Mixture Models & EM algorithm
Probabilistic Graphical Models, Lecture 13

David Sontag
New York University

Slides adapted from Carlos Guestrin, Dan Klein, Luke Zettlemoyer, Dan Weld, Vibhav Gogate, and Andrew Moore

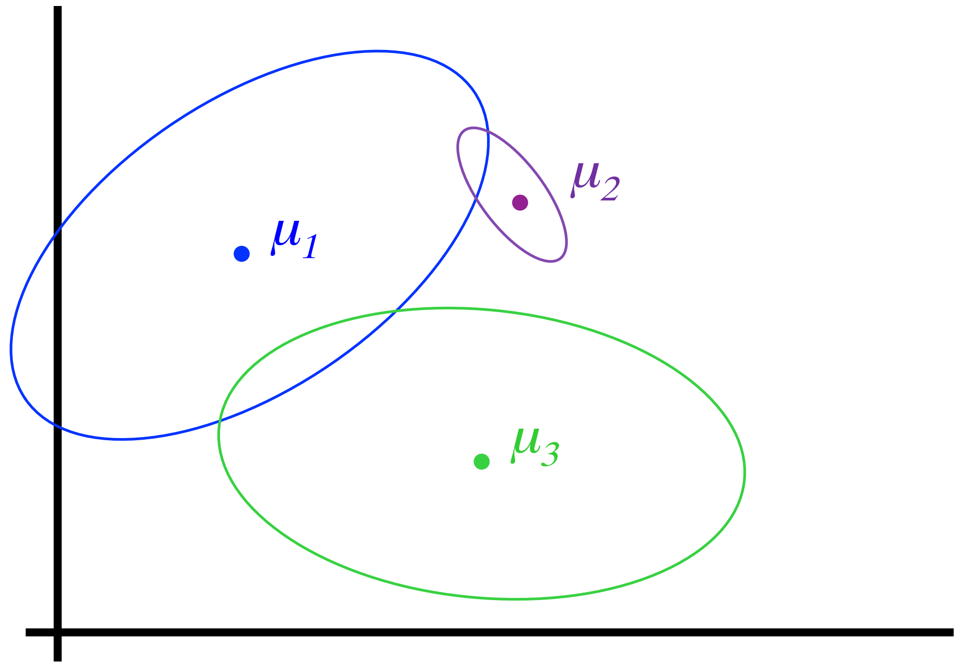
Mixture of Gaussians model

- P(Y): There are k components
- P(X|Y): Each component generates data from a **multivariate Gaussian** with mean μ_i and covariance matrix Σ_i

Each data point is sampled from a ***generative process***:

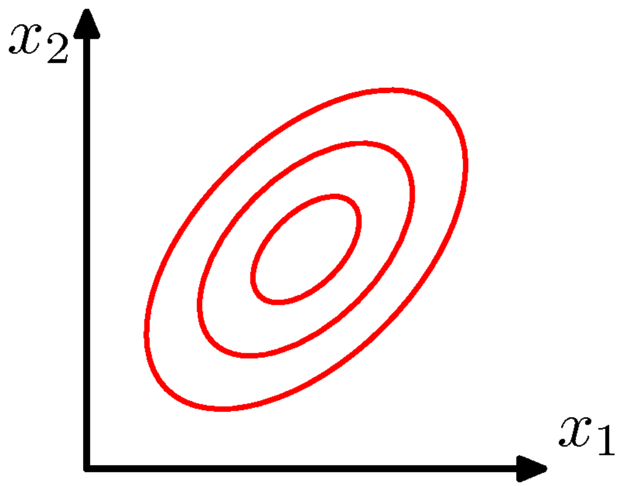
1. Choose component i with probability $P(y=i)$
2. Generate datapoint ~ $N(m_i, \Sigma_i)$

Gaussian mixture model (GMM)



Multivariate Gaussians

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

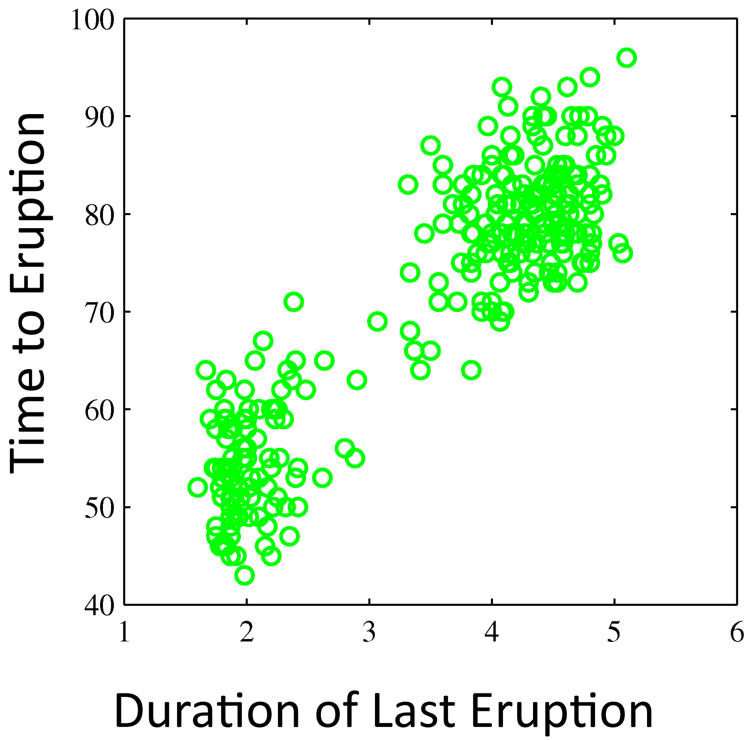


Σ = arbitrary (semidefinite) matrix:

- specifies rotation (change of basis)
- eigenvalues specify relative elongation

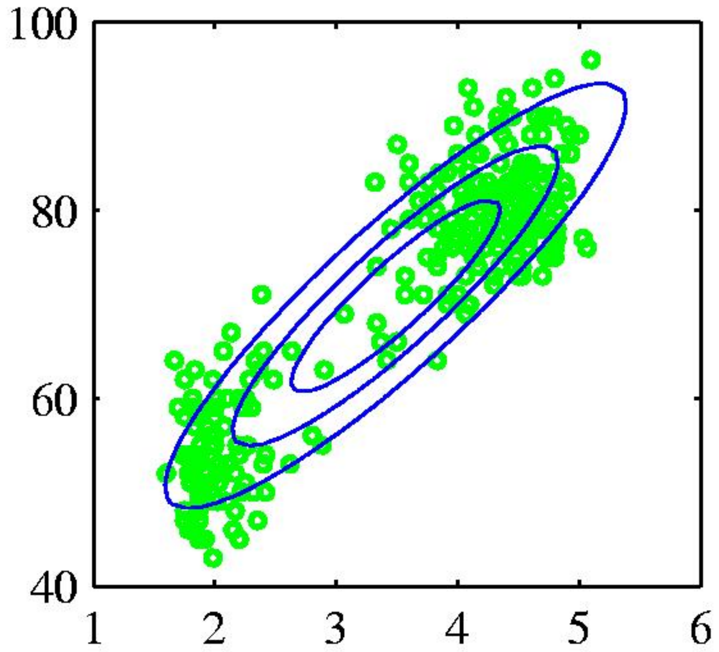
Mixtures of Gaussians (1)

Old Faithful Data Set

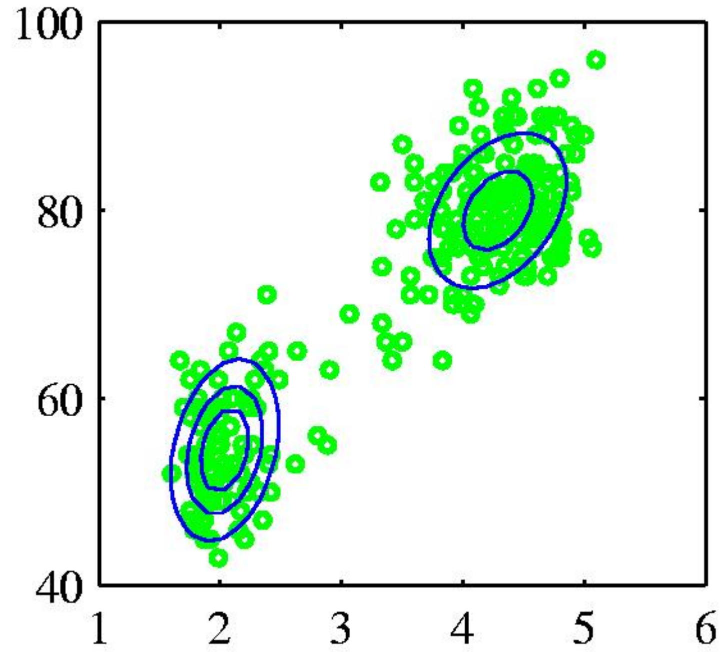


Mixtures of Gaussians (1)

Old Faithful Data Set



Single Gaussian



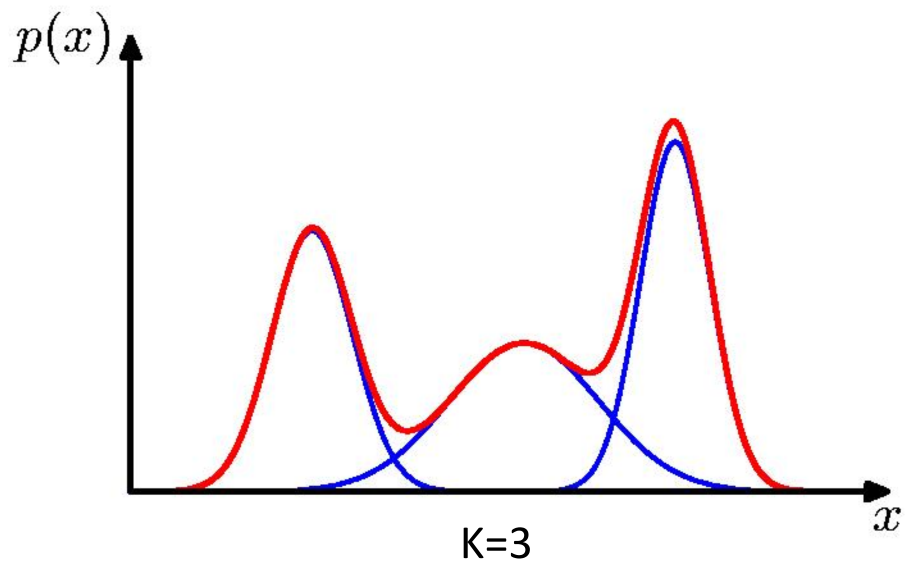
Mixture of two Gaussians

Mixtures of Gaussians (2)

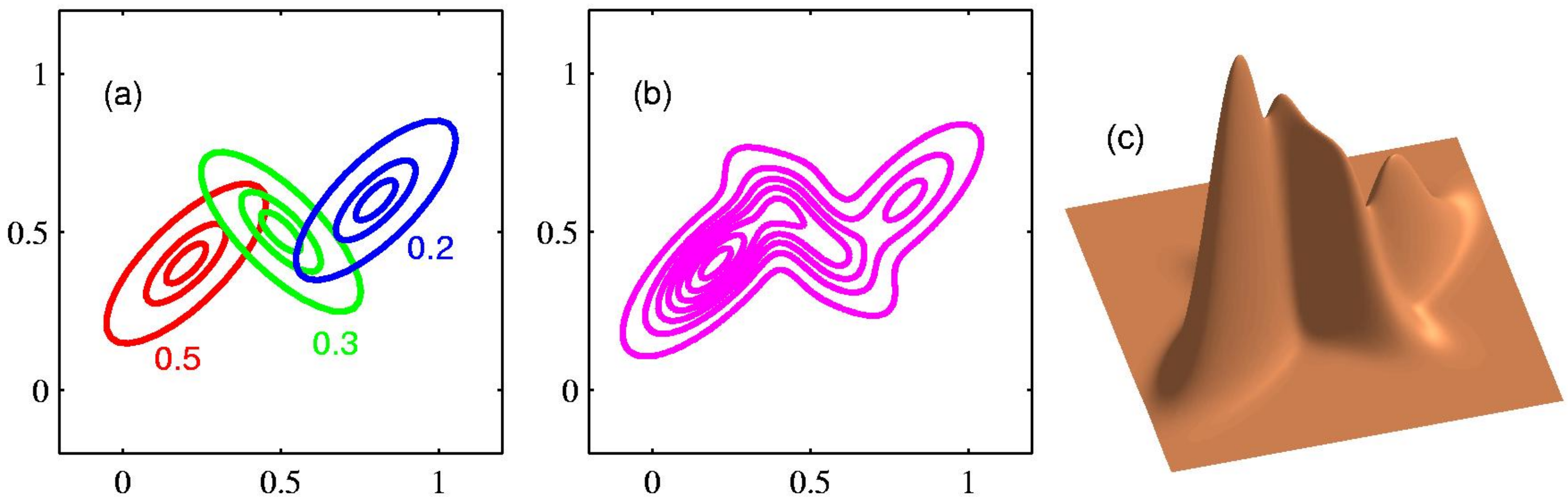
Combine simple models into a complex model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \underset{\substack{\uparrow \\ \text{Mixing coefficient}}}{\pi_k} \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Component}}$$

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

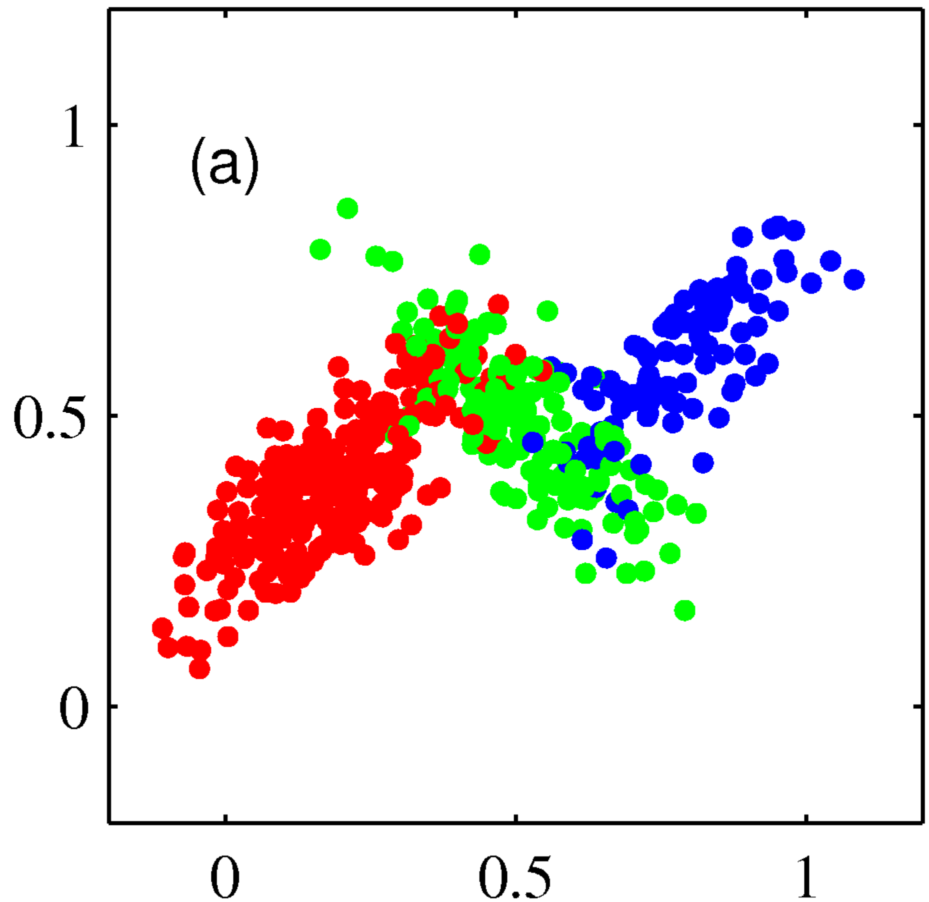


Mixtures of Gaussians (3)



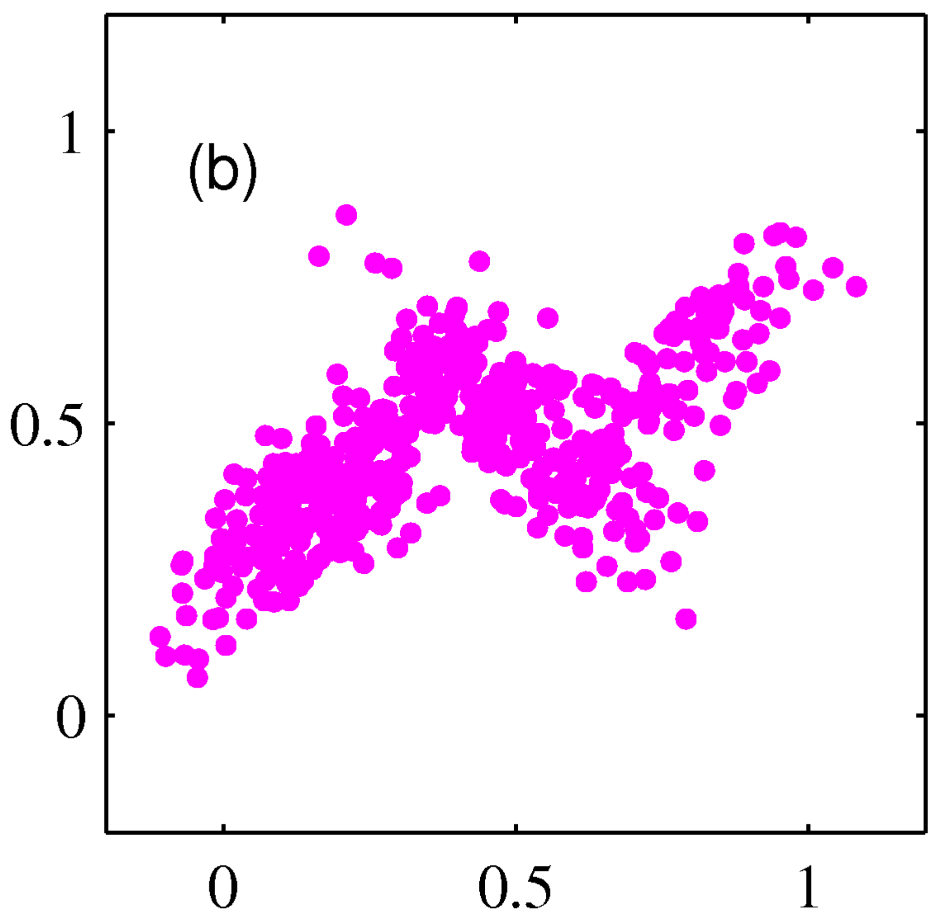
Unsupervised learning

Model data as mixture of multivariate Gaussians



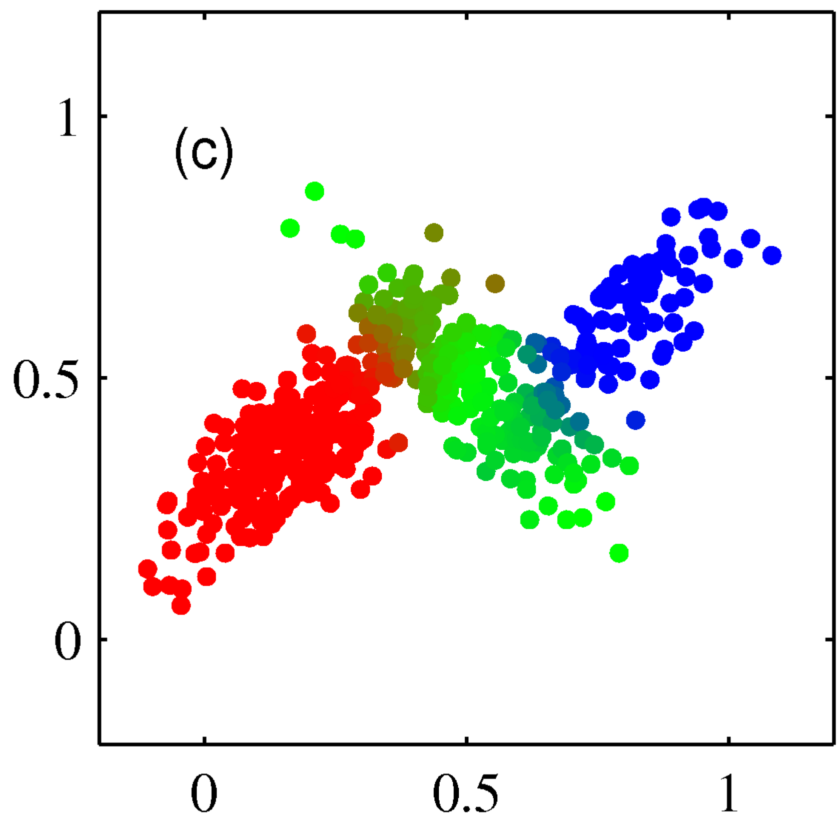
Unsupervised learning

Model data as mixture of multivariate Gaussians



Unsupervised learning

Model data as mixture of multivariate Gaussians



Shown is the *posterior probability* that a point was generated from i[th] Gaussian: $\Pr(Y = i \mid x)$

ML estimation in **supervised** setting

- Univariate Gaussian

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2$$

- ***Mixture*** of ***Multi***variate Gaussians

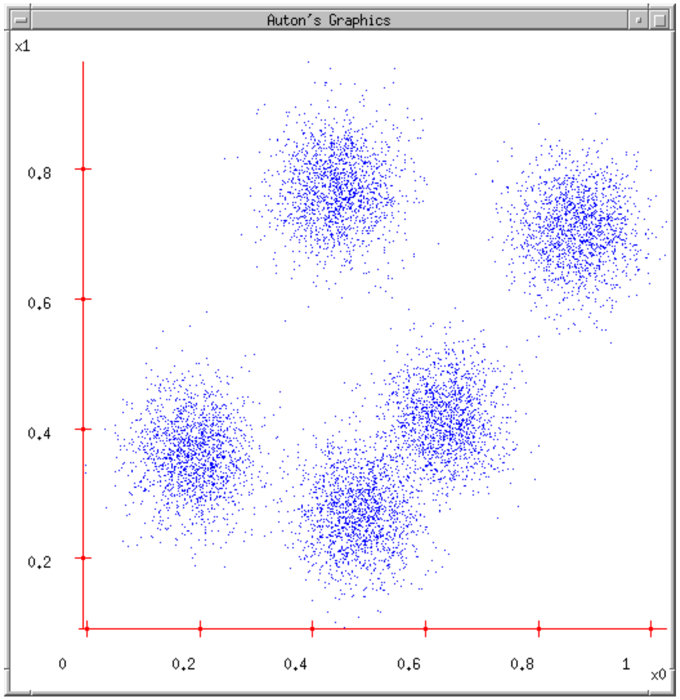
ML estimate for each of the Multivariate Gaussians is given by:

$$\mu^k_{ML} = \frac{1}{n}\sum_{j=1}^{n} x_n \qquad \Sigma^k_{ML} = \frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{x}_j - \mu^k_{ML}\right)\left(\mathbf{x}_j - \mu^k_{ML}\right)^T$$

Just sums over *x* generated from the *k*'th Gaussian

That was easy! But what if ***unobserved data***?

- MLE:
  - – $\arg\max_\theta \prod_j P(y_j, x_j)$
  - – θ: all model parameters
    - eg, class probs, means, and variances
- But we don't know y_j's!!!
- Maximize ***marginal likelihood***:
  - – $\arg\max_\theta \prod_j P(x_j) = \arg\max \prod_j \sum_{k=1}^{K} P(Y_j=k, x_j)$



EM: Two Easy Steps

Objective: $\text{argmax}_\theta \lg \prod_j \sum_{k=1}^{K} P(Y_j=k, x_j \mid \theta) = \sum_j \lg \sum_{k=1}^{K} P(Y_j=k, x_j \mid \theta)$

Data: $\{x_j \mid j=1 .. n\}$

- **E-step**: Compute expectations to "fill in" missing y values according to current parameters, θ
  - For all examples j and values k for Y_j, compute: $P(Y_j=k \mid x_j, \theta)$

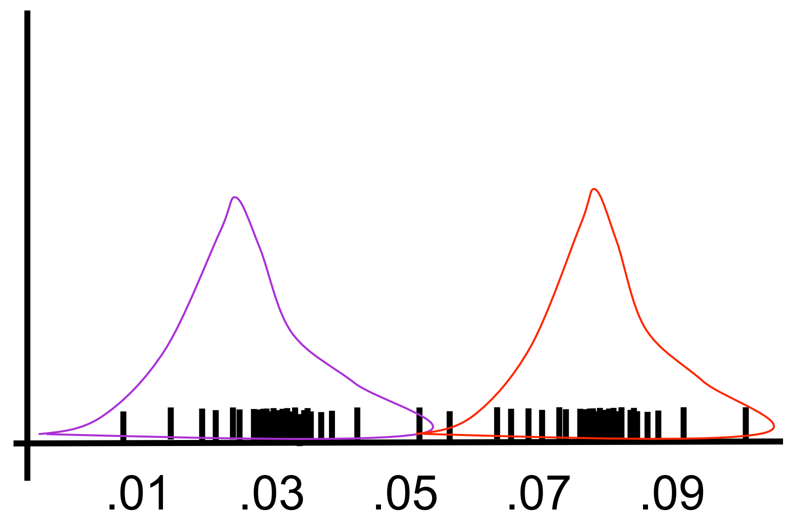
- **M-step**: Re-estimate the parameters with "weighted" MLE estimates
  - Set $\theta = \text{argmax}_\theta \sum_j \sum_k P(Y_j=k \mid x_j, \theta) \log P(Y_j=k, x_j \mid \theta)$

Especially useful when the E and M steps have closed form solutions!!!

Simple example: learn means only!

Consider:

- 1D data
- Mixture of k=2 Gaussians
- Variances fixed to σ=1
- Distribution over classes is uniform
- Just need to estimate μ_1 and μ_2



$$\prod_{j=1}^{m} \sum_{k=1}^{K} P(x, Y_j = k) \propto \prod_{j=1}^{m} \sum_{k=1}^{K=2} \exp\left[-\frac{1}{2\sigma^2}\|x - \mu_k\|^2\right] P(Y_j = k)$$

EM for GMMs: only learning means

Iterate: On the *t*'th iteration let our estimates be

$$\lambda_t = \{ \mu_1^{(t)}, \mu_2^{(t)} \ldots \mu_K^{(t)} \}$$

E-step

Compute "expected" classes of all datapoints

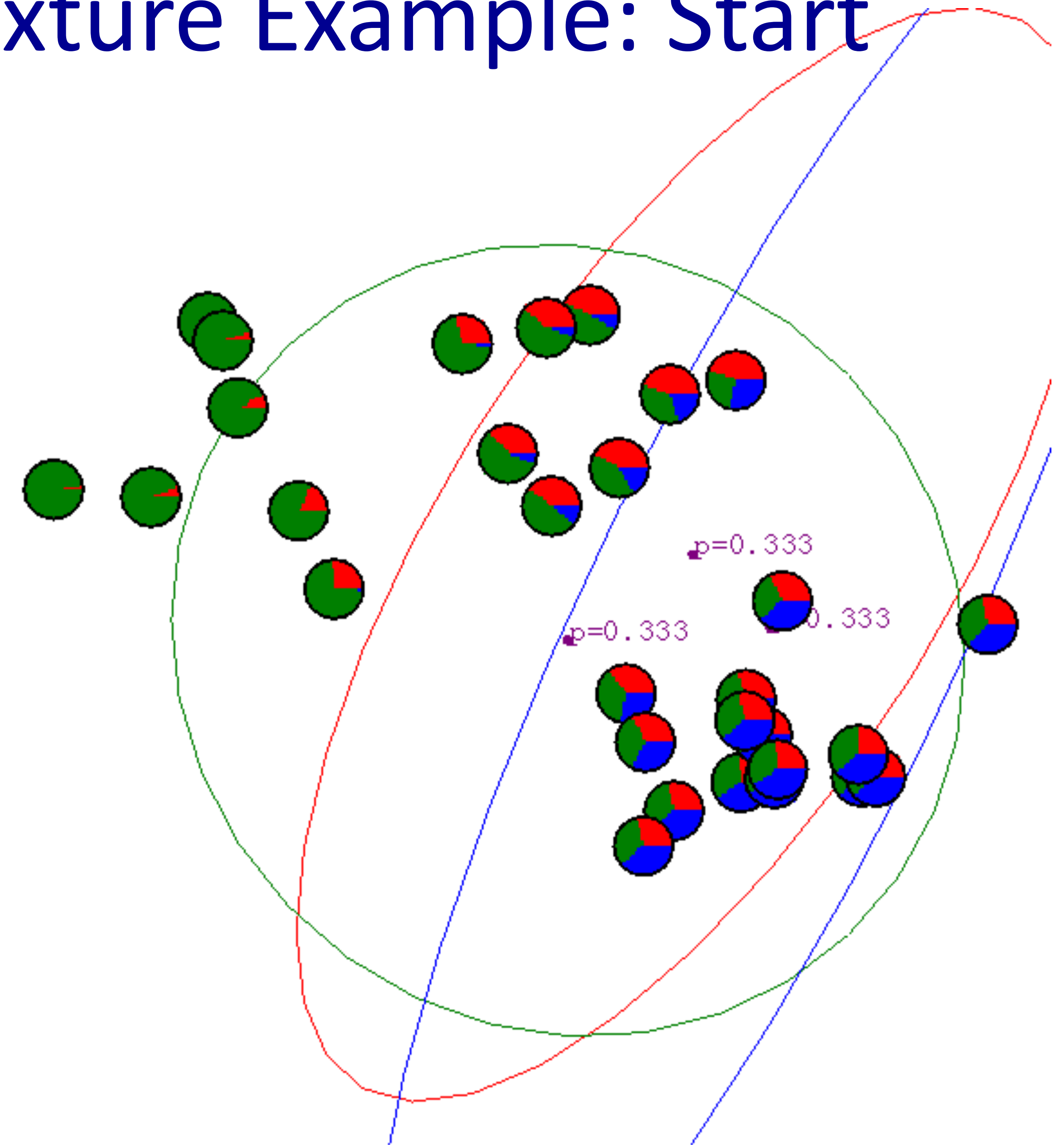
$$P\left(Y_j = k \middle| x_j, \mu_1 \ldots \mu_K\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left\|x_j - \mu_k\right\|^2\right) \mathrm{P}\left(Y_j = k\right)$$

M-step

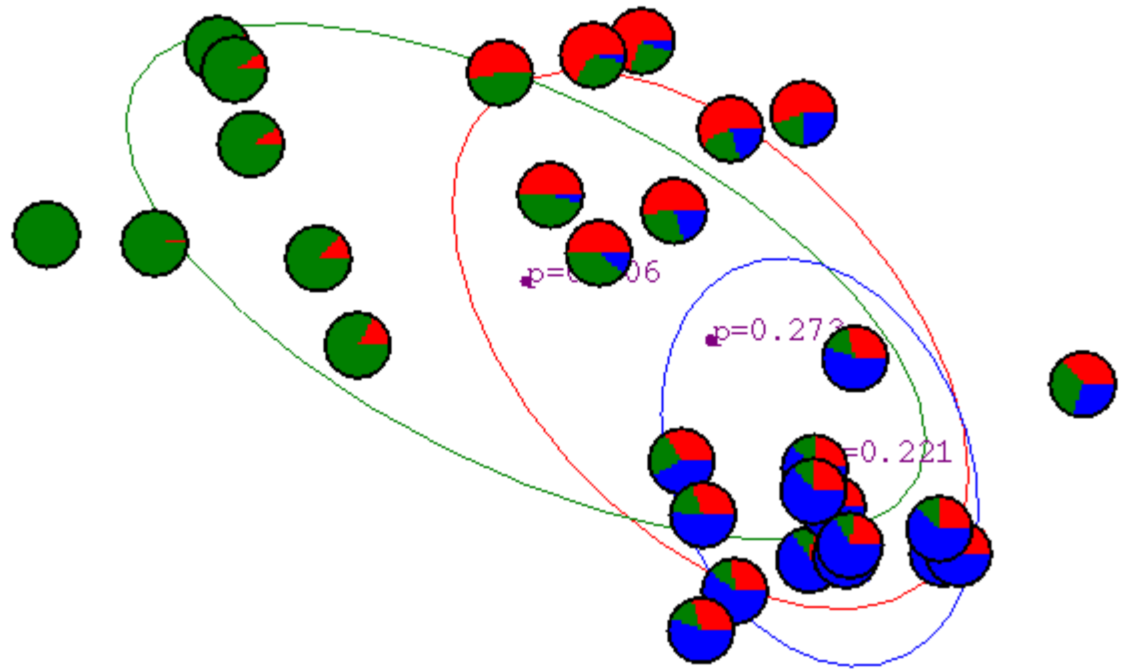
Compute most likely new **μ**s given class expectations

$$\mu_k = \frac{\sum_{j=1}^{m} P\left(Y_j = k \middle| x_j\right) x_j}{\sum_{j=1}^{m} P\left(Y_j = k \middle| x_j\right)}$$

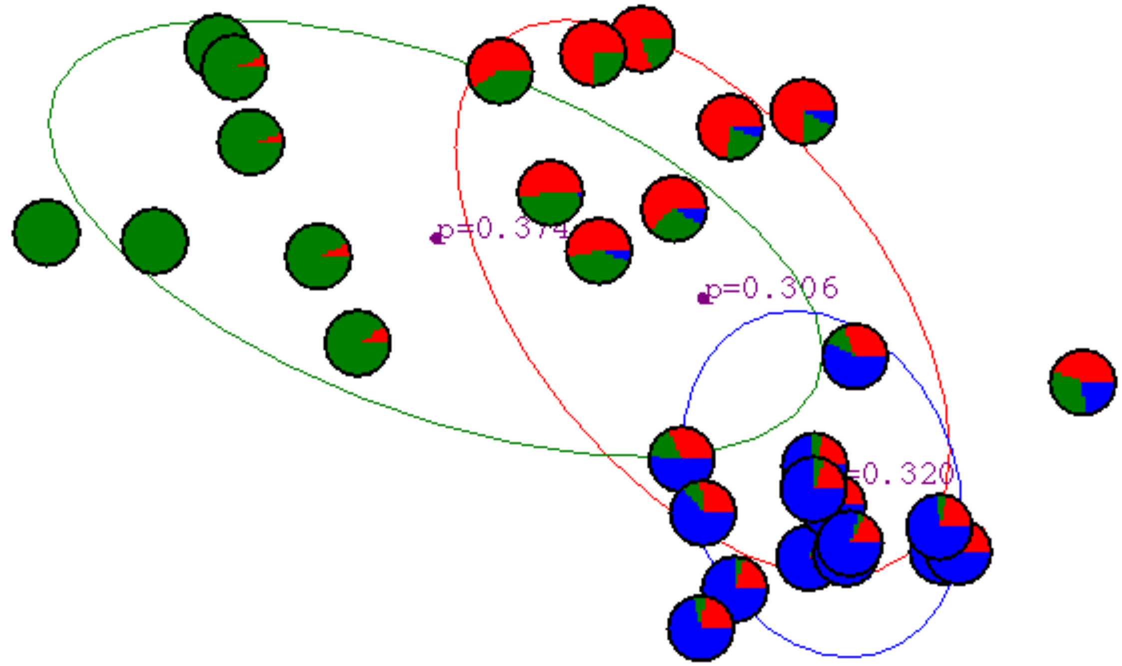
Gaussian Mixture Example: Start



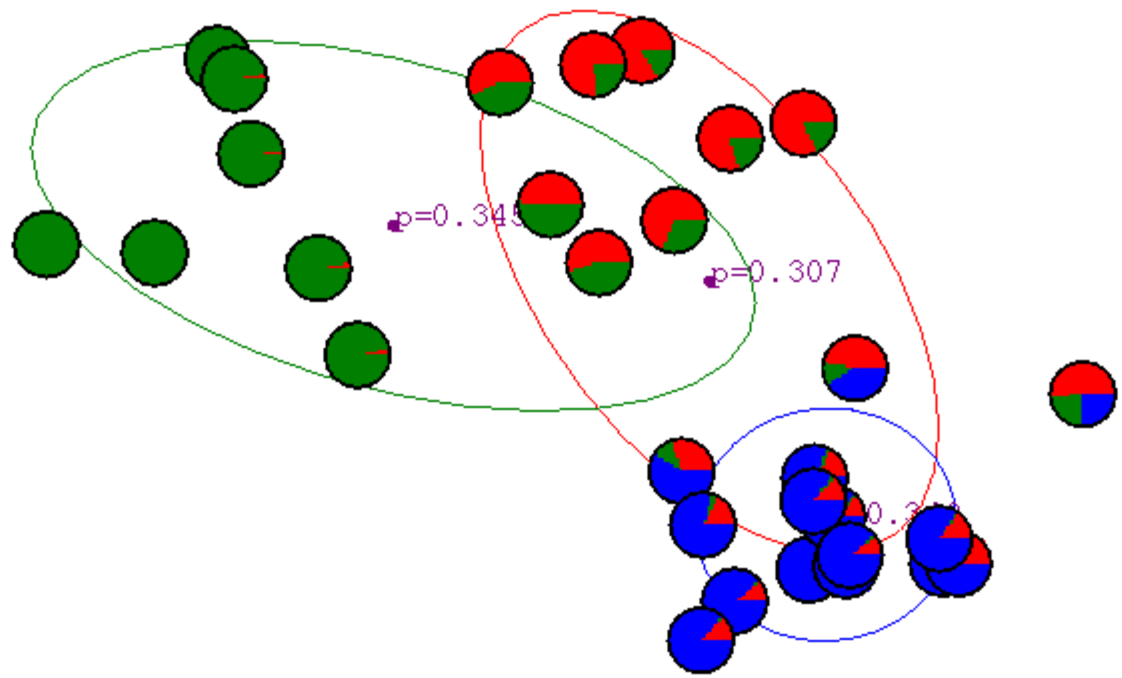
After first iteration



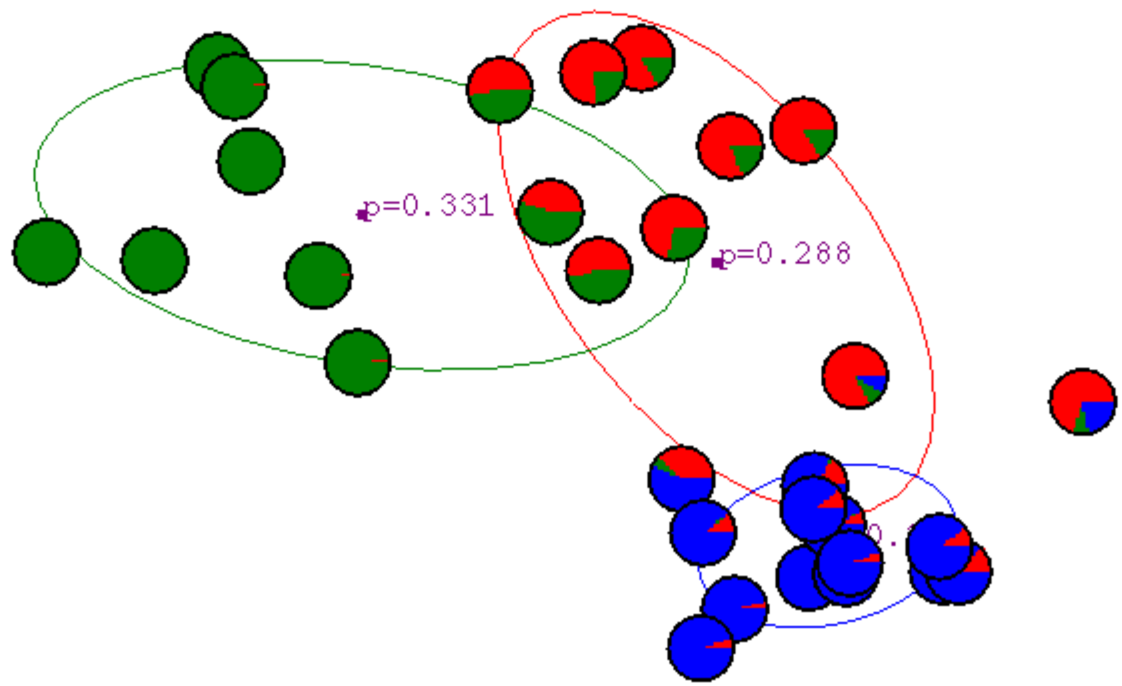
After 2nd iteration



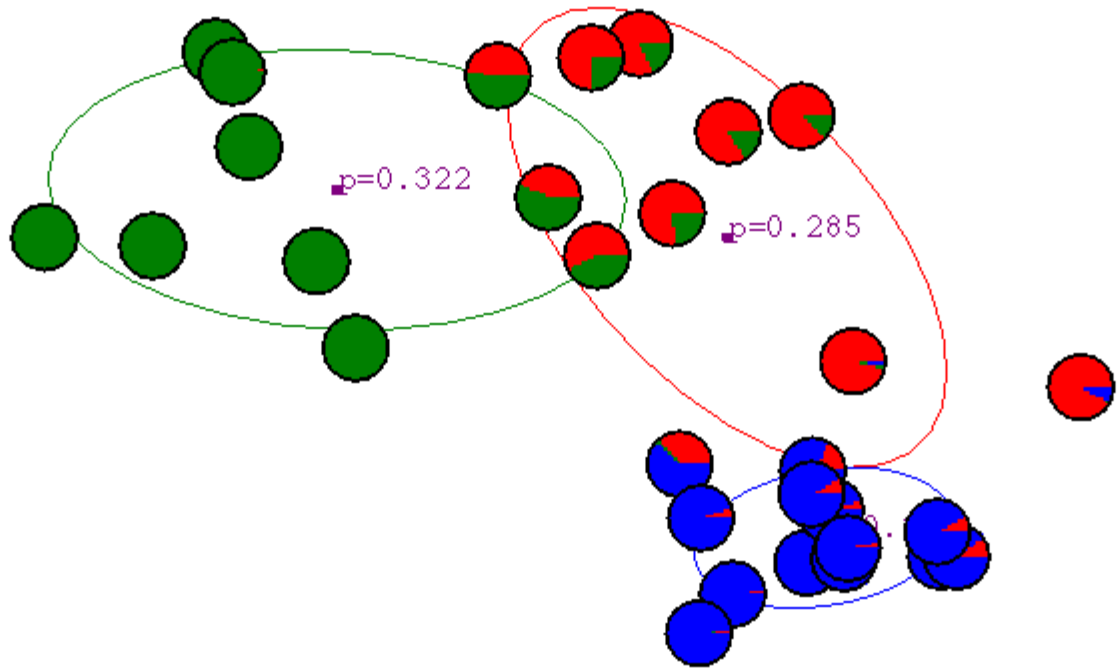
After 3rd iteration



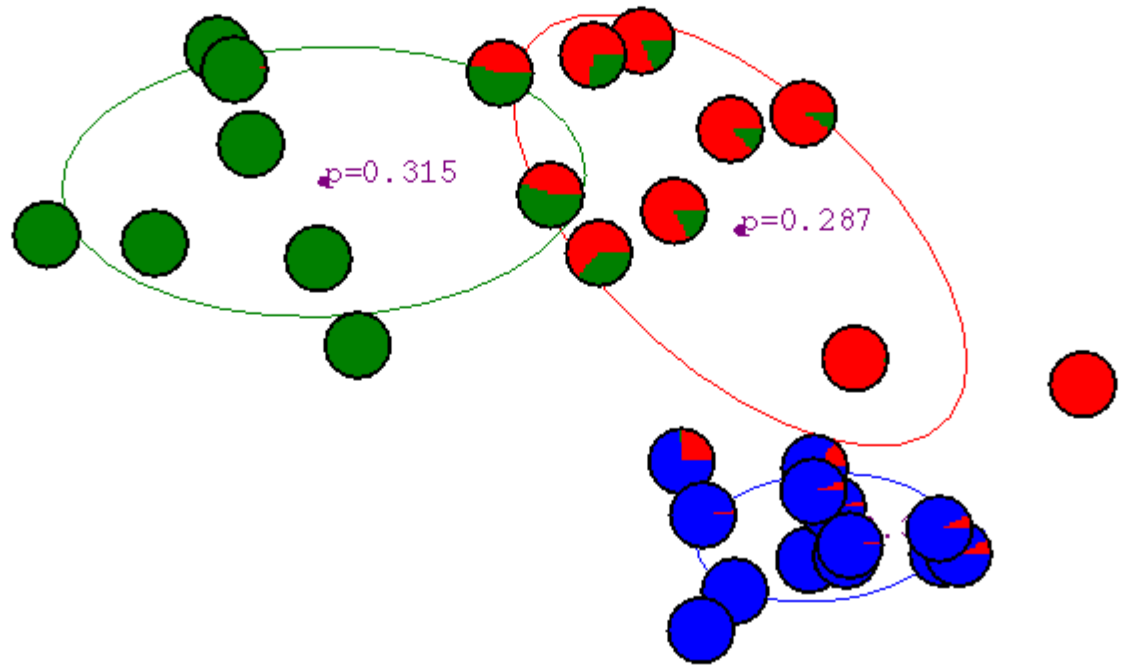
After 4th iteration



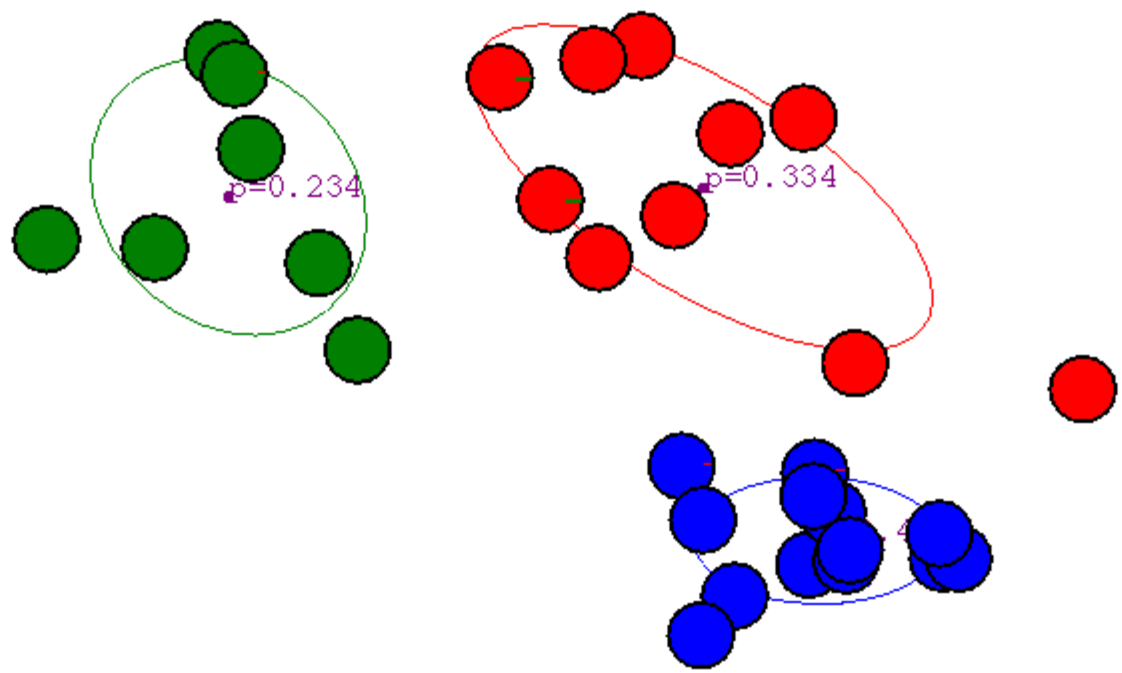
After 5th iteration



After 6th iteration



After 20th iteration



Jensen's inequality

- **Theorem**: log $\sum_{\mathbf{z}}$ P($\mathbf{z}$) f($\mathbf{z}$) ≥ $\sum_{\mathbf{z}}$ P($\mathbf{z}$) log f($\mathbf{z}$)
  - **e.g., Binary case for convex function f:**

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2).$$