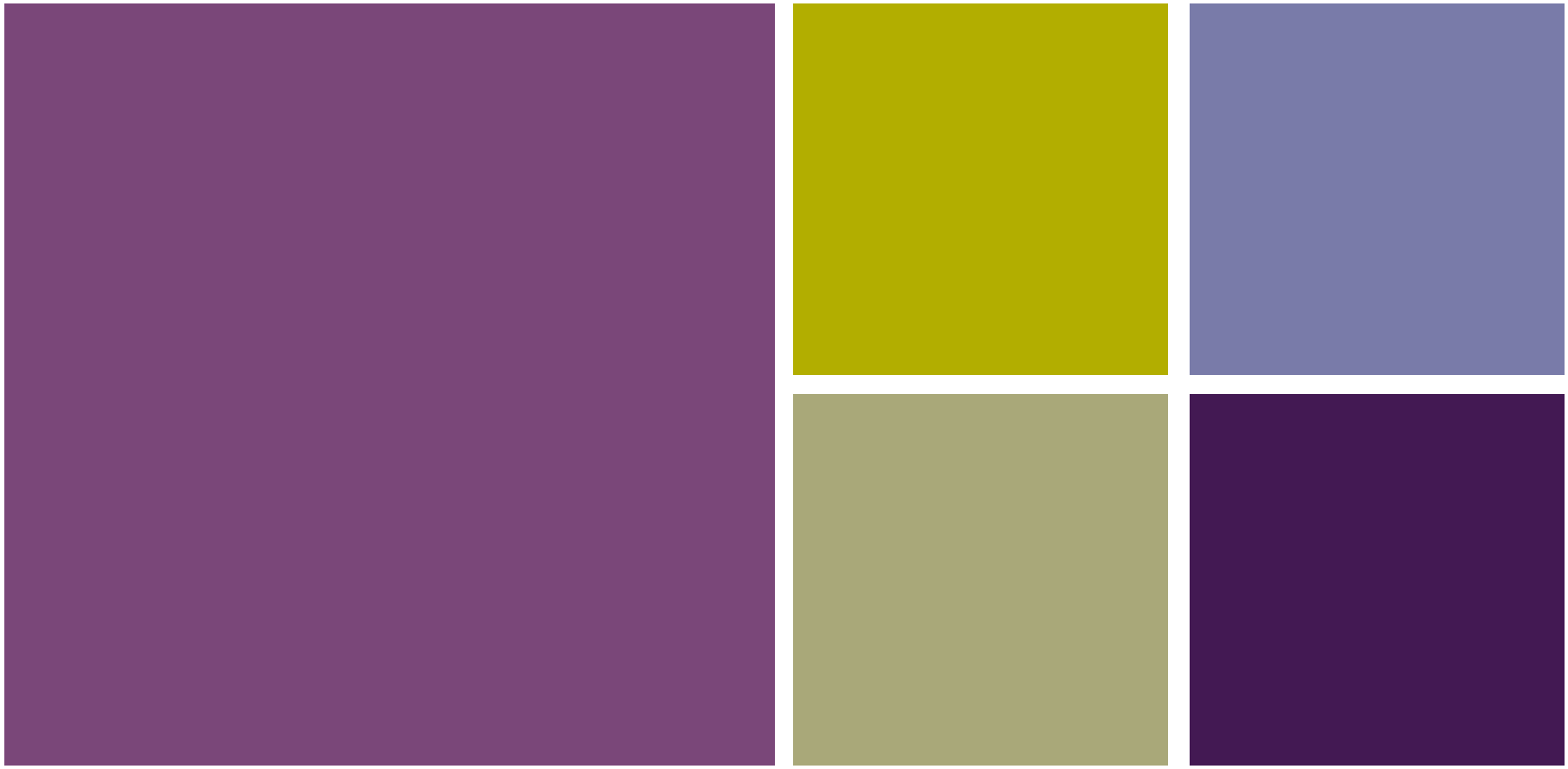


Gaussian Mixture Models &
Expectation Maximization Algorithm

# Pattern Recognition and Image Analysis

Dr. Manal Helal – Fall 2014

Lecture 6

# Objectives

- Preliminaries
- GMM
- EM

# Notation

- Random variables are represented with capital letters,
- Values they take are represented with lowercase letters
- P(x) : Probability of value x
- $p_X$ : probability distribution for random variable X
- $p_X(x)$: represents the probability of value x (according to $p_X$).
- $p_{X|Y}(x|y)$: represents the probability of value x given value y (according to $p_X$ given $p_Y$).
- $X_1^n$ : represent the sequence $X_1, X_2, \ldots, X_n$
- $x_1^n$ : $x_1, x_2, \ldots, x_n$

# Gaussian Mixture Models

- A Gaussian mixture model (GMM) is useful for modeling data that comes from one of several groups: the groups might be different from each other, but data points within the same group can be well-modeled by a Gaussian distribution.

- The Gaussian PDF is:

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

- For short:

$$p_X(x) = \mathcal{N}(x; \mu, \sigma^2)$$

# Estimating the mean

- Given a Gaussian $X_1^n$ i.i.d observations with unknown mean and variance, we can use MLE to estimate the variance. First find the Log-likelihood, then differentiation, then set it to 0.

$$p_{X_1^n}(x_1^n) = \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_1-\mu)^2/2\sigma^2}$$

$$\ln p_{X_1^n}(x_1^n) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2}(x_i - \mu)^2$$

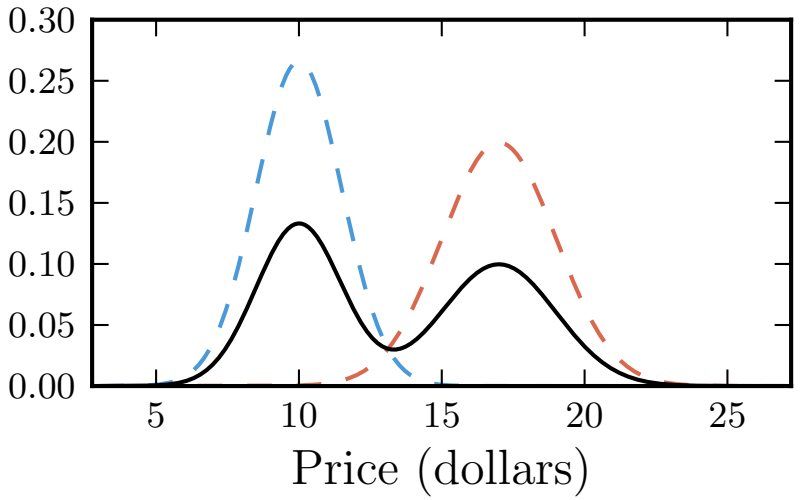
$$\frac{d}{d\mu} \ln p_{X_1^n}(x_1^n) = \sum_{i=1}^{n} \frac{1}{\sigma^2}(x_i - \mu)$$

- The mean is estimated to be: $\widehat{\mu} = \frac{1}{N}\sum_i x_i$

which is the average of the observed sample.

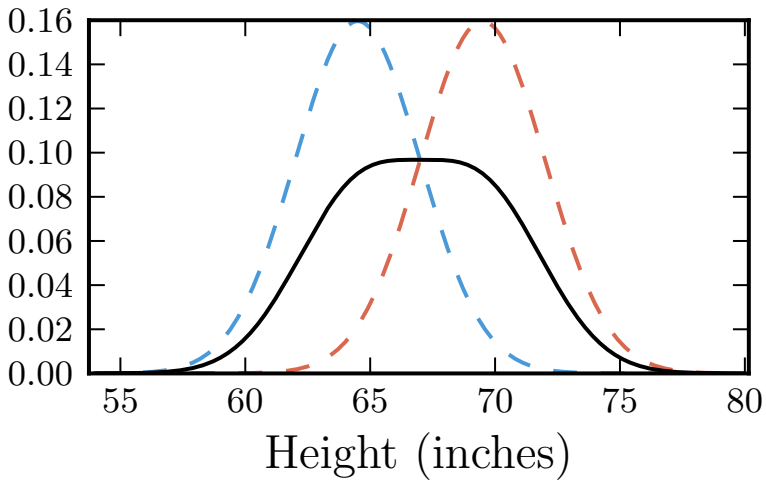
# Example 1

- The price of a randomly chosen paperback book is normally distributed with mean $10.00 and standard deviation $1.00
- The price of a randomly chosen hardback is normally distributed with mean $17 and variance $1.50
- Is the price of any book selected at random from both groups, will be normally distributed?



# Example 2

- The height of a randomly chosen man is normally distributed with a mean around 5'9.5" and standard deviation around 2.5"
- The height of a randomly chosen woman is normally distributed with a mean around 5'4.5" and standard deviation around 2.5"
- Is the height of any person selected at random from both groups, will be normally distributed?



# The Model

- Given people numbered i = 1, . . . , n, their heights as $Y_i \in R$ and an unobserved label $C_i \in \{M,F\}$ for each person representing that person's gender.

- Assuming that the two groups have the same known variance $\sigma^2$, but different unknown means $\mu_M$ and $\mu_F$. The distribution for the class labels is Bernoulli:

$$p_{C_i}(c_i) = q^{\mathbb{1}(c_i=M)}(1-q)^{\mathbb{1}(c_i=F)}$$

- Assuming we know q, and replacing $\pi_M$ = q, and $\pi_F$ = (1-q) and for any number of classes, we will have:

$$p_{C_i}(c_i) = \prod_{c\in\{M,F\}} \pi_c^{\mathbb{1}(c_i=c)}$$

- The conditional distributions within each class are Gaussian:

$$p_{Y_i|C_i}(y_i|c_i) = \prod_c \mathcal{N}(y_i;\mu_c,\sigma^2)^{\mathbb{1}(c_i=c)}$$

# Parameter Estimation: $\mu_M$, $\mu_F$

- Given the model setup in previous slide, compute the joint density of all the data points $pY_1,\ldots,Y_N(y_1,\ldots,y_n)$ in terms of $\mu_M$, $\mu_F$, $\sigma$, and q. Take the log to find the log- likelihood, and then differentiate with respect to $\mu_M$.
- density for a single data point Yi = yi:

$$p_{Y_i}(y_i) = \sum_{c_i} p_{C_i}(c_i) p_{Y_i|C_i}(y_i|c_i)$$

$$= \sum_{c_i} \left(\pi_c \mathcal{N}(y_i; \mu_C, \sigma^2)\right)^{\mathbb{1}(c_i=c)}$$

$$= q\mathcal{N}(y_i; \mu_M, \sigma^2) + (1-q)\mathcal{N}(y_i; \mu_F, \sigma^2)$$

- The joint density of all the observations is:

$$p_{Y_1^n}(y_1^n) = \prod_{i=1}^{n} \left(q\mathcal{N}(y_i; \mu_M, \sigma^2) + (1-q)\mathcal{N}(y_i; \mu_F, \sigma^2)\right)$$

- The log-likelihood of the parameters is then :

$$\ln p_{Y_1^n}(y_1^n) = \sum_{i=1}^{n} \ln \left(\pi_M \mathcal{N}(y_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(y_i; \mu_F, \sigma^2)\right)$$

# Using Hidden Variables



$$\ln p_{Y_1^n}(y_1^n) = \sum_{i=1}^{n} \ln \left( \pi_M \mathcal{N}(y_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(y_i; \mu_F, \sigma^2) \right)$$

- The sum prevents the log-likelihood : $\log_b(a+c) = \log_b a + \log_b \left(1 + \frac{c}{a}\right)$
- This is a mixture of exponential and linear term and difficult to get a closed form maximum likelihood.
- If we knew the hidden labels $C_i$ exactly, then it would be easy to do ML estimates for the parameters:
  - take all the points for which $C_i = M$ and use those to estimate $\mu_M$.
  - repeat for the points where $C_i = F$ to estimate $\mu_F$.
  - start with Bayes' rule:

$$p_{C_i|Y_i}(c_i|y_i) = \frac{p_{Y_i|C_i}(y_i|c_i)p_{C_i}(c_i)}{p_{Y_i}(y_i)}$$

$$= \frac{\prod_{c\in\{M,F\}} \left(\pi_c \mathcal{N}(y_i; \mu_c, \sigma^2)\right)^{\mathbb{1}(c=c_i)}}{\pi_M \mathcal{N}(y_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(y_i; \mu_F, \sigma^2)} = q_{C_i}(c_i)$$

# Solving for $C_i = M$

$$p_{C_i|Y_i}(M|y_i) = \frac{\pi_M \mathcal{N}(y_i; \mu_M, \sigma^2)}{\pi_M \mathcal{N}(y_i; \mu_M, \sigma^2) + \pi_F \mathcal{N}(y_i; \mu_F, \sigma^2)} = q_{C_i}(M)$$

- Rewrite in terms of q$C_i$ and set to zero and solve for $\mu_M$:

$$\sum_{i=1}^{n} q_{C_i}(M) \frac{y_i - \mu_M}{\sigma^2} = 0$$

$$\mu_M = \frac{\sum_{i=1}^{n} q_{C_i}(M) y_i}{\sum_{i=1}^{n} q_{C_i}(M)}$$

- $\mu_M$ is a weighted average of the heights, where each height is weighted by how likely that person is to be male.

- By symmetry, solve for $\mu_F$, creating a circular setup: fix one and solve for the other: EM algorithm.

# Expectation Maximisation Algorithm – Informal Steps

- Initialize the parameters somehow.
- First, fix the parameters (in this case, the means $\mu_M$ and $\mu_F$ of the Gaussians) and solve for the posterior distribution for the hidden variables (in this case, q$C_i$ , the class labels).
- Second, fix the posterior distribution for the hidden variables (again, that's q$C_i$, the class labels), and optimize the parameters (the means $\mu_M$ and $\mu_F$ ) using the expected values of the hidden variables (in this case, the probabilities from q$C_i$ ).
- Repeat the two steps above until the values aren't changing much (i.e., until convergence).

# Preliminaries: Entropy

"I thought of calling it "information", but the word was overly used, so I decided to call it "uncertainty". [...] Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."

- Conversation between Claude Shannon and John von Neumann regarding what name to give to the attenuation in phone-line signals.

# Shannon Entropy

- **Entropy** is the average amount of information contained in each message received. Here, *message* stands for an event, sample or character drawn from a distribution or data stream.
- Entropy thus characterizes our uncertainty about our source of information. The idea here is that the less likely an event is, the more information it provides when it occurs.
- Shannon Entropy H(X) for random variable X with Probability distribution P(X) is defined as:

$$H(X) = \sum_i P(x_i)\, I(x_i) = -\sum_i P(x_i) \log_b P(x_i)$$

I is the information content of X. I(X) is itself a random variable, defined as the negative of the logarithm of the probability distribution.

# Preliminaries: KL Divergence

- The Kullback–Leibler divergence (a.k.a information divergence, information gain, relative entropy, or KLIC) is a non-symmetric measure of the difference between two probability distributions P and Q.
- The KL divergence of $Q$ from $P$, denoted $D_{\mathrm{KL}}(P||Q)$, is a measure of the information lost when $Q$ is used to approximate $P$.
- $D_{\mathrm{KL}}(P||Q)$ is not symmetric to $D_{\mathrm{KL}}(Q||P)$
- For Discrete Probability: $$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\ln\frac{P(i)}{Q(i)}.$$

i.e. it is the expectation of the logarithmic difference between the probabilities $P$ and $Q$, where the expectation is taken using the probabilities $P$. The KL divergence is only defined if $P$ and $Q$ both sum to 1 and if  implies  for all i

- For Continuous Probability: $$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x)\,\ln\frac{p(x)}{q(x)}\,\mathrm{d}x,$$

where $p$ and $q$ denote the densities of $P$ and $Q$.

# Preliminaries: Jensen's Inequality

- Generally it relates the value of a convex function of an integral to the integral of the convex function.
- In our case, we need this form:

$$\log(\mathbb{E}[X]) \geq \mathbb{E}[\log(X)]$$

- For a geometric intuition and a proof and more detail, see Wikipedia

# Preliminaries: Marginal Distribution

- The marginal distribution of a subset of a collection of random variables is the probability distribution of the variables contained in the subset. It gives the probabilities of various values of the variables in the subset without reference to the values of the other variables. This contrasts with a conditional distribution, which gives the probabilities contingent upon the values of the other variables.

Joint and marginal distributions of a pair of discrete, random variables X,Y having nonzero mutual information I(X; Y). The values of the joint distribution are in the 4×4 square, and the values of the marginal distributions are along the right and bottom margins.

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $p_y(Y)\downarrow$ |
|---|---|---|---|---|---|
| $y_1$ | $4/32$ | $2/32$ | $1/32$ | $1/32$ | $8/32$ |
| $y_2$ | $2/32$ | $4/32$ | $1/32$ | $1/32$ | $8/32$ |
| $y_3$ | $2/32$ | $2/32$ | $2/32$ | $2/32$ | $8/32$ |
| $y_4$ | $8/32$ | 0 | 0 | 0 | $8/32$ |
| $p_x(X) \rightarrow$ | $16/32$ | $8/32$ | $4/32$ | $4/32$ | $32/32$ |

# Preliminaries: Marginal Distribution – Cont'd



- Given two random variables X and Y whose joint distribution is known, the marginal distribution of X is simply the probability distribution of X averaging over information about Y. It is the probability distribution of X when the value of Y is not known. This is typically calculated by summing or integrating the joint probability distribution over Y.

- For Discrete Random Variables:

$$\Pr(X = x) = \sum_y \Pr(X = x, Y = y) = \sum_y \Pr(X = x | Y = y) \Pr(Y = y),$$

where Pr(X = x,Y = y) is the joint distribution of X and Y, while Pr(X = x | Y = y) is the conditional distribution of X given Y

- For Continuous Random Variables:

$$p_X(x) = \int_y p_{X,Y}(x, y)\, \mathrm{d}y = \int_y p_{X|Y}(x|y)\, p_Y(y)\, \mathrm{d}y,$$

where $p_{X,Y}(x,y)$ gives the joint distribution of $X$ and $Y$, while $p_{X|Y}(x|y)$ gives the conditional distribution for $X$ given $Y$.

# EM Formal Algorithm

- The EM algorithm is actually maximizing a lower bound on the log likelihood (in other words, each step is guaranteed to improve our answer until convergence).

# EM Steps



1. Maximize the log-likelihood $\log p_Y(y;\theta)$
   a. Marginalizing over C and introducing $q_C$ (c)/$q_C$ (c)
   $$= \log\left(\sum_c q_C(c)\frac{p_{Y,C}(y,c;\theta)}{q_C(c)}\right)$$
   b. Rewriting as an expectation
   $$= \log\left(\mathbb{E}_{q_C}\left[\frac{p_{Y,C}(y,C;\theta)}{q_C(C)}\right]\right)$$
   c. Using Jensen's inequality
   $$\geq \mathbb{E}_{q_C}\left[\log\frac{p_{Y,C}(y,C;\theta)}{q_C(C)}\right]$$
2. M-Step:
   $$\mathbb{E}_{q_C}\left[\log\frac{p_{Y,C}(y,C;\theta)}{q_C(C)}\right] = \mathbb{E}_{q_C}\left[\log p_{Y,C}(y,C;\theta)\right] - \mathbb{E}_{q_C}\left[\log q_C(C)\right]$$
   a. Rearrange
   b. Maximizing with respect to $\theta$ :
   $$\widehat{\theta} \leftarrow \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{q_C}\left[\log p_{Y,C}(y,C;\theta)\right]$$
3. E-Step:
   $$\mathbb{E}_{q_C}\left[\log\frac{p_{Y,C}(y,C;\theta)}{q_C(C)}\right] = \mathbb{E}_{q_C}\left[\log\frac{p_Y(y;\theta)p_{C|Y}(C|y;\theta)}{q_C(C)}\right]$$
   a. Rearrange
   $$= \log p_Y(y;\theta) - \mathbb{E}_{q_C}\left[\log\frac{q_C(C)}{p_{C|Y}(C|y;\theta)}\right]$$
   $$= \log p_Y(y;\theta) - D(q_C(\cdot)||p_{C|Y}(\cdot|y;\theta))$$
   b. Maximizing with respect to qC :
   $$\widehat{q}_C(\cdot) \leftarrow p_{C|Y}(\cdot|y;\theta)$$

# EM Step 1



1. Maximize the log-likelihood:

$$\log p_Y(y;\theta) = \log\left(\sum_c p_{Y,C}(y,c)\right)$$

   a. Log of sum problem!
   b. Solution: if we have an expectation for one variable (C here), we can swap the order using Jensen's inequality.
   c. Introduce a new distribution qC:

$$= \log\left(\sum_c q_C(c)\frac{p_{Y,C}(y,c;\theta)}{q_C(c)}\right)$$

$$= \log\left(\mathbb{E}_{q_C}\left[\frac{p_{Y,C}(y,C;\theta)}{q_C(C)}\right]\right)$$

$$\geq \mathbb{E}_{q_C}\left[\log\frac{p_{Y,C}(y,C;\theta)}{q_C(C)}\right]$$

   d. Using Jensen's inequality :
   e. Using definition of conditional probability

$$= \mathbb{E}_{q_C}\left[\log\frac{p_Y(y;\theta)p_{C|Y}(C|y;\theta)}{q_C(C)}\right]$$

- Now we have a lower bound on log $p_Y(y;\theta)$ that we can optimize pretty easily. Since we've introduced qC, we now want to maximize this quantity with respect to both $\theta$ and qC.

# EM Step 2 – The M Step.



2. The M stands for maximization, since we're maximizing with respect to the parameters.
   a. Find Best Parameters $\theta$ by rearranging using Jensen's inequality :

$$\mathbb{E}_{q_C}\left[\log \frac{p_{Y,C}(y, C; \theta)}{q_C(C)}\right] = \mathbb{E}_{q_C}\left[\log p_{Y,C}(y, C; \theta)\right] - \mathbb{E}_{q_C}\left[\log q_C(C)\right]$$

   b. In general, qC doesn't depend on $\theta$ , so we'll only care about the first term:

$$\widehat{\theta} \leftarrow \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{q_C}\left[\log p_{Y,C}(y, C; \theta)\right]$$

- Now we have a lower bound on log $p_Y$ (y; $\theta$ ) that we can optimize pretty easily. Since we've introduced qC, we now want to maximize this quantity with respect to both $\theta$ and qC.

# EM Step 3 – The E Step.



3. the E stands for expectation, since we're computing qC so that we can use it for expectations.
   a. Find Best qC by rearranging using definition of conditional probability:

$$\mathbb{E}_{q_C}\left[\log \frac{p_Y(y;\theta)p_{C|Y}(C|y;\theta)}{q_C(C)}\right] = \mathbb{E}_{q_C}[\log p_Y(y;\theta)] + \mathbb{E}_{q_C}\left[\log \frac{p_{C|Y}(C|y;\theta)}{q_C(C)}\right]$$

   b. The first term doesn't depend on c, and the second term almost looks like a KL divergence:

$$= \log p_Y(y;\theta) - \mathbb{E}_{q_C}\left[\log \frac{q_C(C)}{p_{C|Y}(C|y;\theta)}\right]$$

$$= \log p_Y(y;\theta) - D(q_C(\cdot)||p_{C|Y}(\cdot|y;\theta))$$

   c. When maximizing this quantity, we want to make the KL divergence as small as possible. KL divergences are always greater than or equal to 0, and they're exactly 0 when the two distributions are equal. So, the optimal qC is $p_{C|Y}(c|y;\theta)$

$$\widehat{q}_C(\cdot) \leftarrow p_{C|Y}(\cdot|y;\theta)$$

# Repeat Until Convergence

By alternating between

$$\widehat{\theta} \leftarrow \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{q_C}\left[\log p_{Y,C}(y, C; \theta)\right]$$

and

$$\widehat{q}_C(\cdot) \leftarrow p_{C|Y}(\cdot|y; \theta)$$

we can maximize a lower bound on the log-likelihood. We've also seen from E-Step that the lower bound is tight (that is, it's equal to the log-likelihood) when we are computing qC .

# The EM Algorithm

**Inputs:** Observation $y$, joint distribution $p_{Y,C}(y, c; \theta)$, conditional distribution $p_{C|Y}(c|y; \theta)$, initial values $\theta^{(0)}$

**function** EM($p_{Y,C}(y, c; \theta), p_{C|Y}(c|y; \theta), \theta^{(0)}$)
    **for** iteration $t \in 1, 2, \ldots$ **do**
        $q_C^{(t)} \leftarrow p_{C|Y}(c|y; \theta^{(t-1)})$ (**E-step**)
        $\theta^{(t)} \leftarrow \text{argmax}_\theta \, \mathbb{E}_{q_C^{(t)}} [p_{Y,C}(y, C; \theta)]$ (**M-step**)
        **if** $\theta^{(t)} \approx \theta^{(t-1)}$ **then**
            return $\theta^{(t)}$

# Applying the algorithm for GMM (again)



- Given an observed random variable Y (heights), some hidden variable C (gender) that Y depends on. The distributions of C and Y have some parameters $\theta$ (the means $\mu_M$ and $\mu_F$) that we don't know.

- The objective is to estimate the parameters $\theta$, given some initial value: suppose we set $\mu M = 3'$ and $\mu F = 5'$. Then the computed posteriors qCi would all favor F over M (since most people are closer to $5'$ than $3'$), and we would end up computing $\mu F$ as roughly the average of all our heights, and $\mu M$ as the average of a few short people.

- For the E-step, we have to compute the posterior distribution $p_{C|Y}(c|y)$:

$$p_{C_i|Y_i}(c_i|y_i) = \frac{p_{Y_i|C_i}(y_i|c_i)p_{C_i}(c_i)}{p_{Y_i}(y_i)}$$

$$= \frac{\prod_{c\in\{M,F\}} \left(\pi_c \mathcal{N}(y_i;\mu_c,\sigma^2)\right)^{\mathbb{1}(c=c_i)}}{\pi_M \mathcal{N}(y_i;\mu_M,\sigma^2) + \pi_F \mathcal{N}(y_i;\mu_F,\sigma^2)} = q_{C_i}(c_i)$$

- For $C_i = M$:

$$p_{C_i|Y_i}(M|y_i) = \frac{\pi_M \mathcal{N}(y_i;\mu_M,\sigma^2)}{\pi_M \mathcal{N}(y_i;\mu_M,\sigma^2) + \pi_F \mathcal{N}(y_i;\mu_F,\sigma^2)} = q_{C_i}(M)$$

# Applying the algorithm for GMM (again) – Cont'd



- To do the M-Step:

$$\mathbb{E}_{q_C}[\ln p_{Y,C}(y,C)] = \mathbb{E}_{q_C}[\ln p_{Y|C}(y|C)p_C(C)]$$

$$= \mathbb{E}_{q_C}\left[\ln \prod_{i=1}^{n} \prod_{c\in\{M,F\}} \left(\pi_c \mathcal{N}(y_i;\mu_c,\sigma^2)\right)^{\mathbb{1}(C_i=c)}\right]$$

$$= \mathbb{E}_{q_C}\left[\sum_{i=1}^{n} \sum_{c\in\{M,F\}} \mathbb{1}(C_i=c)\left(\ln \pi_c + \ln \mathcal{N}(y_i;\mu_c,\sigma^2)\right)\right]$$

$$= \sum_{i=1}^{n} \sum_{c\in\{M,F\}} \mathbb{E}_{q_C}\left[\mathbb{1}(C_i=c)\right]\left(\ln \pi_c + \ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{(y_i-\mu_c)^2}{2\sigma^2}\right)$$

- $\mathbb{E}_{q_C}\left[\mathbb{1}(C_i=c)\right]$ is the probability that Ci is c, according to q.
  Now, we can differentiate with respect to $\mu$ M :

$$\frac{d}{d\mu_M}\mathbb{E}_{q_C}[\ln p_{Y|C}(y|C)p_C(C)] = \sum_{i=1}^{n} q_{C_i}(M)\left(\frac{y_i-\mu_M}{\sigma^2}\right) = 0$$

# Applying the algorithm for GMM (again) – Cont'd



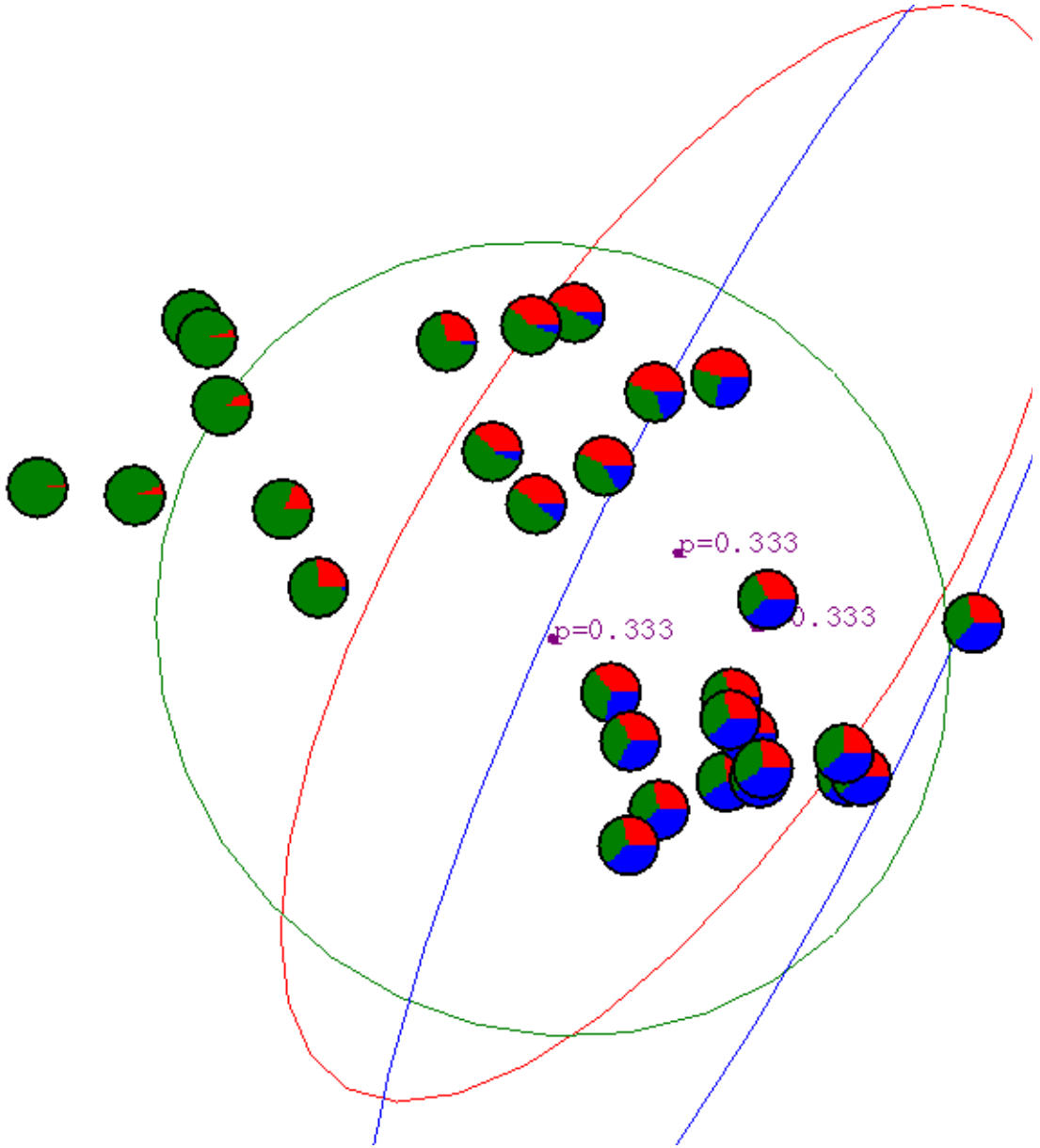
- Using the last form:

$$\sum_{i=1}^{n} q_{C_i}(M)\frac{y_i - \mu_M}{\sigma^2} = 0$$

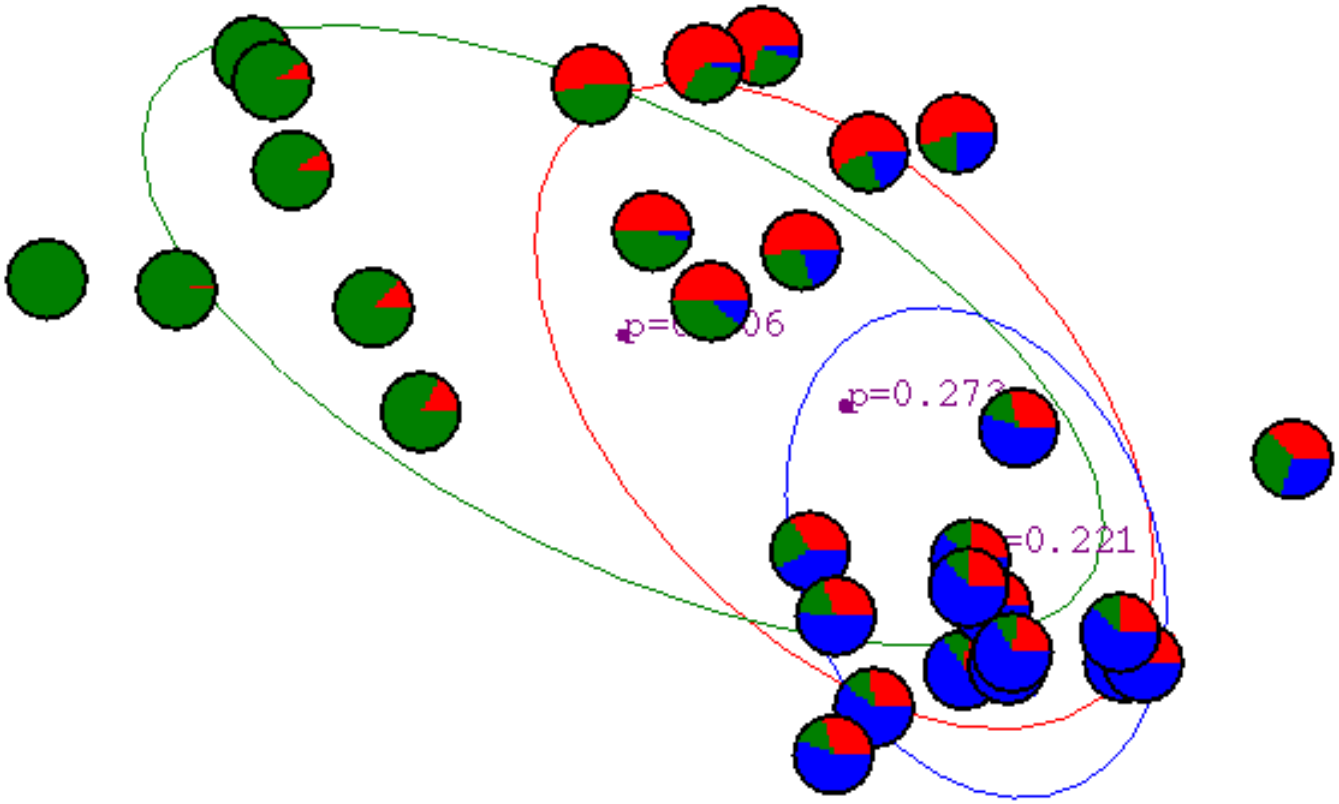
- The solution will be the weighted average as follows:

$$\mu_M = \frac{\sum_{i=1}^{n} q_{C_i}(M) y_i}{\sum_{i=1}^{n} q_{C_i}(M)} \qquad \mu_F = \frac{\sum_{i=1}^{n} q_{C_i}(F) y_i}{\sum_{i=1}^{n} q_{C_i}(F)}$$

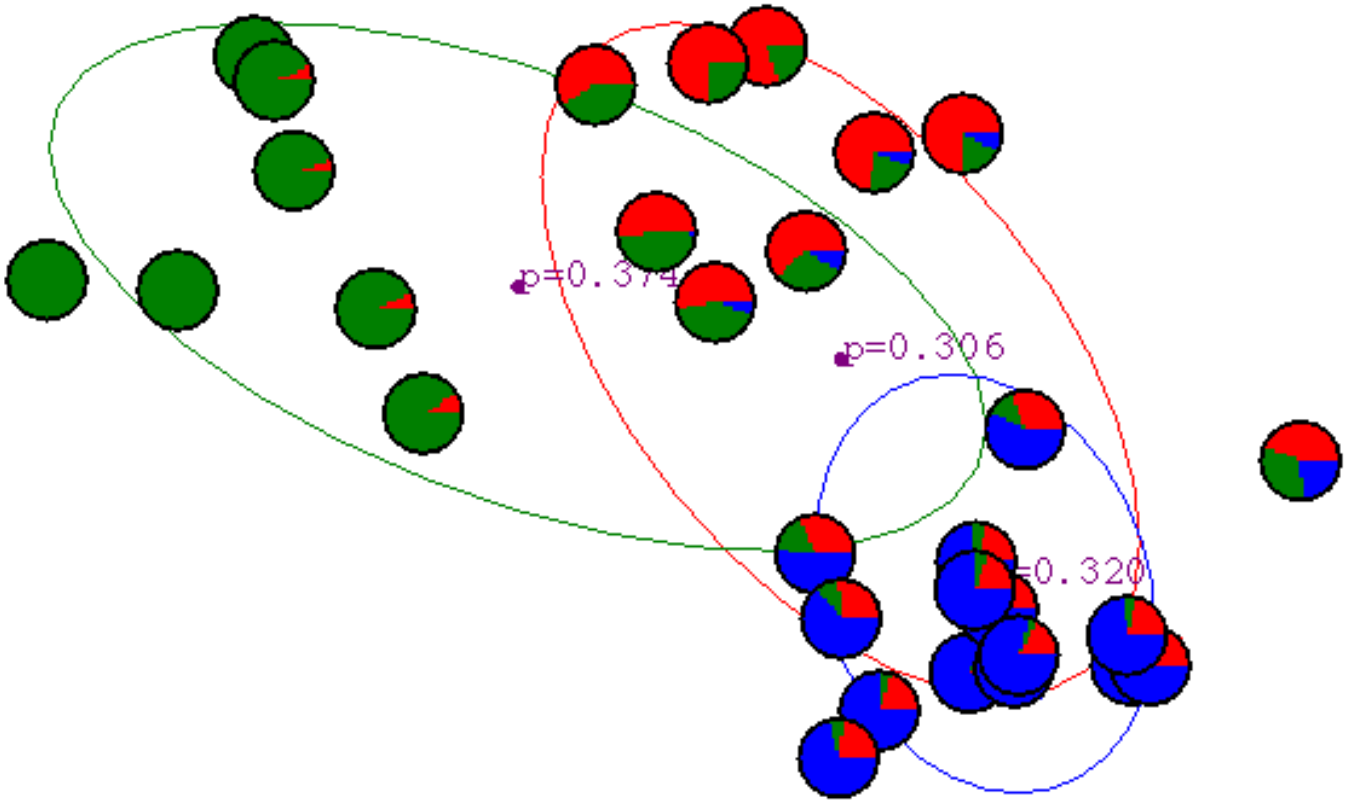
# Example: Gaussian Mixture Example: Start



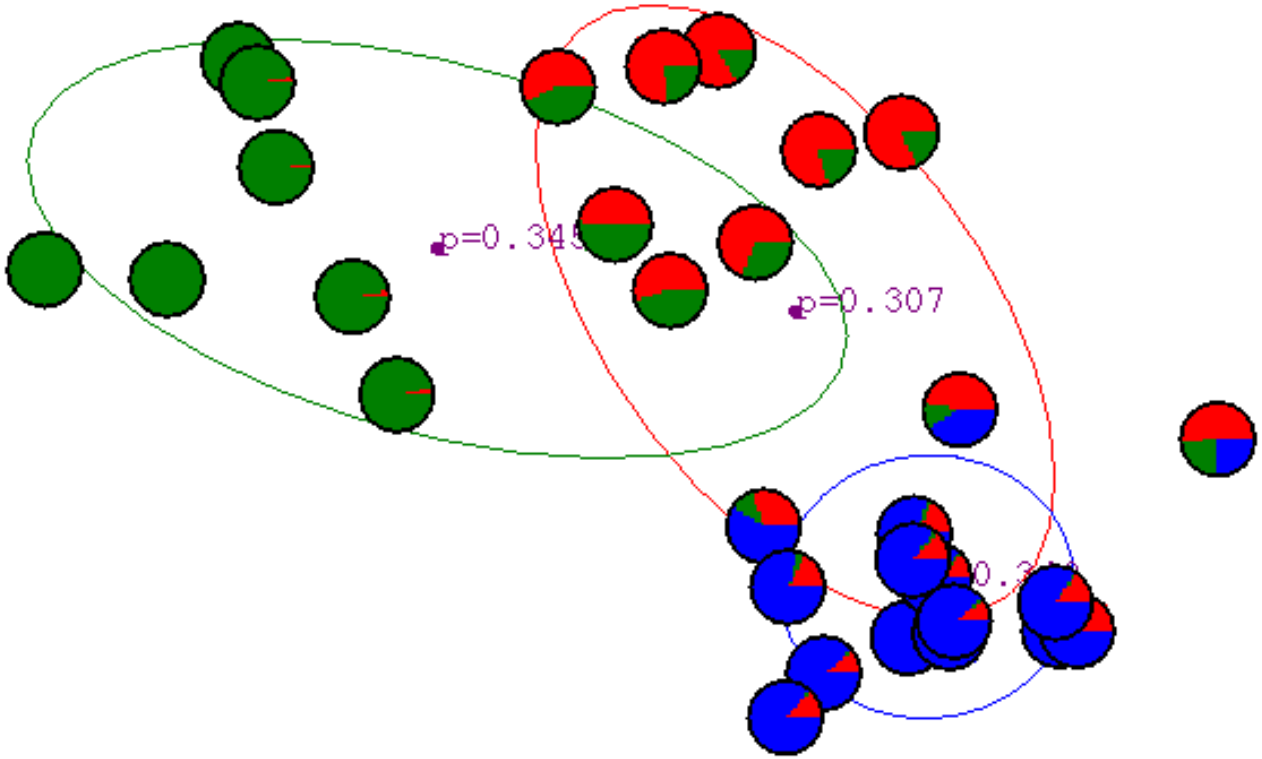
# After first iteration



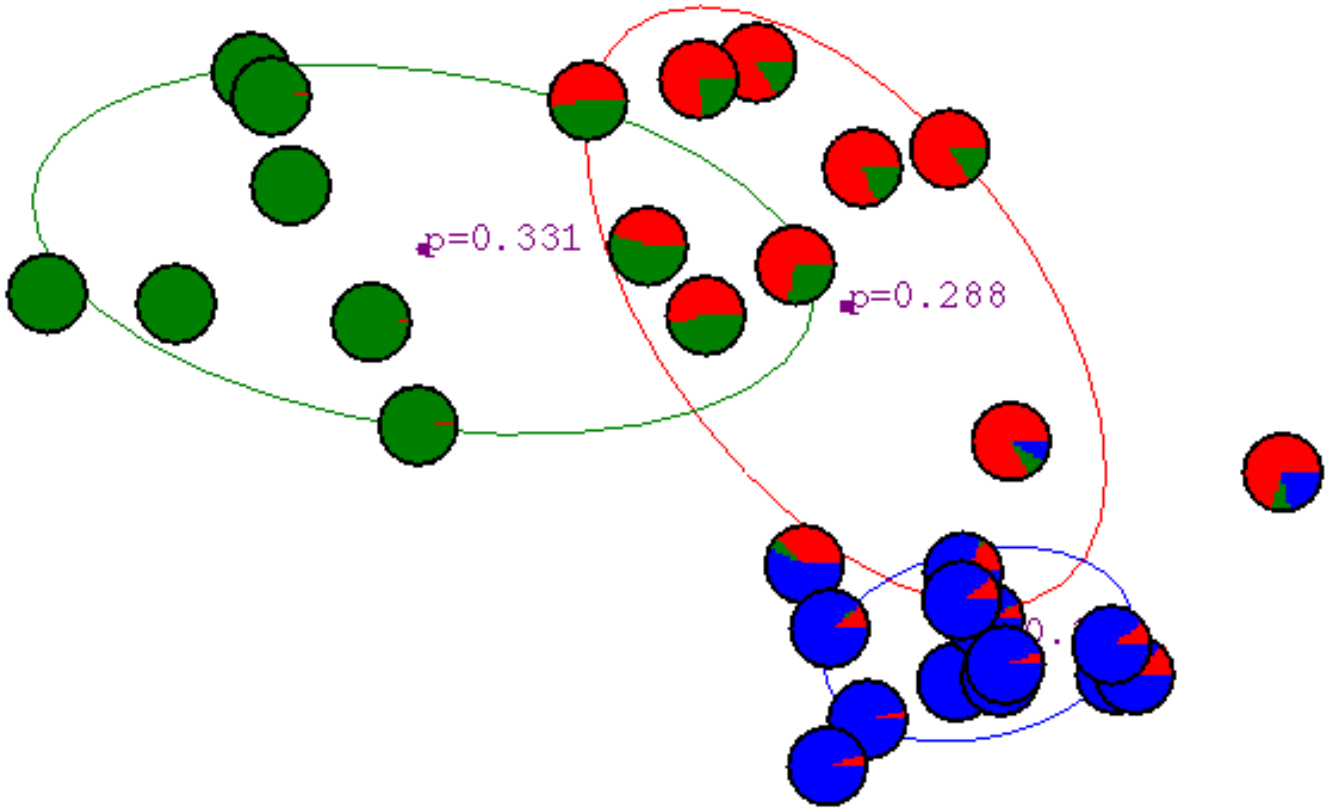
# After 2nd iteration



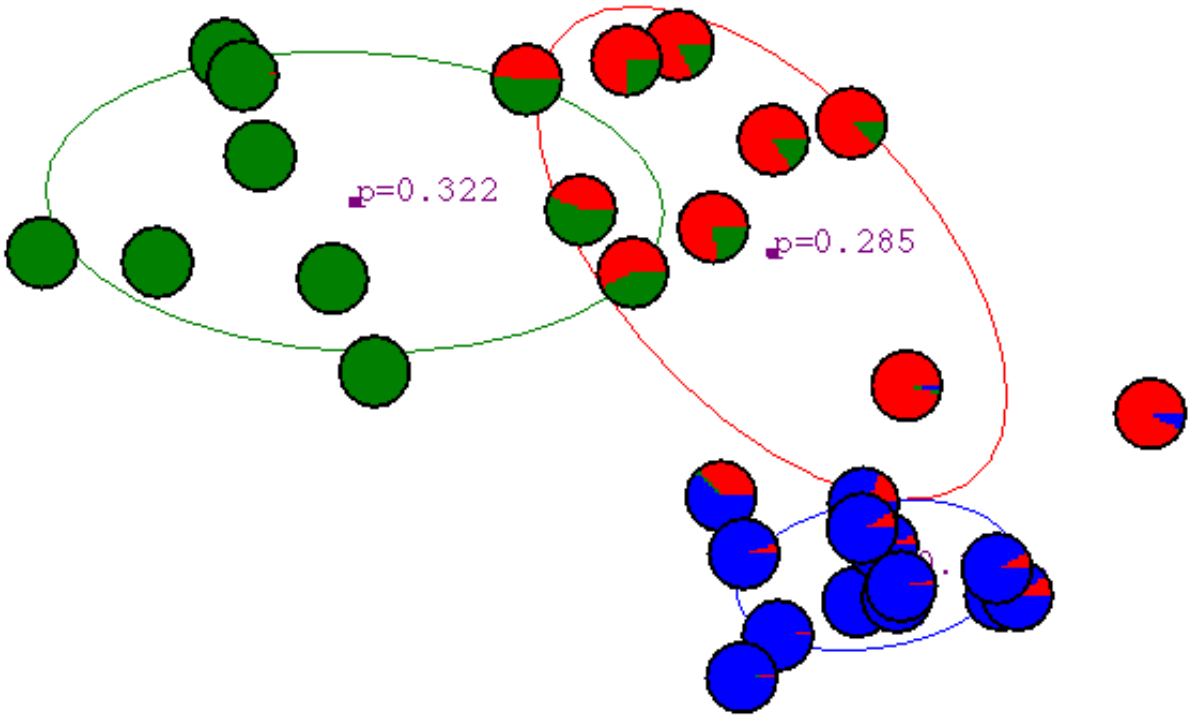
# After 3rd iteration



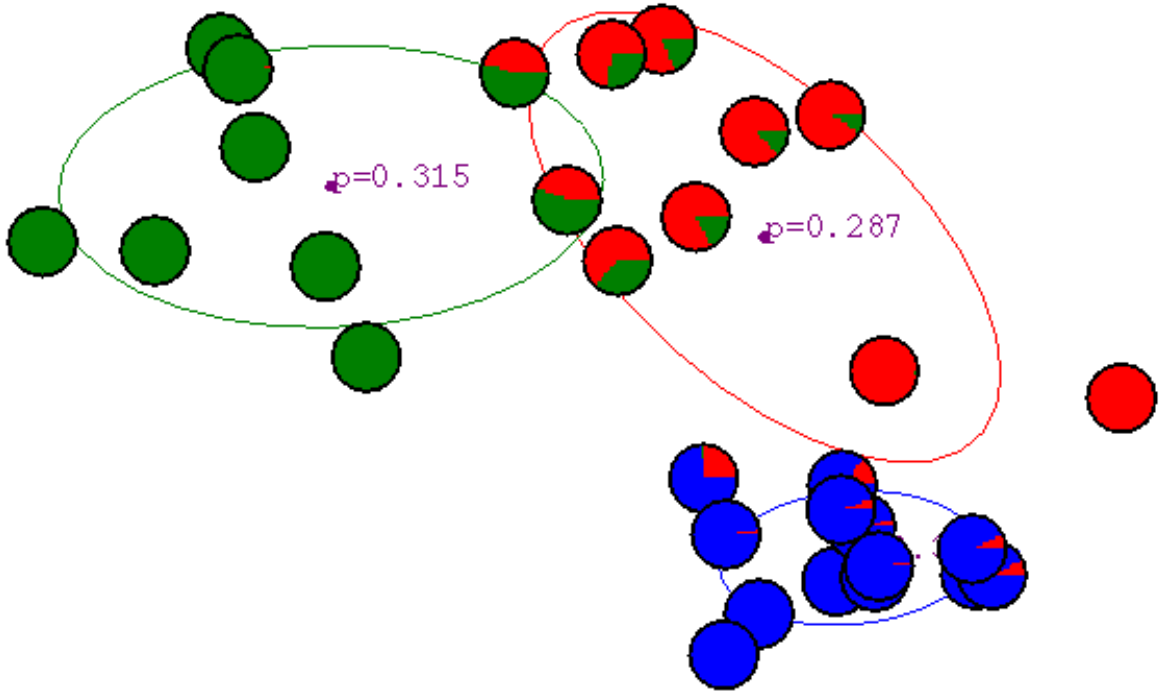
# After 4th iteration



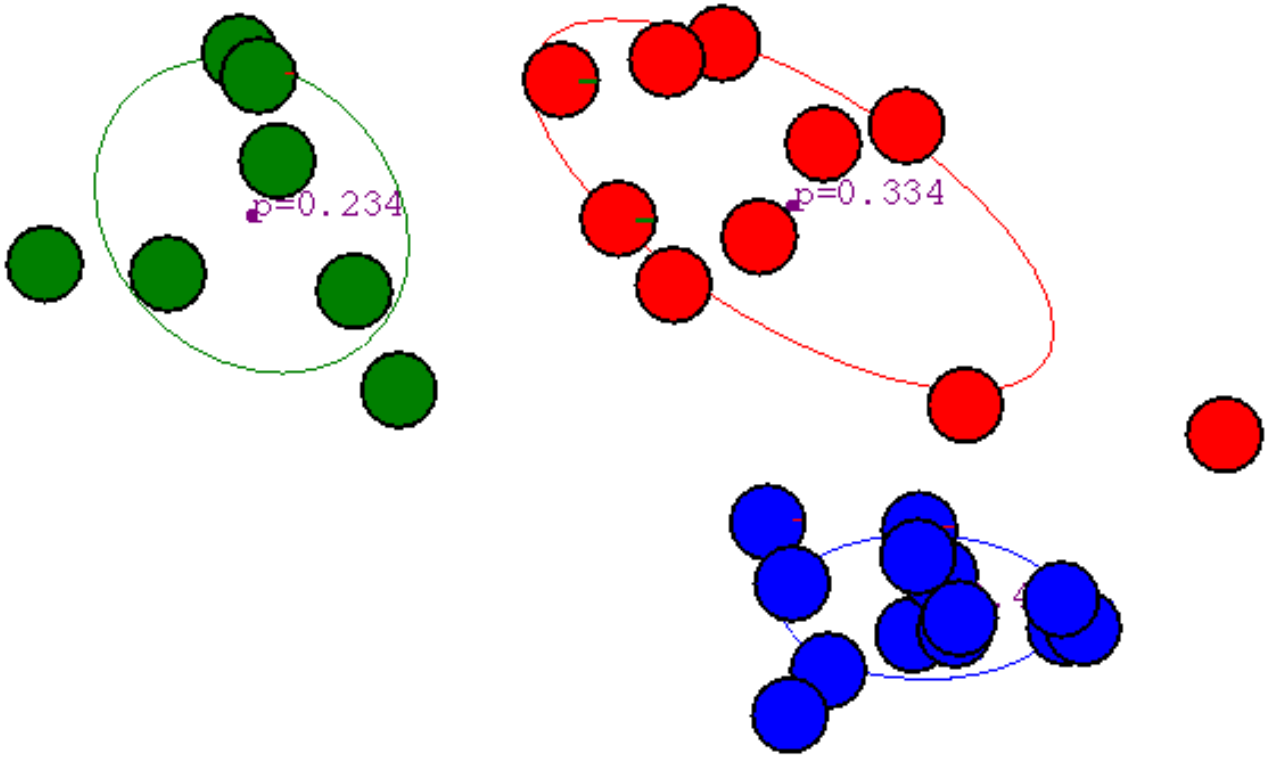
# After 5th iteration

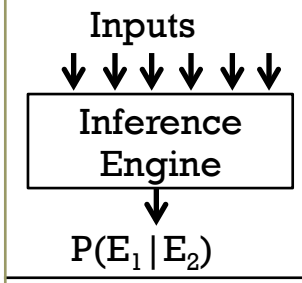
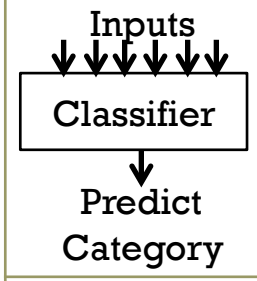
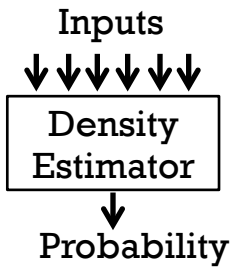
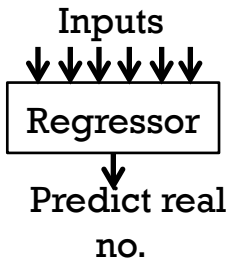


# After 6th iteration



# After 20th iteration



| | Categorical inputs only | Real-valued inputs only | Mixed Real / Cat okay | Methods |
|---|---|---|---|---|
|  | | | | Joint DE, Bayes Net Structure Learning |
|  | Joint BC<br>Naïve BC | Gauss BC | Dec Tree | Dec Tree, Sigmoid Perceptron, Sigmoid N.Net, Gauss/Joint BC, Gauss Naïve BC, N.Neigh, Bayes Net Based BC, Cascade Correlation, GMM-BC |
|  | Joint DE<br>Naïve DE | Gauss DE | | Joint DE, Naïve DE, Gauss/Joint DE, Gauss Naïve DE, Bayes Net Structure Learning, GMMs |
|  | | | | Linear Regression, Polynomial Regression, Perceptron, Neural Net, N.Neigh, Kernel, LWR, RBFs, Robust Regression, Cascade Correlation, Regression Trees, GMDH, Multilinear Interp, MARS |

# Assignment 2

- Repeat Exercise 1.6.1 not using a generated data as shown, but using data you decided for your project.