$ w.r.t. $\mu_k$

- Solve $\nabla_{\mu_k} \mathcal{Q}^t(\theta) = 0$

$$\mu_k^{t+1} \leftarrow \frac{\sum_{n=1}^{N} \gamma_{nk}^t x_n}{\alpha^{t+1} N}$$

M-Step 3: maximize $\mathcal{Q}^t(\theta)$ w.r.t. $\Sigma_k$

- Solve $\nabla_{\Sigma_k} \mathcal{Q}^t(\theta) = 0$

$$\Sigma_k^{t+1} \leftarrow \frac{\sum_{n=1}^{N} \gamma_{nk}^t (x_n - \mu_k^t)^T (x_n - \mu_k^t)}{\alpha^{t+1} N}$$

# EM for GMM model

$$\mathcal{Q}^t(\theta) = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}^t\left[\log\alpha_k + \log f(x_n\,|\,\mu_k,\Sigma_k)\right]$$

$$\overset{\mu_k}{=} -\sum_{n=1}^{N}\frac{\gamma_{nk}^t}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)$$

$$f(x\,|\,\mu,\Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}{(2\pi)^{-\frac{D}{2}}\sqrt{|\,\Sigma\,|}}$$

$$\nabla_{\mu_k}\mathcal{Q}^t(\theta) = 0$$

$$\Rightarrow -\sum_{n=1}^{N}\frac{\gamma_{nk}^t}{2}\left[\nabla_{\mu_k}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)\right] = 0$$

$$\Rightarrow \sum_{n=1}^{N}\gamma_{nk}^t\Sigma_k^{-1}(x_n-\mu_k) = 0$$

$$\Rightarrow \sum_{n=1}^{N}\gamma_{nk}^t(x_n-\mu_k) = 0$$

$$\Rightarrow (\sum_{n=1}^{N}\gamma_{nk}^t)\mu_k = \sum_{n=1}^{N}\gamma_{nk}^t x_n$$

$$\Rightarrow \mu_k = \frac{\sum_{n=1}^{N}\gamma_{nk}^t x_n}{\sum_{n=1}^{N}\gamma_{nk}^t}$$

# Model Selection

$$p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n | z_n, \boldsymbol{\theta}) p(z_n \mid \boldsymbol{\theta})$$

**Need to specify two components**

1. Likelihood

   $p(\boldsymbol{x}_n | z_n, \boldsymbol{\theta})$

2. Mixture distribution (i.e. number of clusters)

   $p(z_n \mid \boldsymbol{\theta})$

How do we know that we have made “good” choices?

# Model Selection


**Strategy 1: Cross-validation**

Split data in to folds.

*For each fold*

- Perform EM to learn $\theta$ from training set $X^{\text{train}}$
- Calculate test set likelihood $p(X^{\text{test}} \mid \theta)$