# Clustering: Unsupervised Learning

## The $k$-means clustering algorithm

In the clustering problem, we are given a training set $\{x^{(1)}, \ldots, x^{(m)}\}$, and want to group the data into a few cohesive "clusters." Here, $x^{(i)} \in \mathbb{R}^n$ as usual; but no labels $y^{(i)}$ are given. So, this is an unsupervised learning problem.

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.

2. Repeat until convergence: {

   For every $i$, set

   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each $j$, set

   $$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

   }

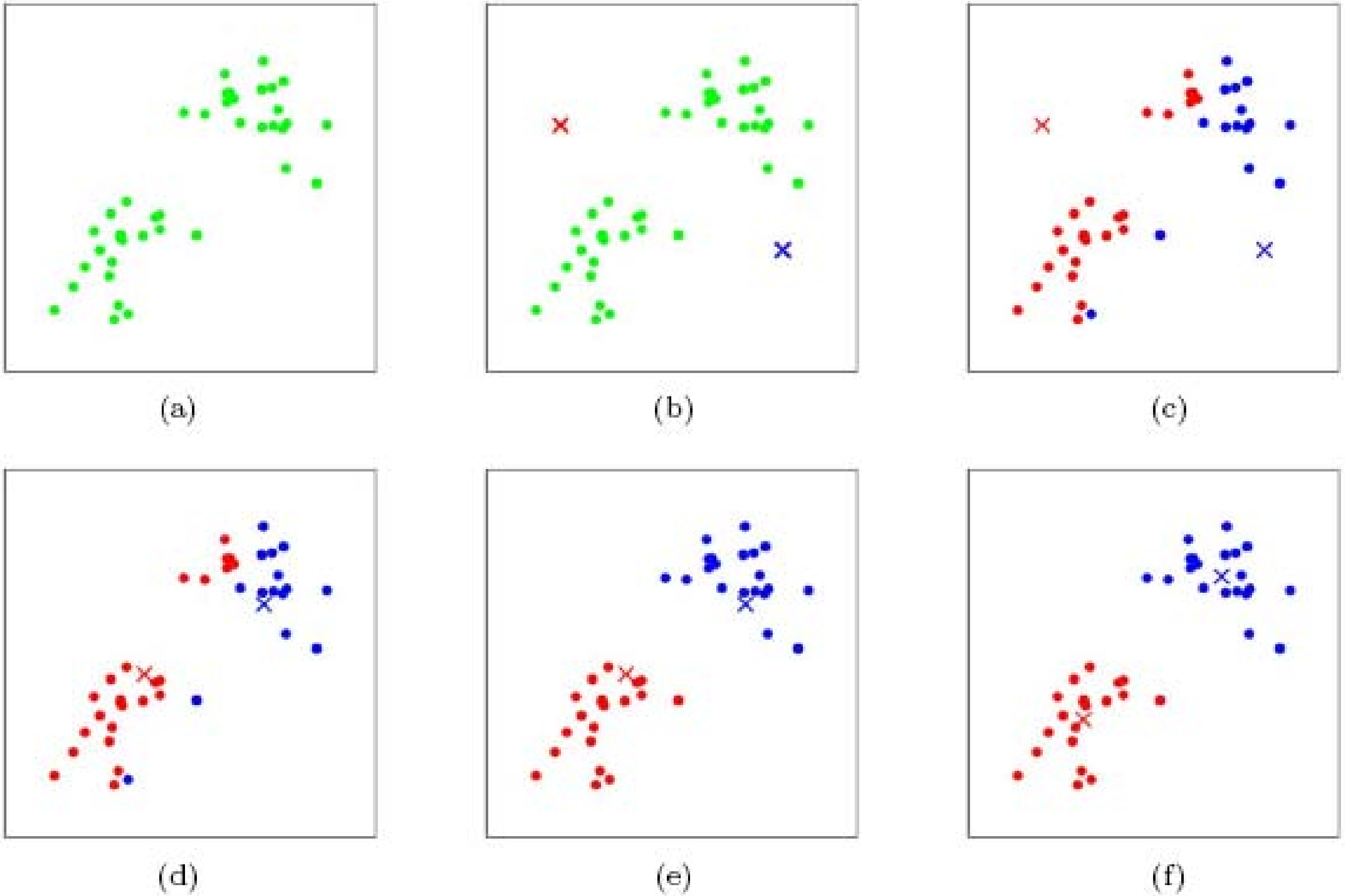
# K-means Clustering - continued

k is the number of clusters we want to find and is a parameter of the algorithm

$\mu_j$ is the current centroid of cluster *j* .It is the current best guess for the position of the center of the cluster

The inner-loop of the algorithm repeatedly carries out two steps: (i) "Assigning" each training example $x^{(i)}$ to the closest cluster centroid $\mu_j$, and (ii) Moving each cluster centroid $\mu_j$ to the mean of the points assigned to it.

# K-means in Action

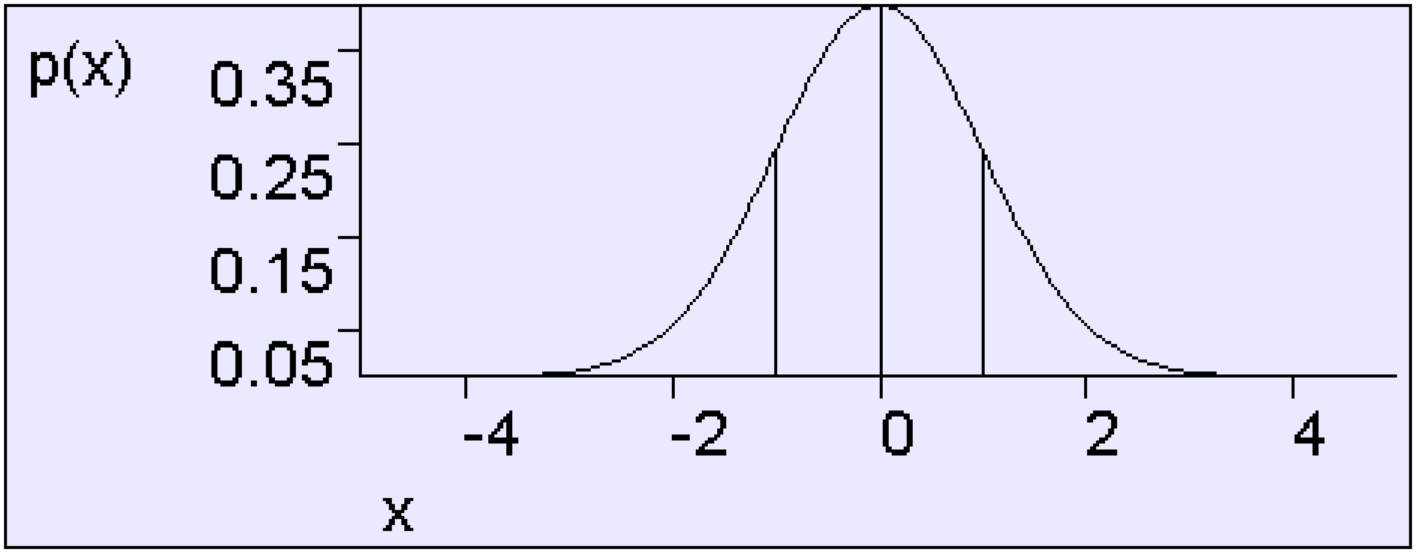



Training examples are shown as dots, cluster centroids are shown as crosses.

# Gaussians – A Quick Review

# Unit variance Gaussian

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{x^2}{2} \right)$$




$$E[X] = 0$$

$$\mathrm{Var}[X] = 1$$

$$H[X] = -\int_{x=-\infty}^{\infty} p(x) \log p(x) dx = 1.4189$$

# Bivariate Gaussians

Write r.v. $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$ Then define $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to mean

$$p(\mathbf{x}) = \frac{1}{2\pi \|\boldsymbol{\Sigma}\|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Where the Gaussian's parameters are...

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2{}_y \end{pmatrix}$$

Where we insist that $\Sigma$ is symmetric non-negative definite

# Bivariate Gaussians

Write r.v. $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$ Then define $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to mean

$$p(\mathbf{x}) = \frac{1}{2\pi \,||\, \boldsymbol{\Sigma} \,||^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \ \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Where the Gaussian's parameters are...

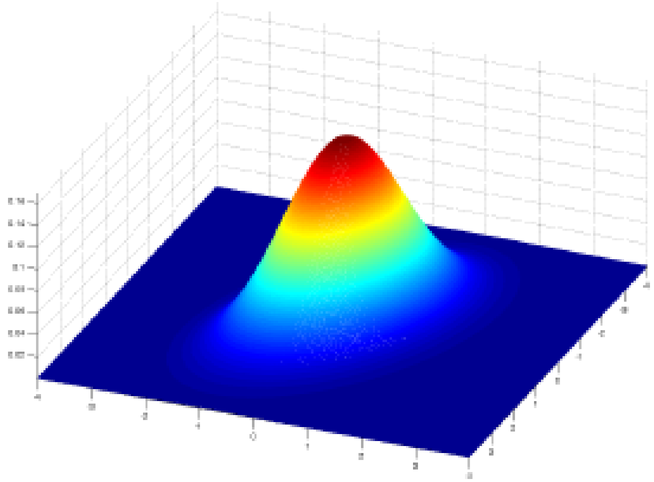
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2{}_y \end{pmatrix}$$

It turns out that E[X] = $\mu$ and Cov[X] = $\Sigma$.

# Multivariate Gaussian distributions

- Gaussian distribution of a random vector $\mathbf{x}$ in $\mathbb{R}^d$:

$$\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

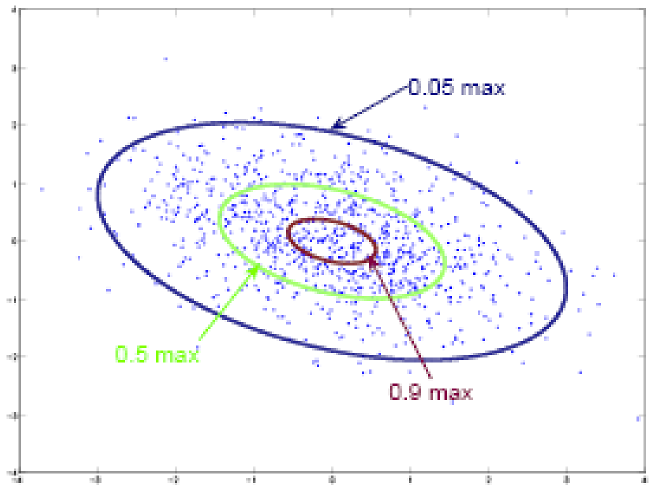


- The $\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}$ factor ensures it's a pdf (integrates to one).

# Multivariate Gaussians: intuition

$$\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

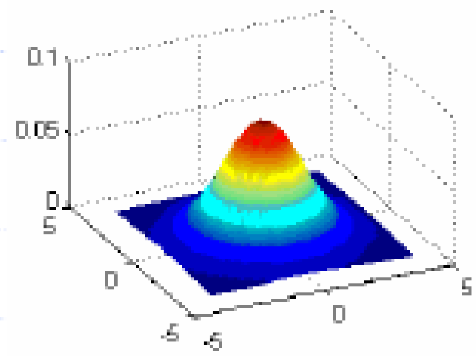
- This is the joint density of $x_1, \ldots, x_d$.
- density falls off exponentially as a function of distance to the mean $\|\mathbf{x}-\mu\|$;
- the *covariance matrix* $\mathbf{\Sigma}$ determines the shape of the density;



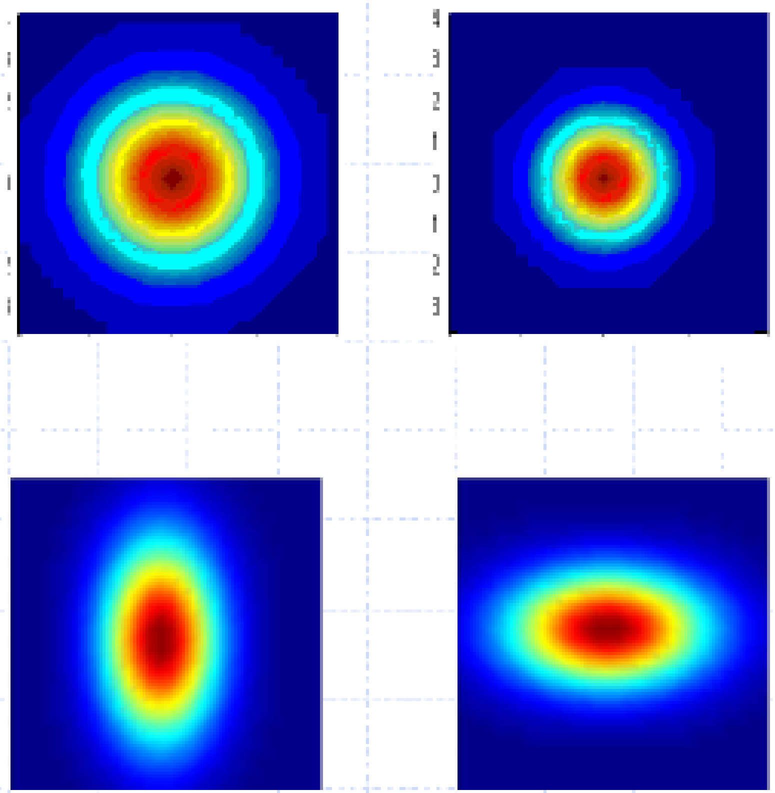

- The higher $d$ the faster $p(\mathbf{x})$ falls off.
- The determinant $|\mathbf{\Sigma}|$ measures the "spread" (analogous to $\sigma^2$).

- Spherical:

$$\Sigma = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$




Other Shapes:



# Clustering: Unsupervised Learning

we are given a training set $\{x^{(1)}, \ldots, x^{(m)}\}$, $x^{(i)} \in \mathbb{R}^n$ **,** No Labels with the x's

model the data by specifying a joint distribution $p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$

z's missing labels and parameter $\phi_j$ gives $p(z^{(i)} = j)$ $\sum_{j=1}^{k} \phi_j = 1$

$x^{(i)}|z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$ , K is number of values that the $z^{(i)}$'s can take on

This is called the **mixture of Gaussians** model.

$z^{(i)}$'s are **latent** random variables, meaning that they're hidden/unobserved.

The parameters of our model are thus $\phi$, $\phi$ and $\Sigma$.

Assuming we know the labels:

likelihood = $$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)}|z^{(i)}; \mu, \Sigma) + \log p(z^{(i)}; \phi)$$

# MLE for Gaussian Mixture Model

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)}|z^{(i)}; \mu, \Sigma) + \log p(z^{(i)}; \phi)$$

Maximizing this with respect to $\phi$, $\mu$ and $\Sigma$ gives the parameters:

$$\begin{aligned}
\phi_j &= \frac{1}{m}\sum_{i=1}^{m} 1\{z^{(i)} = j\}, \\
\mu_j &= \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}, \\
\Sigma_j &= \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}
\end{aligned}$$

# EM for Gaussian Mixture Model

(E-step) For each $i, j$, set

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

(M-step) Update the parameters:

$$\begin{aligned}
\phi_j &:= \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}, \\
\mu_j &:= \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}, \\
\Sigma_j &:= \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}
\end{aligned}$$

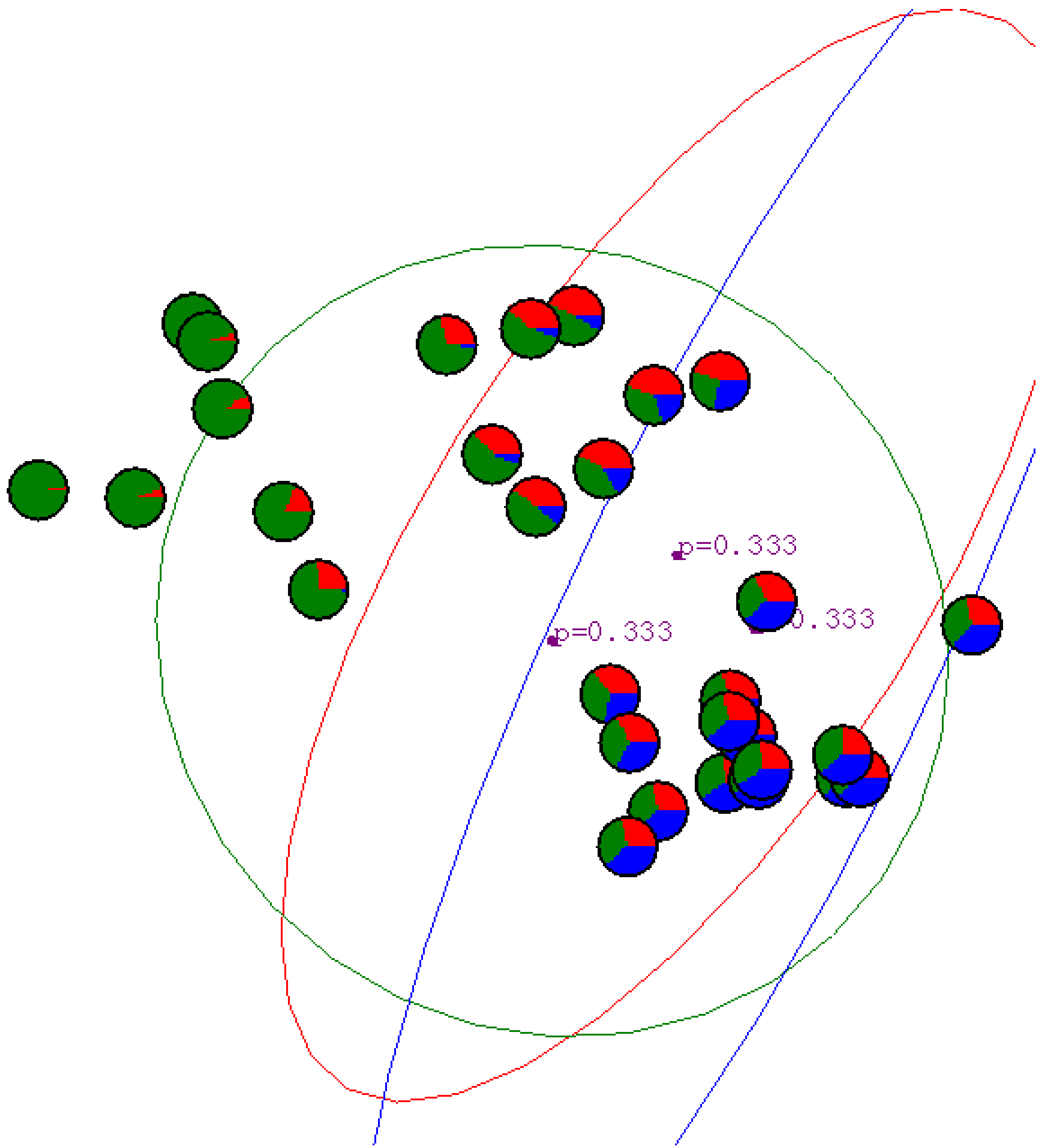
# EM Continued

In the E-step, we calculate the posterior probability of our parameters the $z^{(i)}$'s, given the $x^{(i)}$ and using the current setting of our parameters. I.e., using Bayes rule, we obtain:

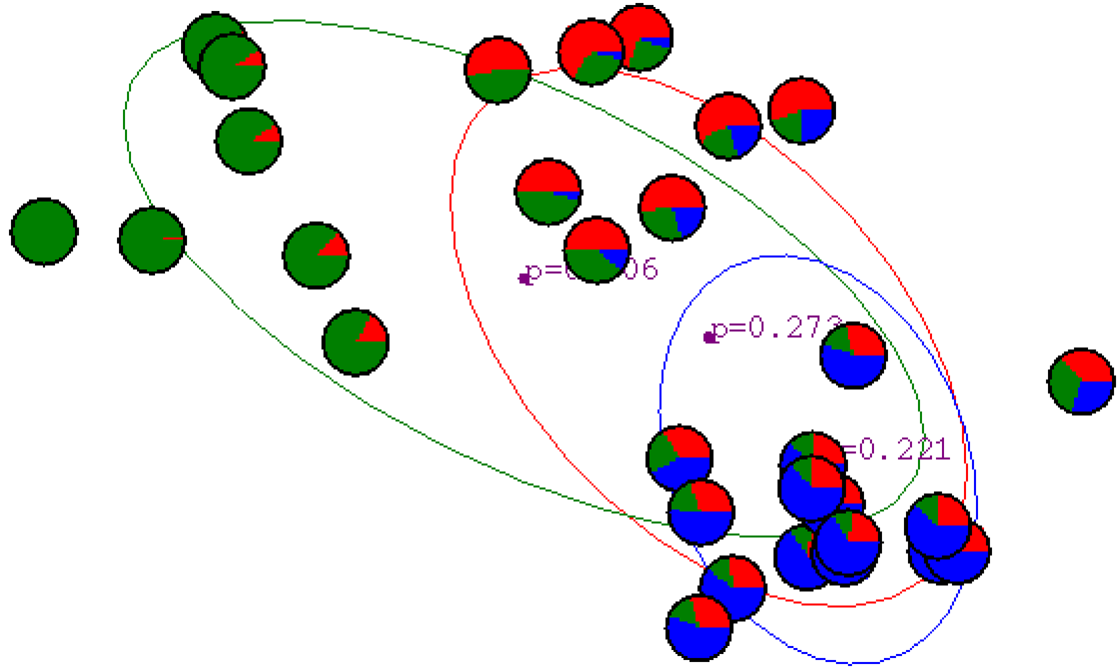
$$p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)} | z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}$$

Here, $p(x^{(i)} | z^{(i)} = j; \mu, \Sigma)$ is given by evaluating the density of a Gaussian with mean $\mu_j$ and covariance $\Sigma_j$ at $x^{(i)}$; $p(z^{(i)} = j; \phi)$ is given by $\phi_j$, and so on. The values $w_j^{(i)}$ calculated in the E-step represent our "soft" guesses for the values of $z^{(i)}$

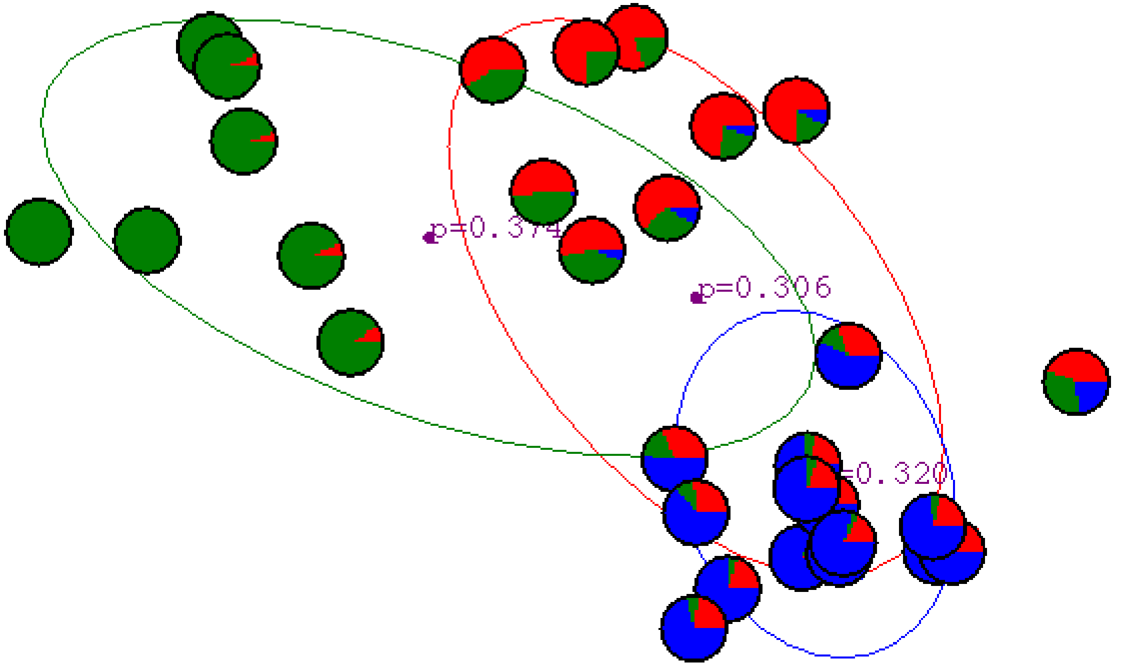
# Gaussian Mixture Example Start



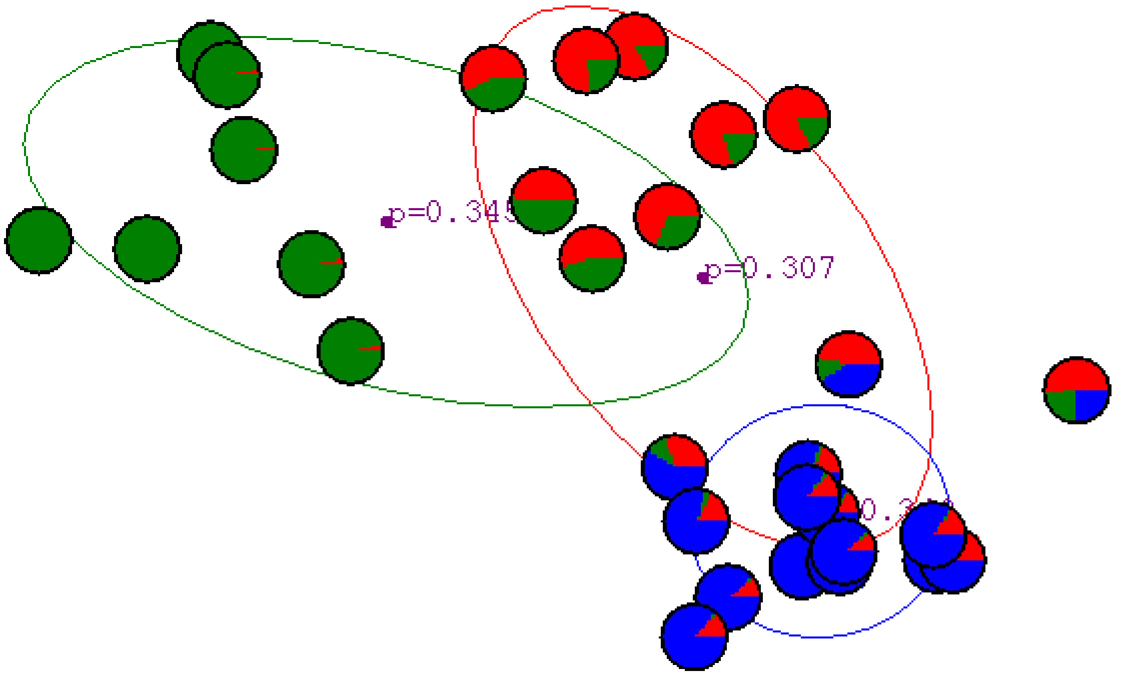
# After first iteration



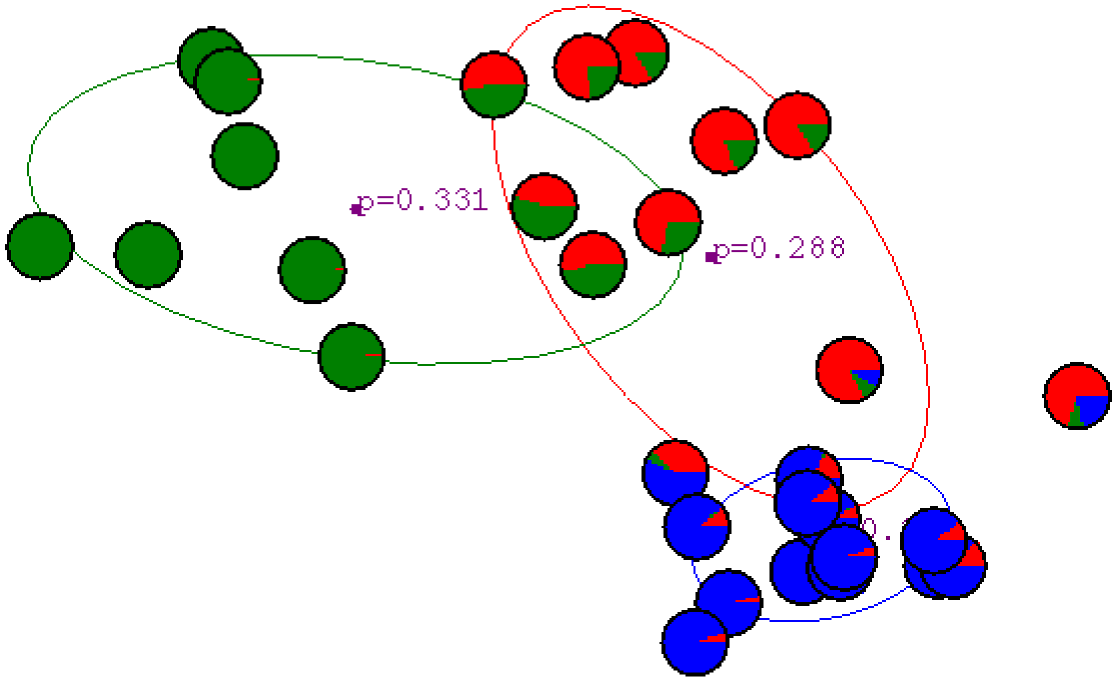
# After 2nd iteration



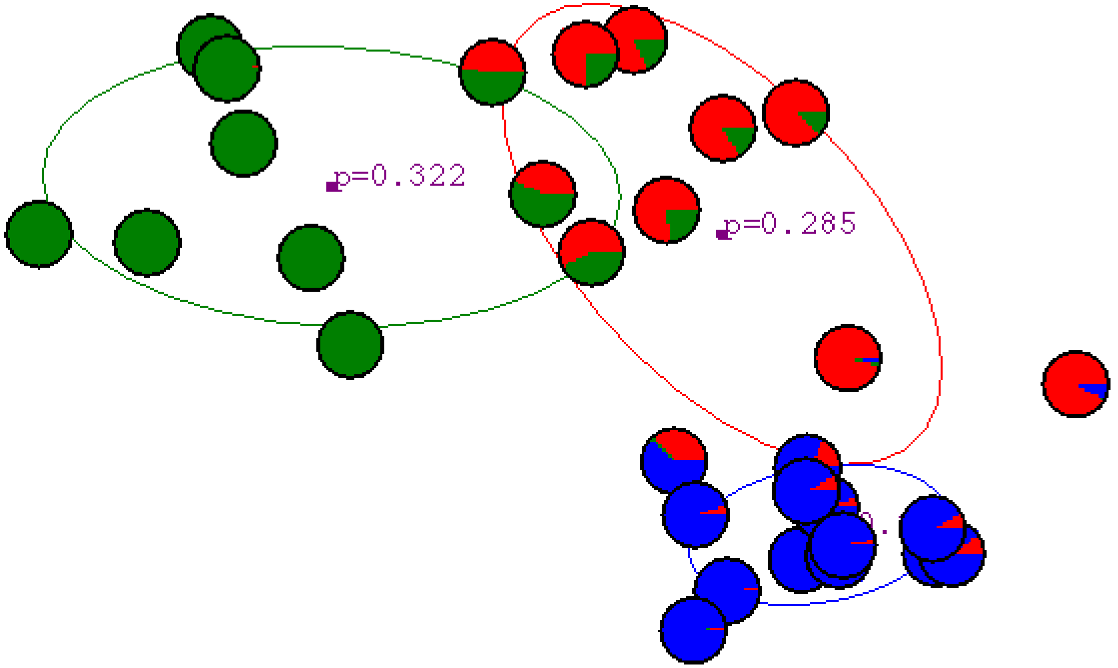
# After 3rd iteration



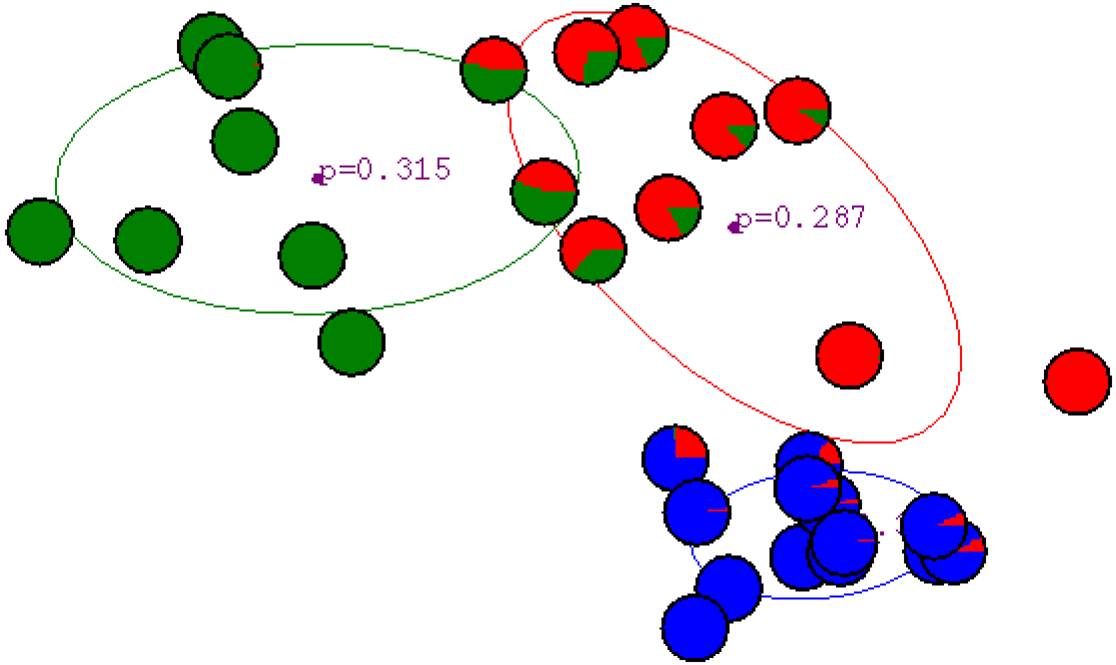
# After 4th iteration



# After 5th iteration



# After 6th iteration



# After 20th iteration