بنام خدا

سمینار درس بازشناسی الگو

کاربرد گوسیها و مخلوط آنها در بازشناسی الگو

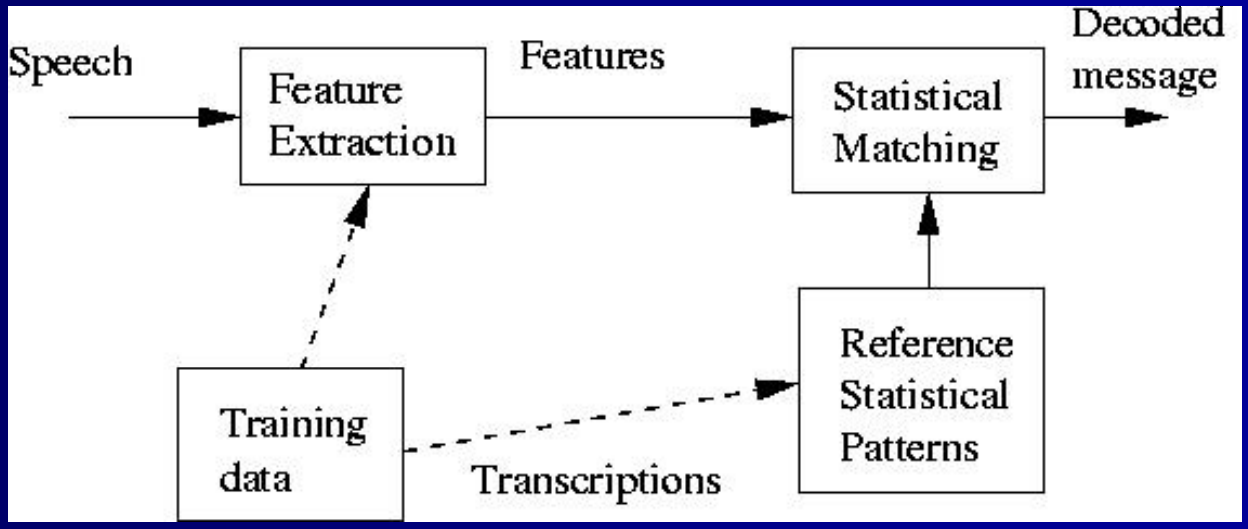
استاد: دکتر کبیر

ارایه: سامان پروانه

فهرست مطالب:

- دلیل استفاده از مدلسازی در بازشناسی
- تابع گوسی
- تابع چگالی تابع گوسی
- تخمین پارامترهای تابع گوسی
- مخلوط گوسیها
- تخمین پارامترها
- بررسی همسانی

چرا داده‌ها را مدلسازی می‌کنیم؟

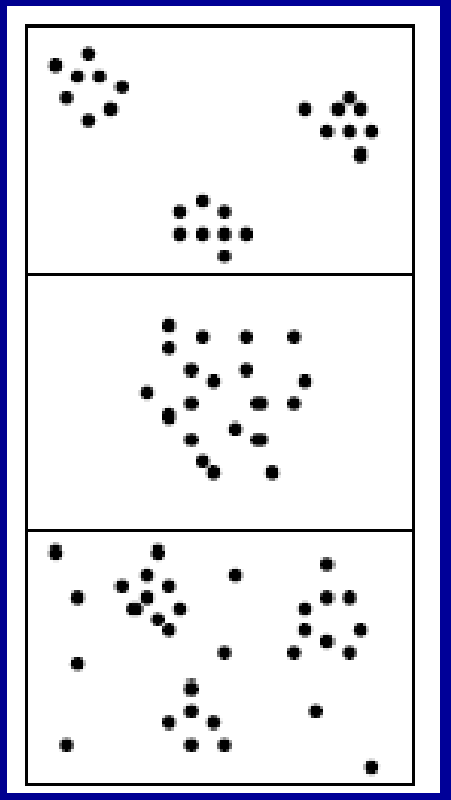


آموزش بدون ناظر:

بعضی اوقات آسان

بعضی اوقات غیرممکن

بعضی اوقات در وضعیت بینابینی

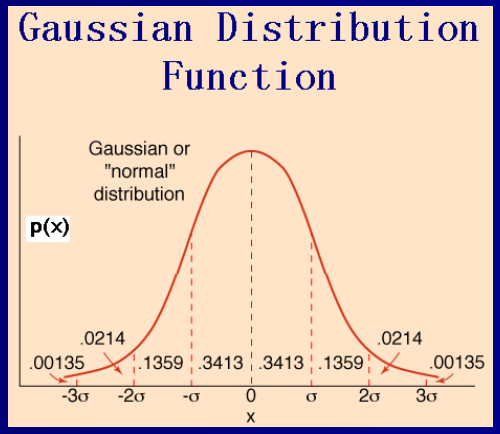


تابع گوسی:

- یک توزیع متدول است:

- بر طبق تئوری حد مرکزی: جمع متغیرهای مستقل به گوسی میل می‌کند.
- نویز و خطا معمولاً با توزیع گوسی مدل می‌شود.
- معمولاً وقتی اطلاعات دقیق وجود ندارد از این توزیع استفاده می‌شود برای مثال در بازدهٔ بورس، قد مردان و ...

- این توزیع، توزیع نرمال نیز نامیده می‌شود و دارای شکل زنگی-شکل است.



تابع چگالی احتمال گوسی:

X یک متغیر تصادفی d-بعدی است. $\mathbf{x} \in \mathbb{R}^d$

$$g_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^d \sqrt{\det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

ماتریس کوواریانس

بردار میانگین

$\mu_i = E(x_i)$ شامل میانگین هر بعد است. $\boldsymbol{\mu}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$

c_{ii} واریانس

c_{ij} کوواریانس

تابع چگالی احتمال گوسی (ادامه):

$$c_{ij} = E\left((x_i - \mu_i)^{\mathsf{T}}(x_j - \mu_j)\right)$$

x_i and x_j, $i \neq j$ → $c_{ij} = 0$ → متعامد

چه موقع ماتریس کوواریانس قطری است؟

$\sqrt{\Sigma}$: انحراف استاندارد متغیر تصادفی X است.

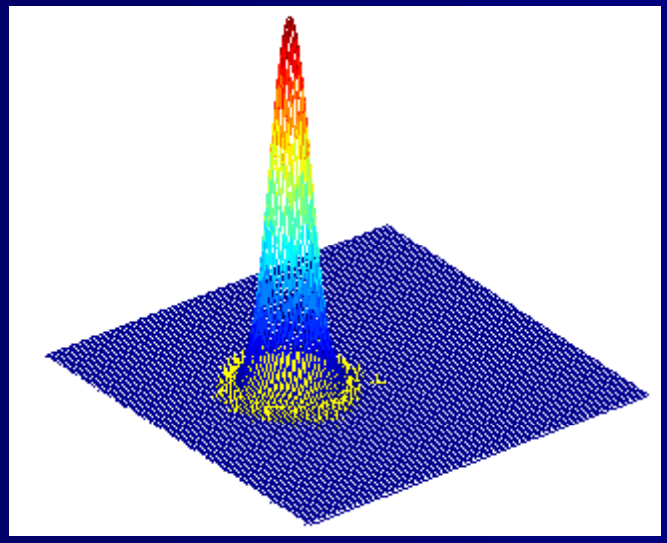
$\mathbf{x} \supset \mathcal{N}(\mathbf{0}, \mathbf{I})$ → $\mathbf{y} = \boldsymbol{\mu} + \sqrt{\boldsymbol{\Sigma}}\,\mathbf{x}$ → $\mathbf{y} \supset \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

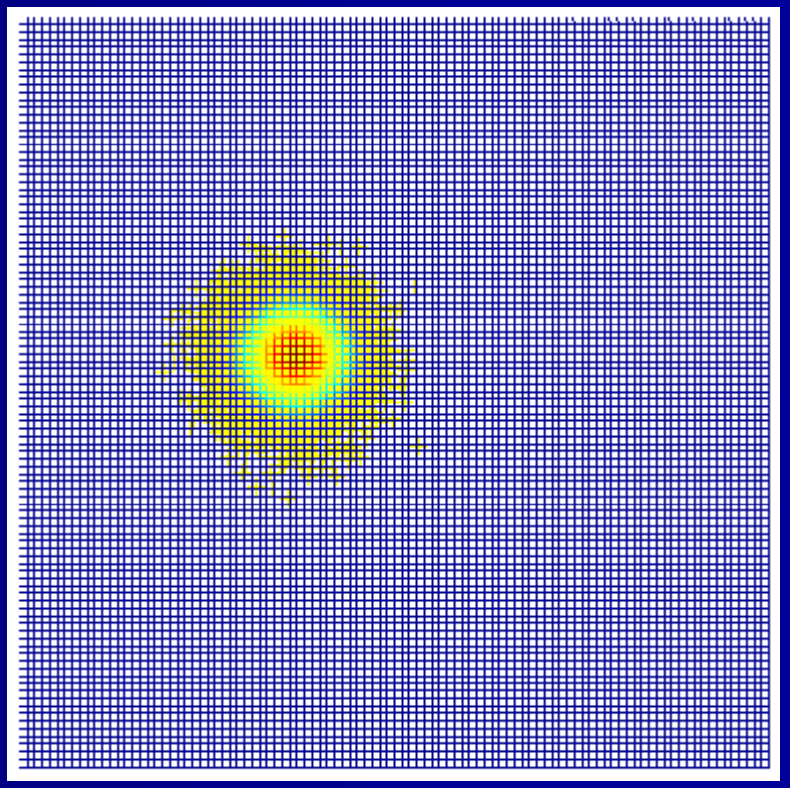
نمونهٔ فرآیند دایره‌ای:

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

$$\mu = \begin{bmatrix} 730 \\ 1090 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 8000 & 0 \\ 0 & 8000 \end{bmatrix}$$



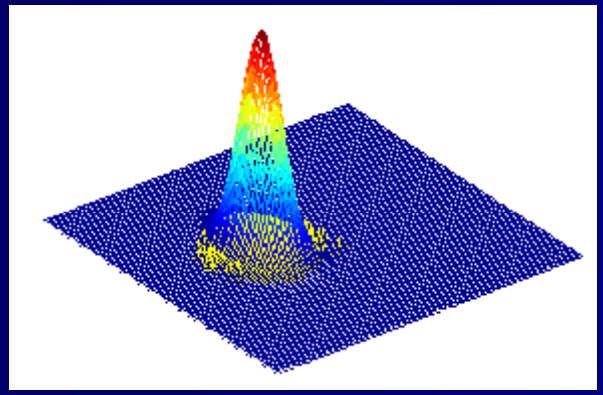


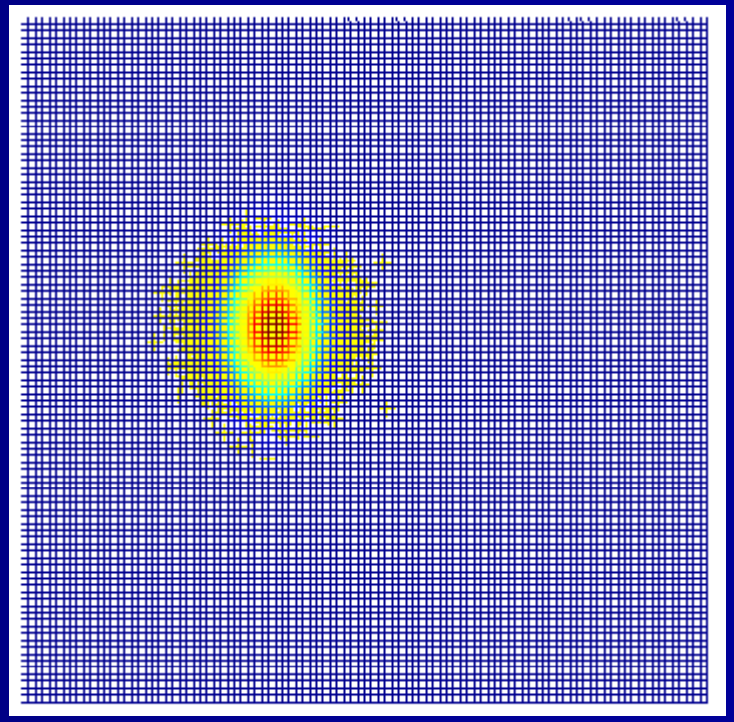
نمونهٔ کوواریانس قطری :

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

$$\mu = \begin{bmatrix} 730 \\ 1090 \end{bmatrix}$$

$$\Sigma_2 = \begin{bmatrix} 8000 & 0 \\ 0 & 18500 \end{bmatrix}$$



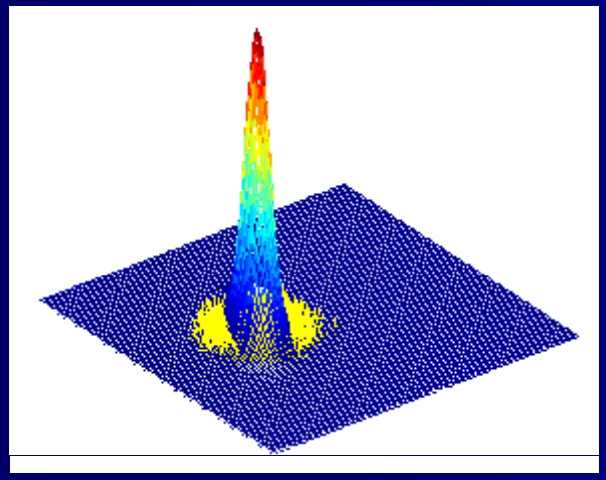


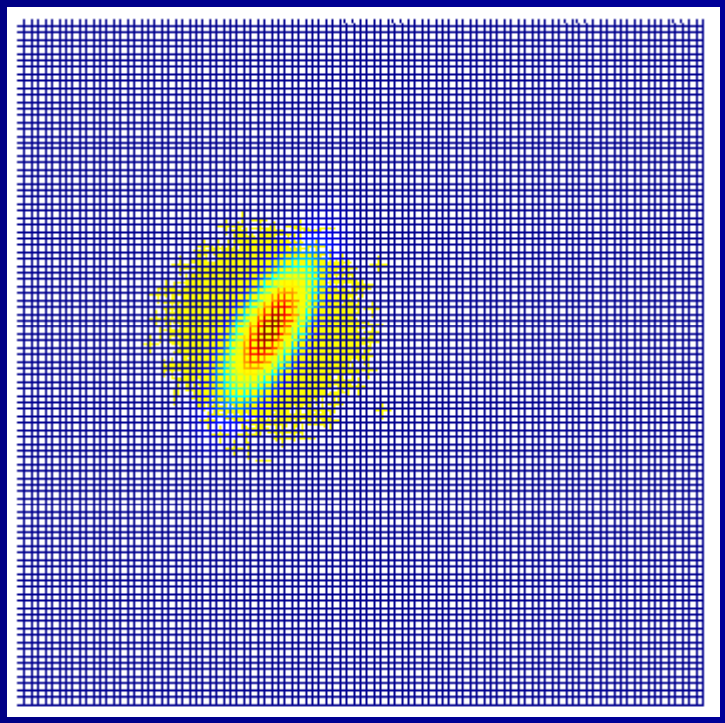
نمونهٔ کوواریانس کامل:

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

$$\mu = \begin{bmatrix} 730 \\ 1090 \end{bmatrix}$$

$$\Sigma_3 = \begin{bmatrix} 8000 & 8400 \\ 8400 & 18500 \end{bmatrix}$$





تخمین پارامترهای تابع گوسی از داده:

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

تخمین‌گر میانگین:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

→ `mean(X)`

تخمین‌گر بدون بایاس کوواریانس:

$$\hat{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)^{\mathsf{T}}(\mathbf{x}_i - \mu)$$

→ `cov(X)`

همسانی یک نمونه نسبت به مدل گوسی:

همسانی (Likelihood) برای طبقه‌بندی استفاده می‌شود.

-با معلوم بودن یک مدل تولید داده (معلوم بودن پارامترها) همسانی برابر است با:

$$p(\mathbf{x}_i|\Theta)$$

-برای مدل گوسی معادل با ارزیابی رابطهٔ زیر است:

$$g_{(\mu,\Sigma)}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^{\mathbf{d}}\sqrt{\det(\Sigma)}}\, e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)}$$

- همسانی همزمان (Joint Likelihood): برای یک مجموعه از نمونه‌هایی که بصورت یکنواخت توزیع شده‌اند، حاصلضرب همسانی هر کدام از نمونه‌ها است.

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \rightarrow p(X|\Theta) = \prod_{i=1}^{N} p(\mathbf{x}_i|\Theta) = \prod_{i=1}^{N} p(\mathbf{x}_i|\mu,\Sigma) = \prod_{i=1}^{N} g_{(\mu,\Sigma)}(\mathbf{x}_i)$$

Log-likelihood بجای Likelihood:

۱- حاصلضرب را به حاصلجمع تبدیل می‌کند.

$$p(X|\Theta) = \prod_{i=1}^{N} p(\mathbf{x}_i|\Theta) \quad \Leftrightarrow \quad \log p(X|\Theta) = \log \prod_{i=1}^{N} p(\mathbf{x}_i|\Theta) = \sum_{i=1}^{N} \log p(\mathbf{x}_i|\Theta)$$

۲- در حالت گوسی، از محاسبات نمایی جلوگیری می‌کند.

$$\begin{aligned} p(\mathbf{x}|\Theta) &= \frac{1}{\sqrt{2\pi}^d \sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)} \\ \log p(\mathbf{x}|\Theta) &= \frac{1}{2}\left[-d\log(2\pi) - \log(\det(\Sigma)) - (\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right] \end{aligned}$$

چون Log-likelihood، صعودی یکنوا است در نتیجه روابط مشابه Likelihood است و می‌تواند بصورت مستقیم برای طبقه‌بندی استفاده بشود.

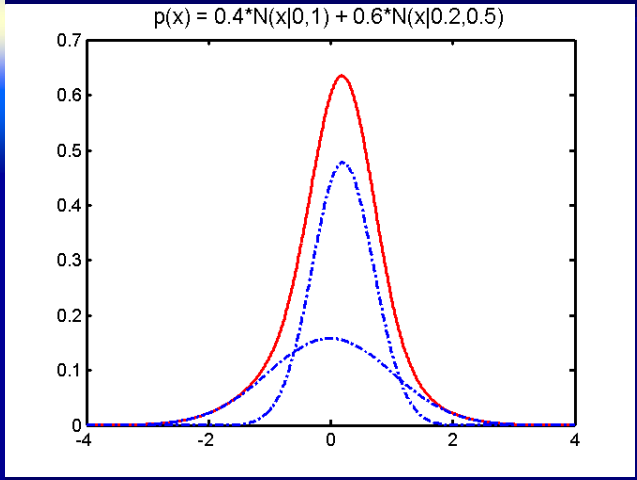
$$p(x|\Theta_1) > p(x|\Theta_2) \quad \Leftrightarrow \quad \log p(x|\Theta_1) > \log p(x|\Theta_2)$$

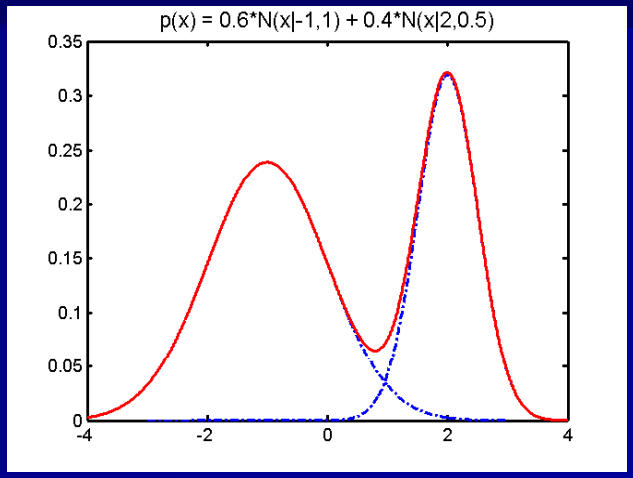
مدل مخلوط گوسی (GMM):

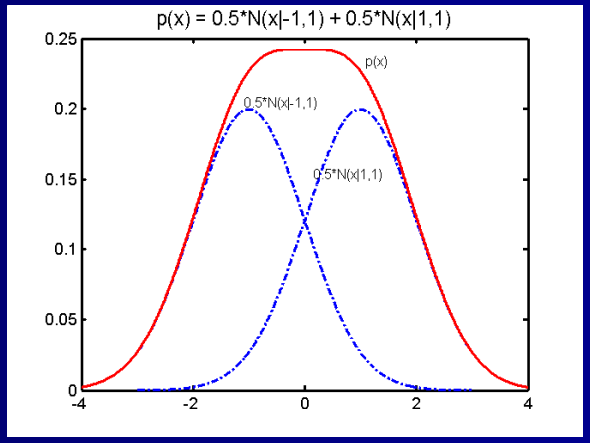
- مدل مخلوط گوسیها
  - بر اساس مدلسازی آماری خوشه‌ها است.
  - خوشه بندی←یک مسئلهٔ تخمین پارامتر

- مشابه K-means که جمع فاصلهٔ نمونه‌ها تا مراکز خوشه را مینیمم می‌کرد، از الگوریتم EM استفاده می‌کنیم تا همسانی (Likelihood) مدل مخلوط گوسی را ماکزیمم کنیم.

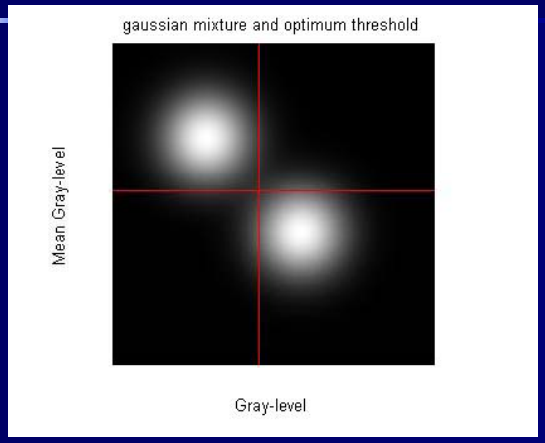
چند مخلوط گوسی نمونه (حالت یک‌بعدی):

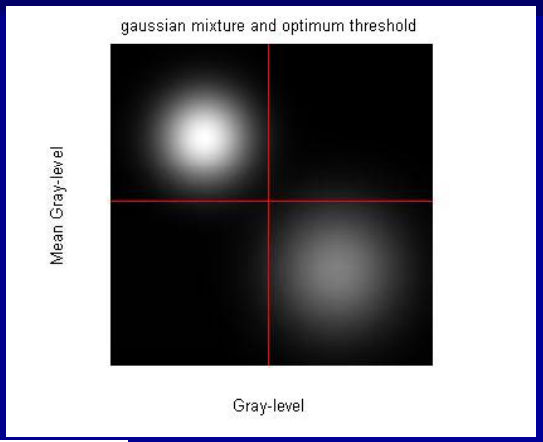


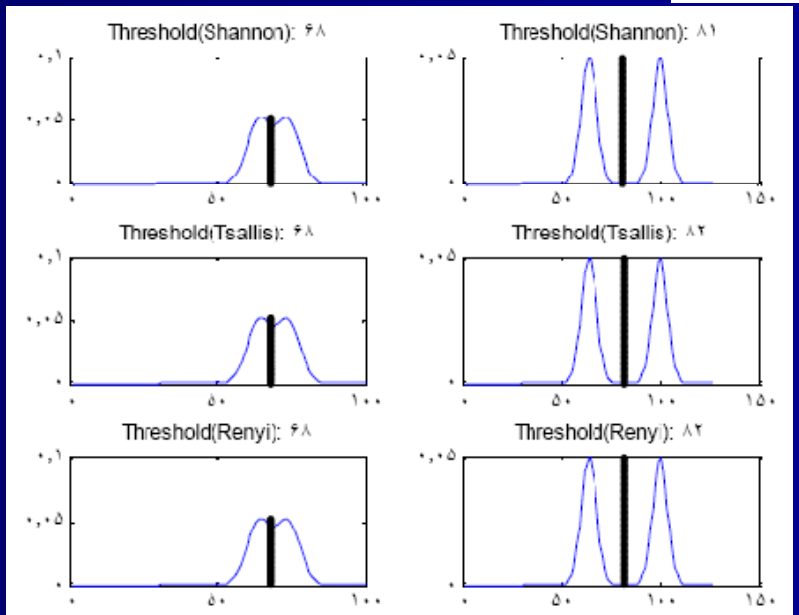




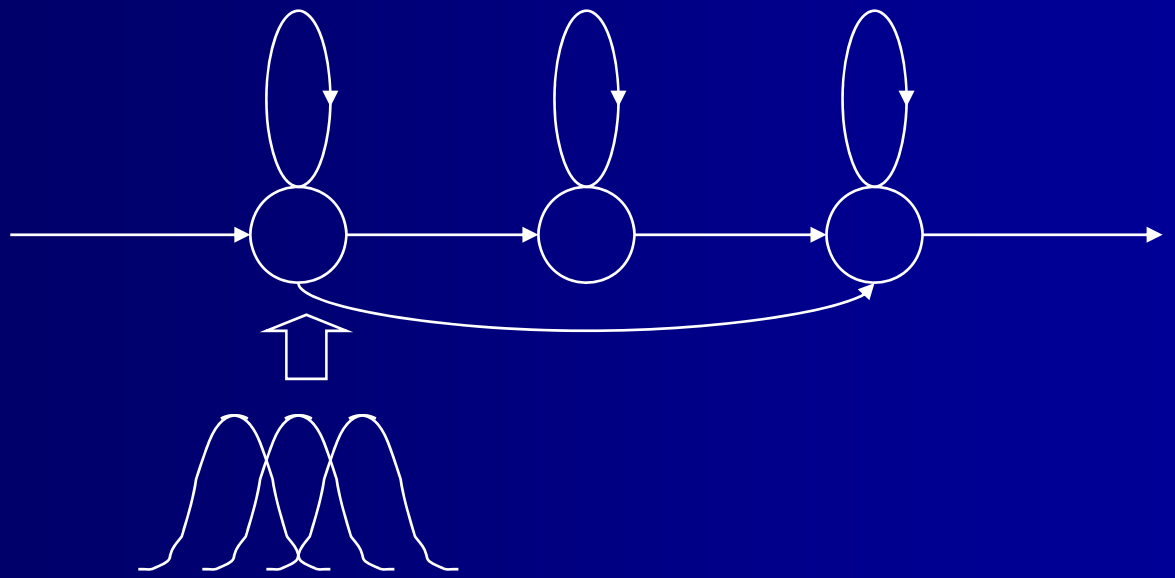
کاربرد مخلوط گوسی در پردازش تصویر:







کاربرد مخلوط گوسی در HMM:



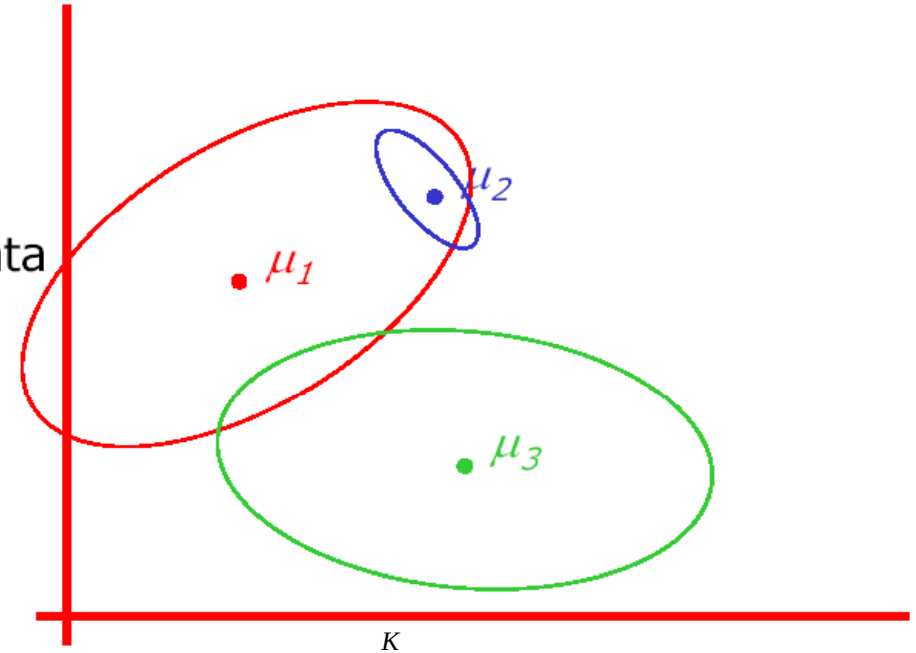
مخلوط گوسی:

The General GMM assumption

- There are k components. The i'th component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix Σ_i

Assume that each datapoint is generated according to the following recipe:

1. Pick a component at random. Choose component i with probability $P(\omega_i)$.
2. Datapoint ~ $N(\mu_i, \Sigma_i)$

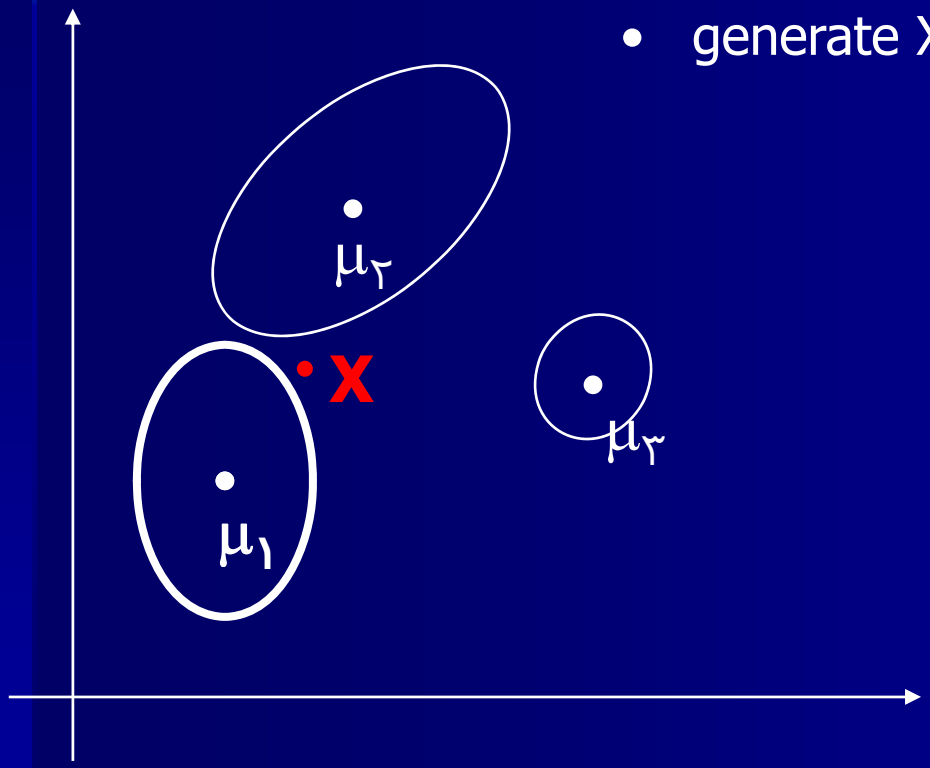


$$p(x) = \sum_{i=1}^{K} P(w_i) N(x \mid \mu_i, \Sigma_i)$$

$$P(w_i) > 0, \qquad \sum_{i=1}^{K} P(w_i) = 1$$

مخلوط گوسی (ادامه):

- choose component with probability w_k
- generate $X \sim N(\mu_k, \sigma_k)$



چند نکته دربارهٔ مخلوط گوسی:

- تابع چگالی p(x) جمع (مخلوط) وزن‌یافتهٔ تعدادی تابع چگالی گوسی است.
- این جمع توابع چگالی با متغیری که جمع وزن‌یافتهٔ متغیرهای دیگر است متفاوت می‌باشد: $Y = aX_1 + bX_2$

ترکیب مخلوط گوسی:

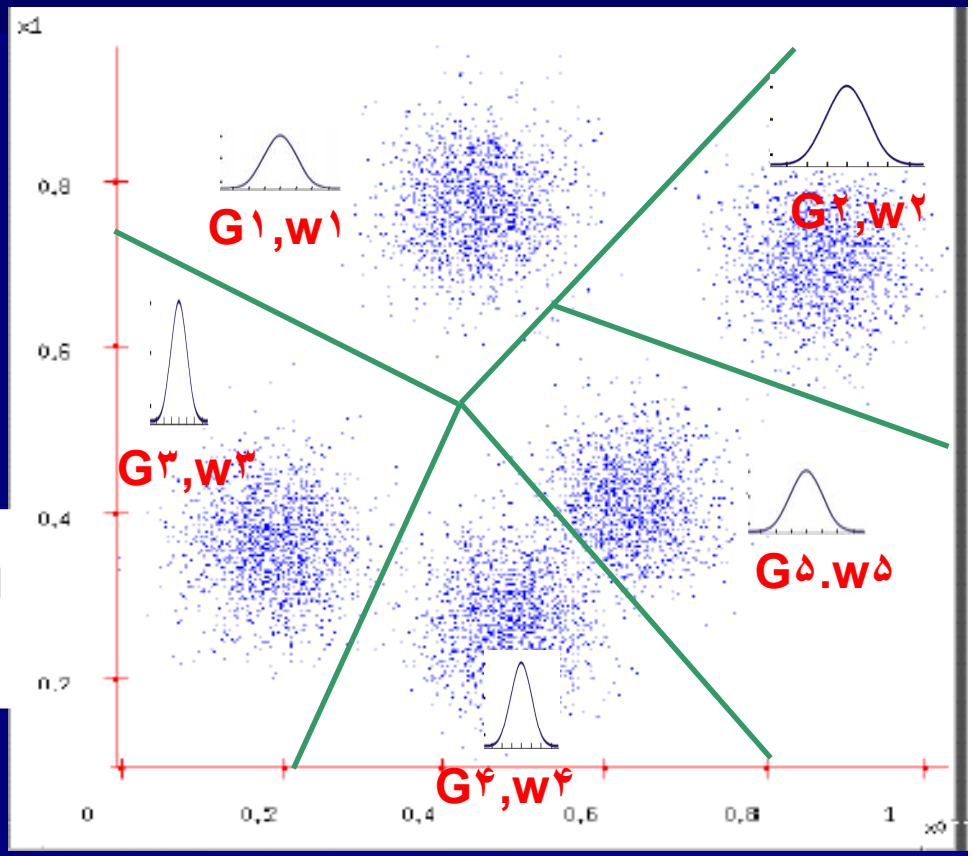
$$P(C_j \mid X) = P(X \mid C_j) \cdot \frac{P(C_j)}{P(X)}$$

$$P(X \mid C_j) = \sum_{k=1}^{Nc} w_k G_k$$

$$G_k \equiv p(X \mid G_i) = \frac{1}{(2\pi)^{d/2} \mid V_i \mid^{1/2}} \cdot e^{[-1/2(X-\mu_i)^T Vi^{-1}(X-\mu_i)]}$$

Variables: μ_i, V_i, w_k

We use EM (estimate-maximize) algorithm to approximate this variables.



تعداد گوسیها چند عدد باشد؟

Problem of scoring models with different complexities

- Model too flexible $\Rightarrow$ overfit the data $\Rightarrow$ high variance
- Model too restrictive $\Rightarrow$ can't fit the data $\Rightarrow$ high bias
- Bias-variance tradeoff: compromise

احتمال مشترک و شرطی:

اگر دو متغیر X و Y را داشته باشیم و بخواهیم احتمالاتی که شامل X و Y است را ترکیب کنیم:

- احتمال مشترک (Joint Probability): آنرا با $p(x,y)$ نمایش می‌دهیم و احتمال این است که X مقدار x^1 را بگیرد و Y مقدار y^1 را بگیرد. در اصل احتمال این است که هر دوی این حالتها رخ بدهند:

  X برابر x^1 بشود و Y برابر y^1 بشود.

- احتمال شرطی (Joint Probability): آنرا با $p(x/y)$ نمایش می‌دهیم و بدین مفهوم است که درصورتیکه Y مقدار y را گرفت، X مقدار x را بگیرد (صرفنظر از غیرممکن بودن اینکه Y مقدار y را بگیرد)

- رابطه: $p(x,y) = p(x|y)p(y)$

اهمیت احتمال شرطی:

- احتمال شرطی در حوزه‌های مختلفی از جمله طبقه‌بندی، تئوری تصمی‌گیری، پیش‌بینی و موقعیتهای مشابه دارای اهمیت است.
- دلیل این اهمیت در کاربردهای بالا این است که طبقه‌بندی، تصمیم‌گیری، پیش‌بینی و ... را بر اساس یک ملاک انجام می‌دهیم. بنابراین می‌خواهیم احتمال یک نتیجه را با معلوم بودن ملاک بدانیم. $p(r/e)$.

- در طبقه‌بندی، ملاک مقادیر اندازه‌گیریها یا ویژگیهای x است که طبقه‌بندی باید بر اساس آنها باشد. نتایج ممکن نیز، کلاسها (w) هستند .

- مشکل در این است که تخمین احتمال شرطی $p(w_i/x)$ بطور مستقیم از روی داده خیلی سخت است.

تخمین GMM:

برای GMM در حالت کلی log-likelihood برابر است با:

$$L(\Theta)=\sum_{i=1}^{N}\log p(x_i\mid\Theta)=\sum_{i=1}^{N}\log\left[\sum_{j=1}^{K}P(w_j)N(x_i\mid\mu_j,\Sigma_j)\right]$$

$$\Theta=\{P(w_1),\ldots,P(w_K),\mu_1,\ldots,\mu_K,\Sigma_1,\ldots,\Sigma_K\}$$

همسانی را با کمک رابطهٔ $\partial L/\partial\Theta=0$ ماکزیمم می‌کنیم ولی باید توجه داشت که حل آن سخت است.

راه‌حل: از الگوریتم ساده شدهٔ EM (Expectation-Maximization) استفاده می‌کنیم.

- مرحلهٔ E (Expectation) کمک می‌کند تا اطلاعات ناکامل را بازیابی کنیم.
- در مرحلهٔ M (Maximization)، همسانی را با استفاده از اطلاعات بازیابی شده ماکزیمم می‌کنیم.

الگوریتم EM:

دادهٔ آموزش ناکامل است.(χ)

مجموعه دادهٔ کامل: $\{\chi, \gamma\}$ متغیرهای مخفی: γ

همسانی دادهٔ کامل: $L(\theta \mid \chi, \gamma)$ که تابع γ است.

مثال ساده (با اطلاعات کامل):

اگر وقایع "نمرات یک کلاس" باشد:

w_1 = Gets an A $P(A) = ½$

w_2 = Gets a B $P(B) = \mu$

w_3 = Gets a C $P(C) = 2\mu$

w_4 = Gets a D $P(D) = ½ - 3\mu$

(دقت: $0 \le \mu \le 1/6$)

فرض کنید می‌خواهیم μ را از داده تخمین بزنیم. در یک کلاس مفروض داریم:

a A's

b B's

c C's

d D's

تخمین بیشترین همسانی برای μ با معلوم بودن a,b,c,d چقدر است؟

با ML براحتی حل می‌شود!

ML = Maximum Likelihood

P(A) = ½ P(B) = μ P(C) = ۲μ P(D) = ½-۳μ

$P(a,b,c,d \mid \mu) = K(½)^a(\mu)^b(۲\mu)^c(½-۳\mu)^d$

$\log P(a,b,c,d \mid \mu) = \log K + a\log ½ + b\log \mu + c\log ۲\mu + d\log (½-۳\mu)$

FOR MLE μ, SET $\frac{\partial \text{LogP}}{\partial \mu} = ۰$

$$\frac{\partial \text{LogP}}{\partial \mu} = \frac{b}{\mu} + \frac{۲c}{۲\mu} - \frac{۳d}{۱/۲ - ۳\mu} = ۰$$

Gives max like $\mu = \frac{b+c}{۶(b+c+d)}$

A	B	C	D
۱۴	۶	۹	۱۰

So if class got

MLE $\mu = \frac{۱}{۱۰}$

مثال ساده (با اطلاعات مخفی):

Someone tells us that

Number of High grades (A's + B's) = *h*

Number of C's = *c*

Number of D's = *d*

What is the max. like estimate of μ now?

REMEMBER

P(A) = ½

P(B) = μ

P(C) = ۲μ

P(D) = ½-۳μ

EXPECTATION

If we know the value of μ we could compute the expected value of *a* and *b*

Since the ratio a:b should be the same as the ratio ½ : μ

$$a = \frac{\frac{1}{2}}{\frac{1}{2}+\mu} h \qquad b = \frac{\mu}{\frac{1}{2}+\mu} h$$

MAXIMIZATION

If we know the expected values of *a* and *b* we could compute the maximum likelihood value of μ

$$\mu = \frac{b+c}{6(b+c+d)}$$

کاربرد EM برای مثال بدیهی:

We begin with a guess for μ

We iterate between EXPECTATION and MAXIMALIZATION to improve our estimates of μ and *a* and *b*.

Define μ(t) the estimate of μ on the t'th iteration

b(t) the estimate of *b* on t'th iteration

REMEMBER

P(A) = ½

P(B) = μ

P(C) = 2μ

P(D) = ½-3μ

$$\mu(0) = \text{initial guess}$$

$$b(t) = \frac{\mu(t)h}{1/2 + \mu(t)} = \mathrm{E}[b \mid \mu(t)]$$

E-step

$$\mu(t+1) = \frac{b(t)+c}{6(b(t)+c+d)}$$

$$= \max \text{ like est of } \mu \text{ given } b(t)$$

M-step

Continue iterating until converged.

Good news: Converging to local optimum is assured.

Bad news: I said "local" optimum.

همگرایی EM:

- Convergence proof based on fact that Prob(data | μ) must increase or remain same between each iteration [NOT OBVIOUS]
- But it can never exceed ١ [OBVIOUS]

So it must therefore converge [OBVIOUS]

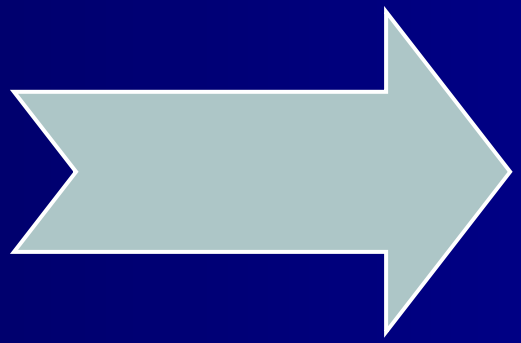
In our example, suppose we had

h = ٢٠

c = ١٠

d = ١٠

μ(٠) = ٠



t	μ(t)	b(t)
٠	٠	٠
١	٠.٠٨٣٣	٢.٨٥٧
٢	٠.٠٩٣٧	٣.١٥٨
٣	٠.٠٩٤٧	٣.١٨٥
۴	٠.٠٩۴٨	٣.١٨٧
۵	٠.٠٩۴٨	٣.١٨٧
۶	٠.٠٩۴٨	٣.١٨٧

Convergence is generally linear: error decreases by a constant factor each time step.

الگوریتم EM برای تخمین GMM:

- مرحلهٔ E: هر مشاهده بر حسب همسانی هر کدام از گوسیهای متناظر، یک وزن را برای هر خوشه انتساب می‌دهد.

$$P(w_j \mid x_i, \Theta^{(t)}) = \frac{p(x_i \mid \boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)}) P(w_j)}{\sum_{v=1}^{K} p(\mathbf{x}_i \mid \boldsymbol{\mu}_v^{(t)}, \Sigma_v^{(t)}) P(w_v)}$$

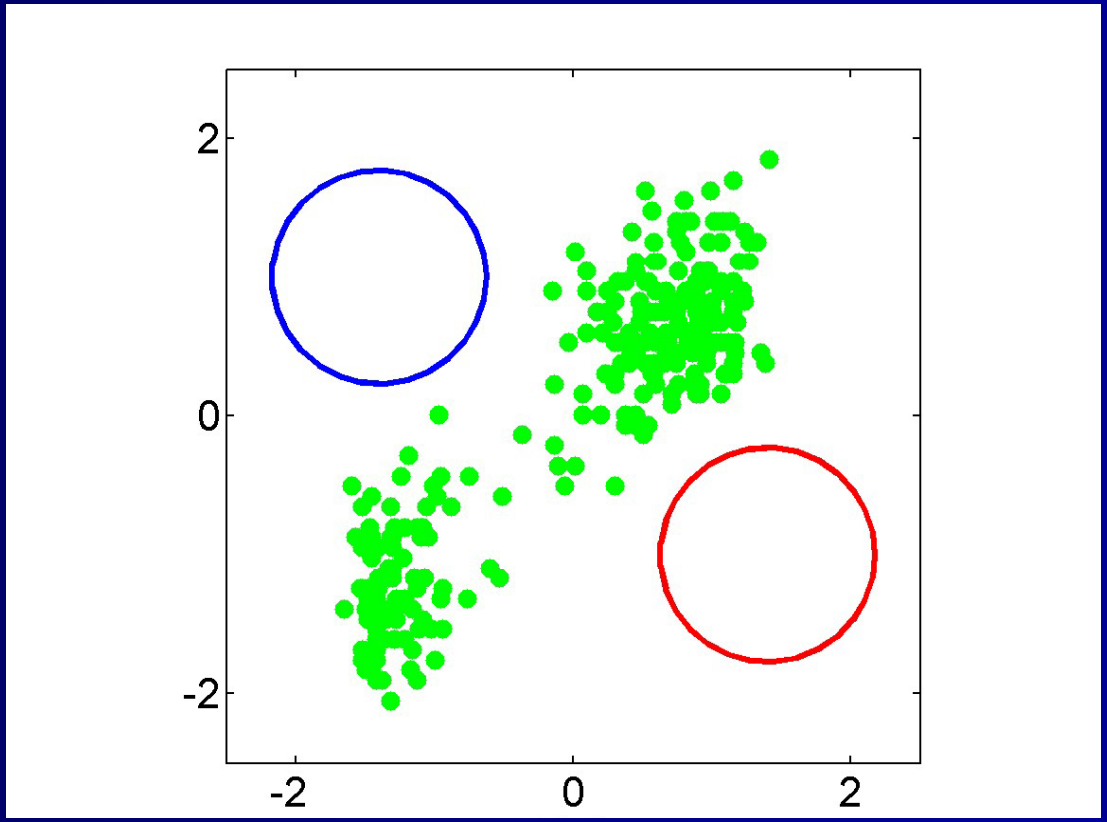
- مرحلهٔ M: هر مشاهده با استفاده از ML در میانگینها و کوواریانسهای هر خوشه مشارکت می‌کند.

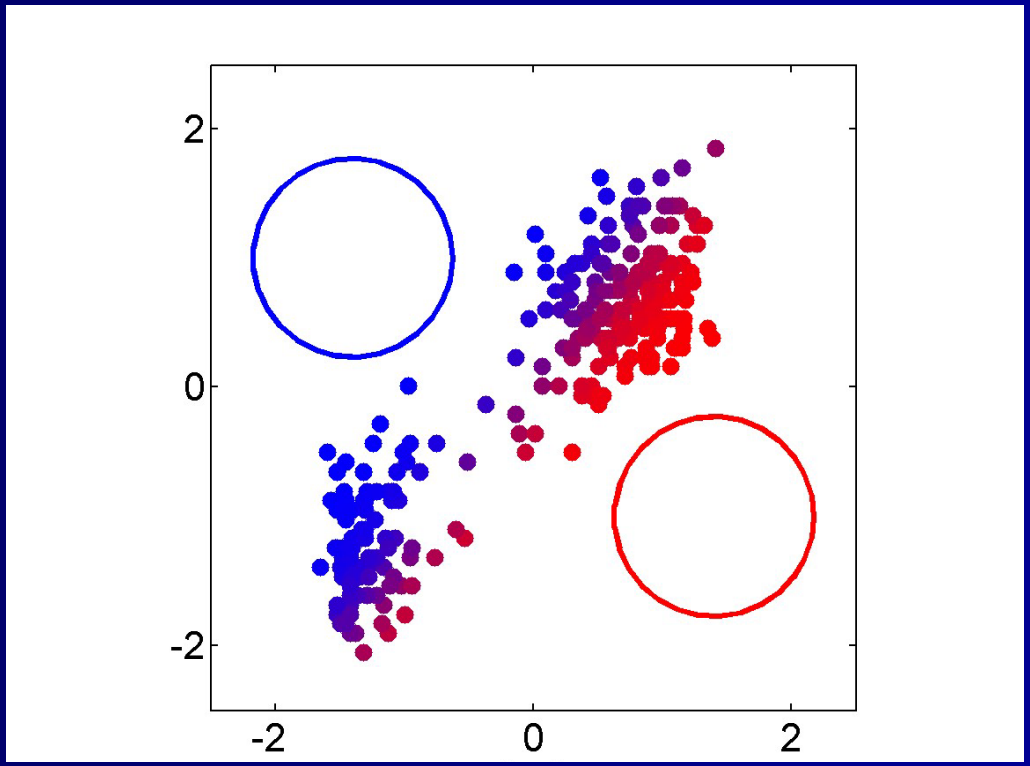
$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{N} P(w_j \mid x_i, \Theta^{(t)}) x_i}{\sum_{i=1}^{N} P(w_j \mid x_i, \Theta^{(t)})}$$

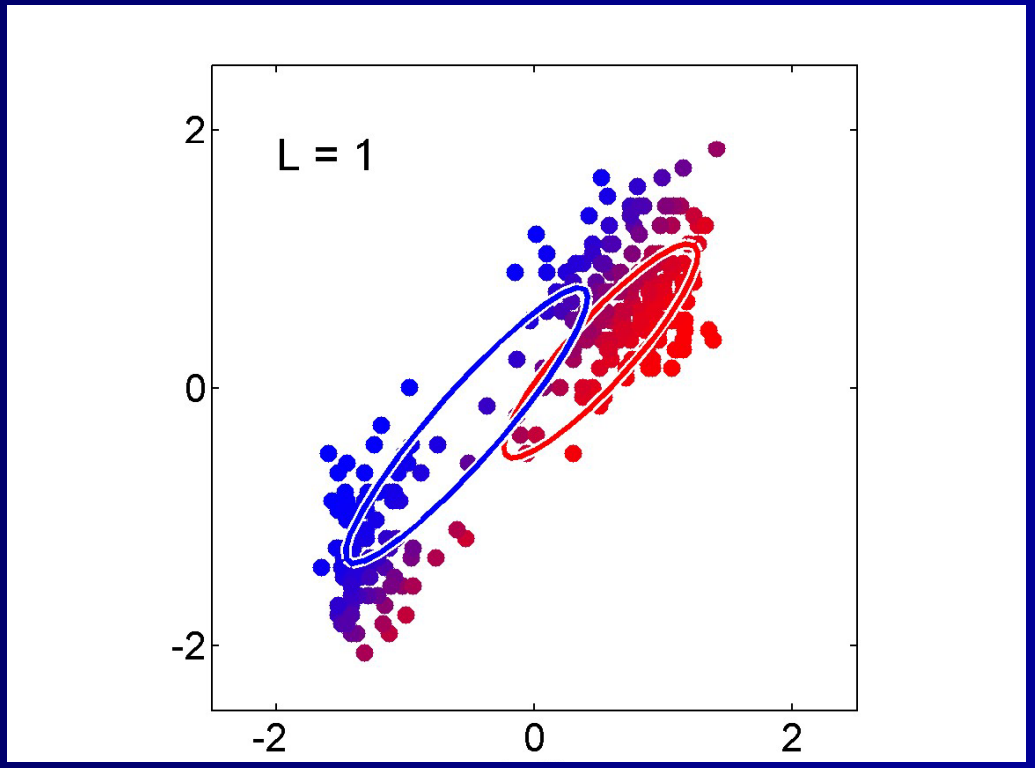
$$\hat{\Sigma}_j^{(t+1)} = \frac{1}{\sum_{i=1}^{N} P(w_j \mid x_i, \Theta^{(t)})} \sum_{i=1}^{N} P(w_j \mid x_i, \Theta^{(t)})(x_i - \hat{\boldsymbol{\mu}}_j^{(t+1)})(x_i - \hat{\boldsymbol{\mu}}_j^{(t+1)})^T$$

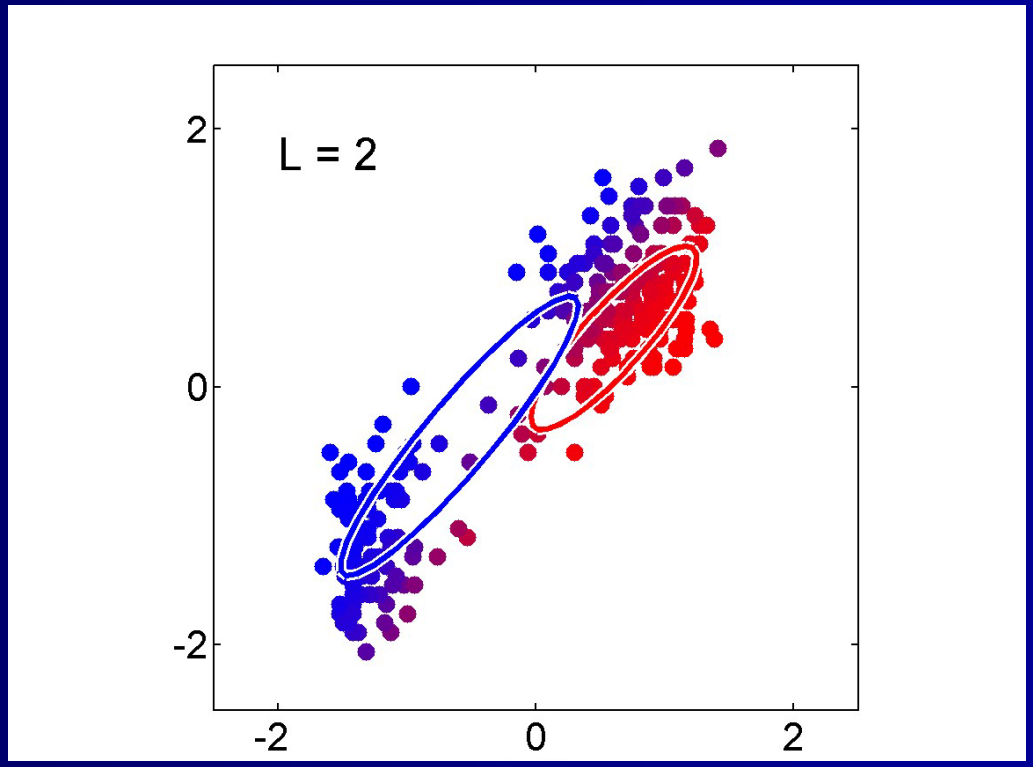
$$P(w_j)^{(t+1)} = \sum_{i=1}^{N} P(w_j \mid x_i, \Theta^{(t)}) \Big/ N$$

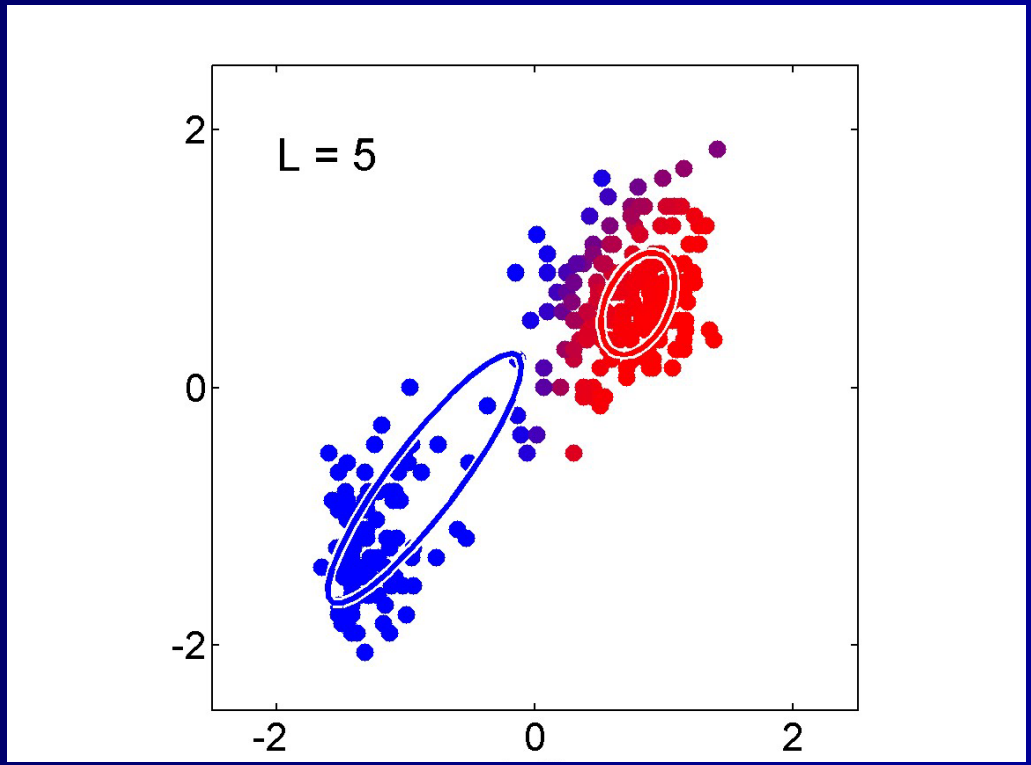
- مراحل E و M را تا همگرایی ادامه می‌دهد.

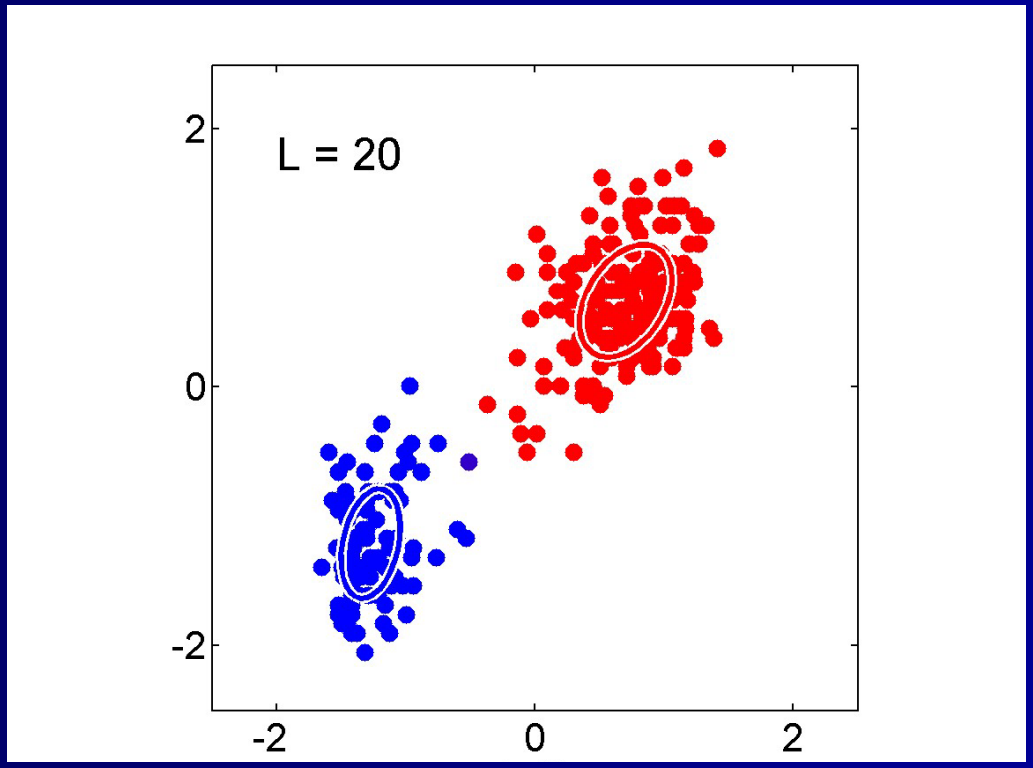

2
0
-2
-2
0
2


2
0
-2
-2
0
2


L = 1
2
0
-2
-2
0
2


L = 2
2
0
-2
-2
0
2


L = 5
2
0
-2
-2
0
2


L = 20
2
0
-2
-2
0
2

K-means بعنوان حالت خاص EM:

- با فرض $P(w_j) = ۱/K, \Sigma_j = \sigma^۲ I$ داریم:

EM → K-means

- الگوریتم K-means:

۱- مقداردهی اولیه و تصادفی k مرکز خوشه.

۲- هر داده را بر اساس معیار مینیمم فاصله به خوشه‌ها متناظر می‌کنیم.

$$P(w_j \mid x_i) = \begin{cases} ۱ & \text{if } \| x_i - \boldsymbol{\mu}_j \| \leq \| x_i - \boldsymbol{\mu}_v \| \\ ۰ & \text{otherwise} \end{cases}$$

۳- مراکز خوشه را با استفاده از میانگین مجدداً محاسبه می‌کنیم:

$$\hat{\mu}_j^{(t+۱)} = \frac{\sum_{i=۱}^{N} P(w_j \mid x_i) x_i}{\sum_{i=۱}^{N} P(w_j \mid x_i)}$$

۴- به مرحلهٔ ۲ می‌رویم تا تغییری در مراکز خوشه رخ ندهد.

خوشه‌بندی نرم و سخت:

- اغلب GMM را بعنوان روش خوشه‌بندی نرم می‌دانند.
  - $P(w_j \mid x_i, \Theta)$

احتمال پسین اینکه iامین مشاهده از jامین گوسی ایجاد بشود.

- برای هر نقطهٔ x_i، GMM احتمالات پسین $\{P(w_1 \mid x_i, \Theta), \ldots, P(w_K \mid x_i, \Theta)\}$ را پیدا می‌کند که برچسب خوشه را برای هر x_i می‌توانیم با کمک رابطهٔ زیر بیابیم:

$$C(x_i) = \arg\max_j P(w_j \mid x_i, \Theta)$$

- K-means یک روش خوشه‌بندی سخت است.
  - جواب فقط ۰ یا ۱ است.

GMM بعنوان روش خوشه‌بندی نرم:

- K-Means

R clusters per class

Iteration:

Step۱: Set each data to one cluster

{ ۰,۱}

Step۲: Tune cluster centers

- GMM

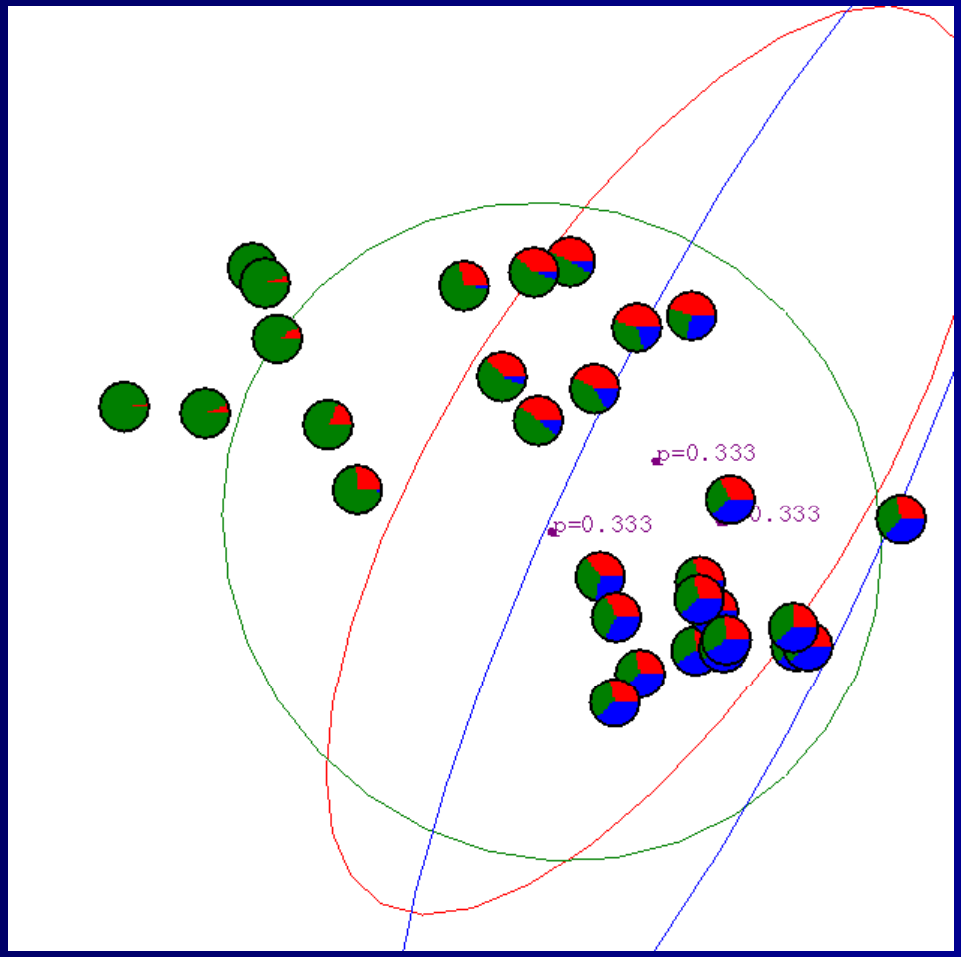
R Gaussians per class

Iteration:

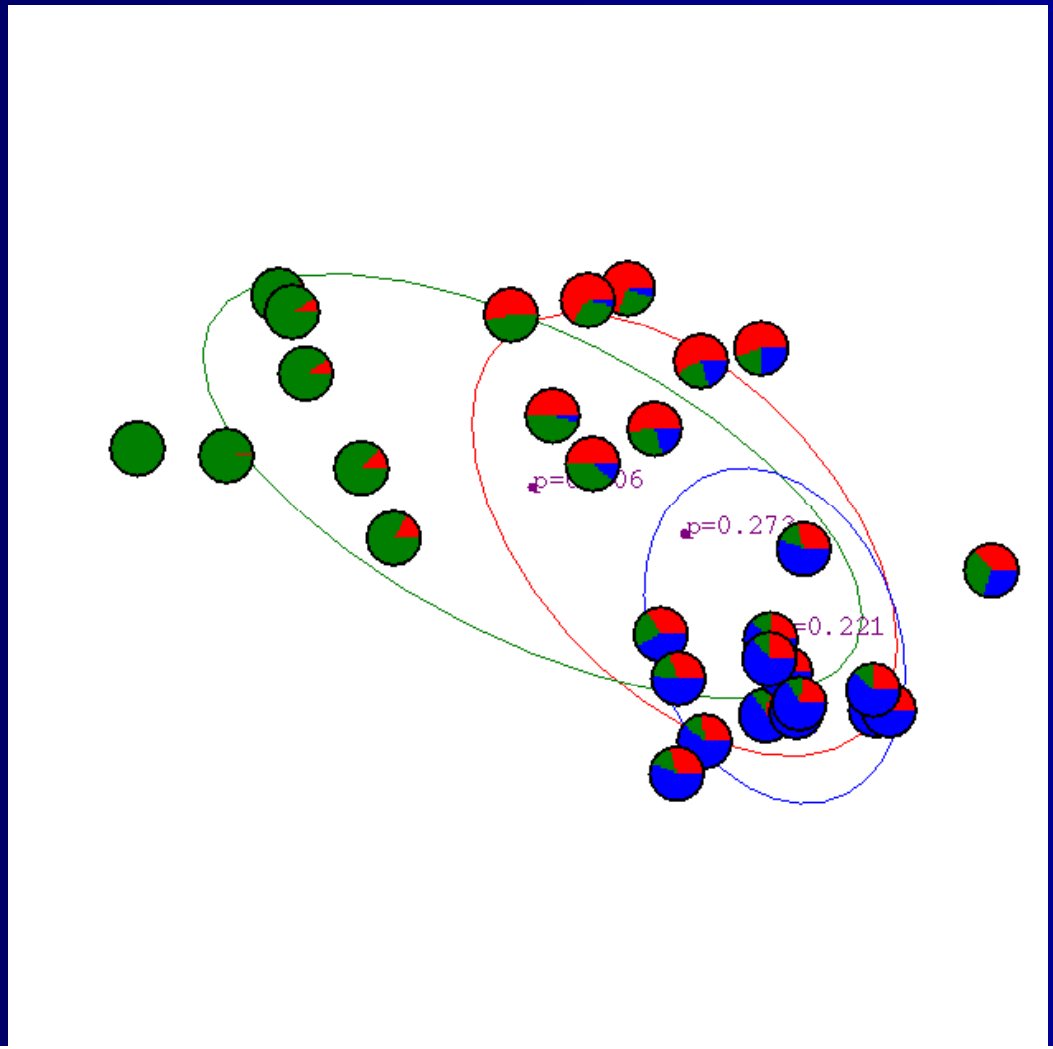
E step: Set each data's responsibilities for all clusters [۰,۱]

M step: Tune parameters of Gaussians

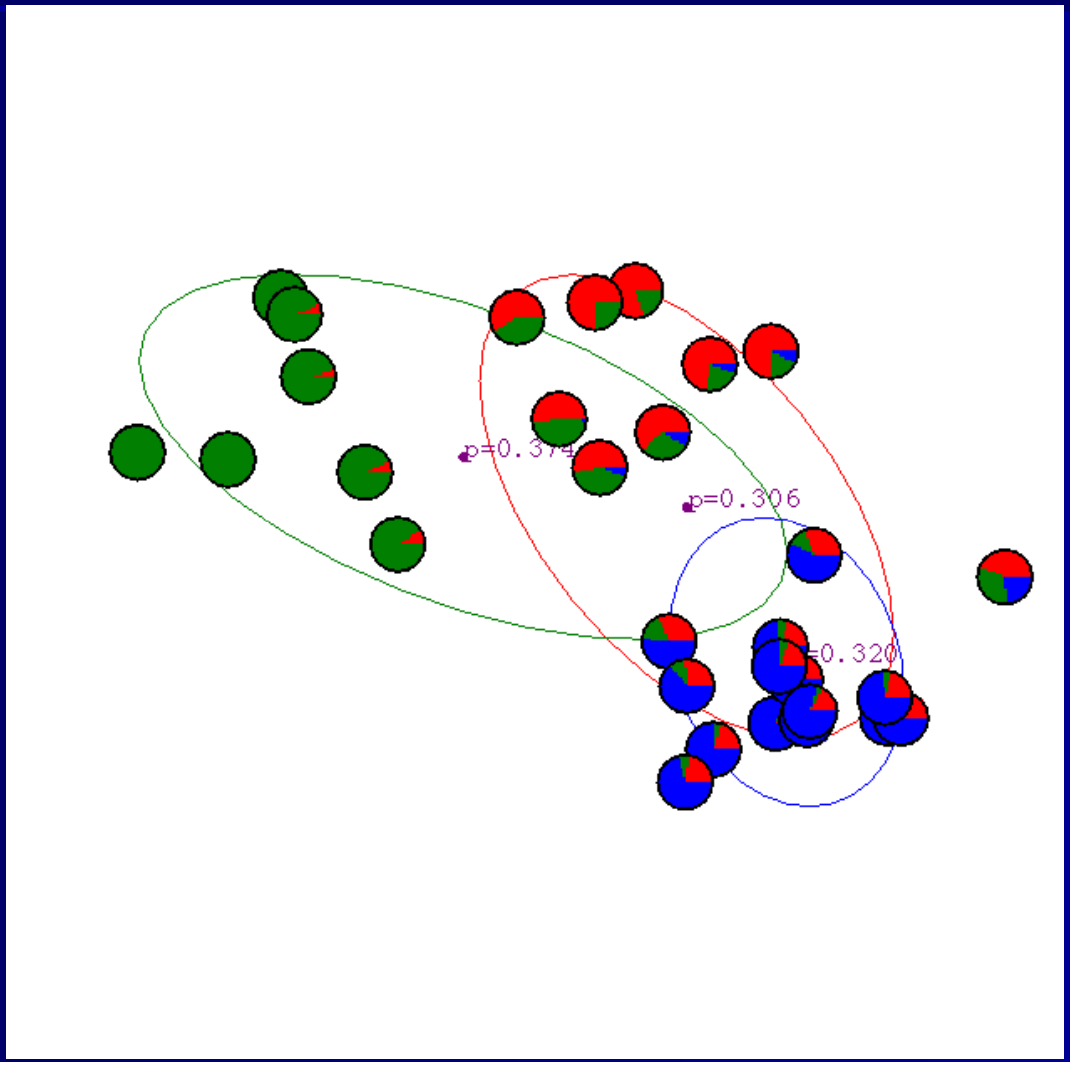
Gaussian Mixture Example: Start



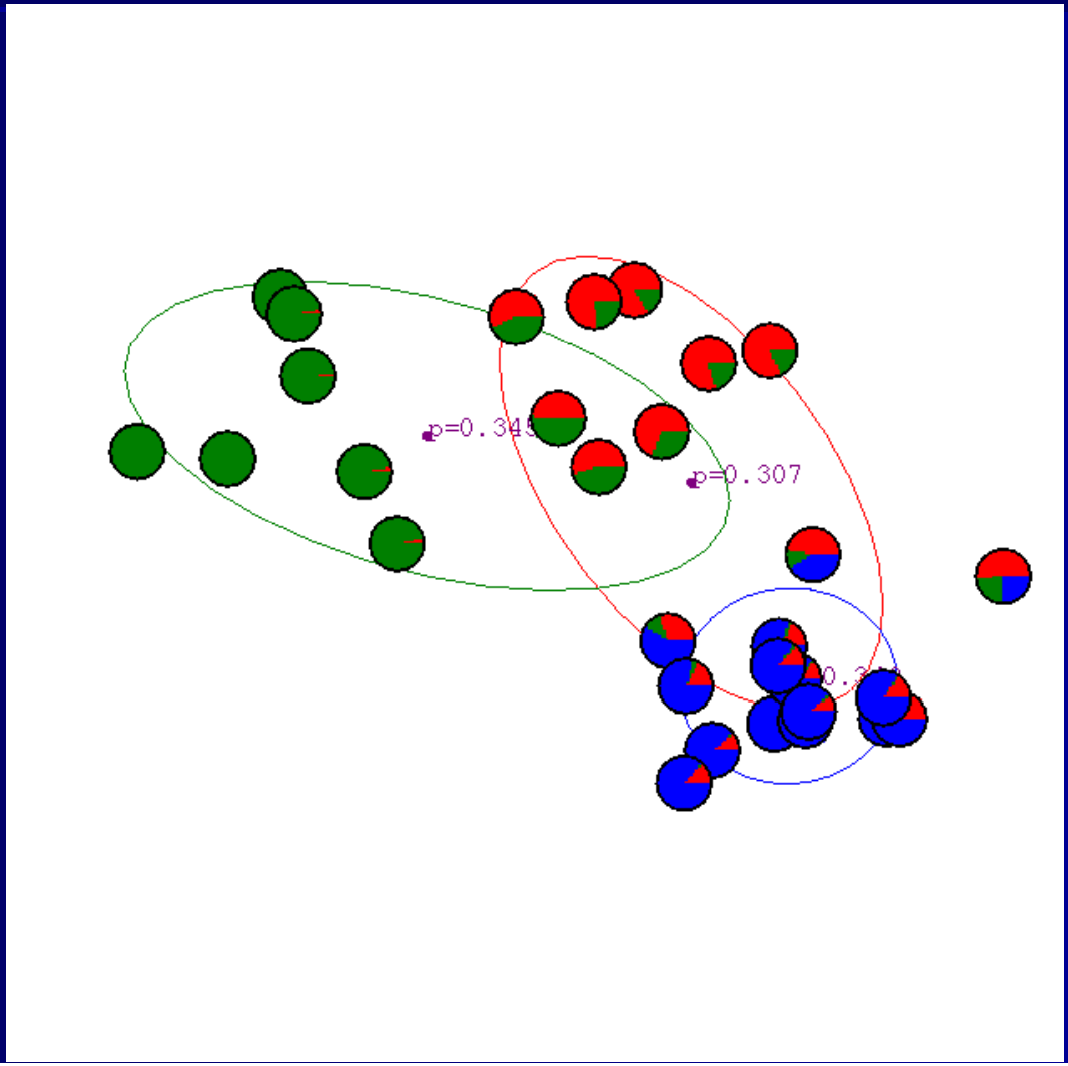
After First Iteration



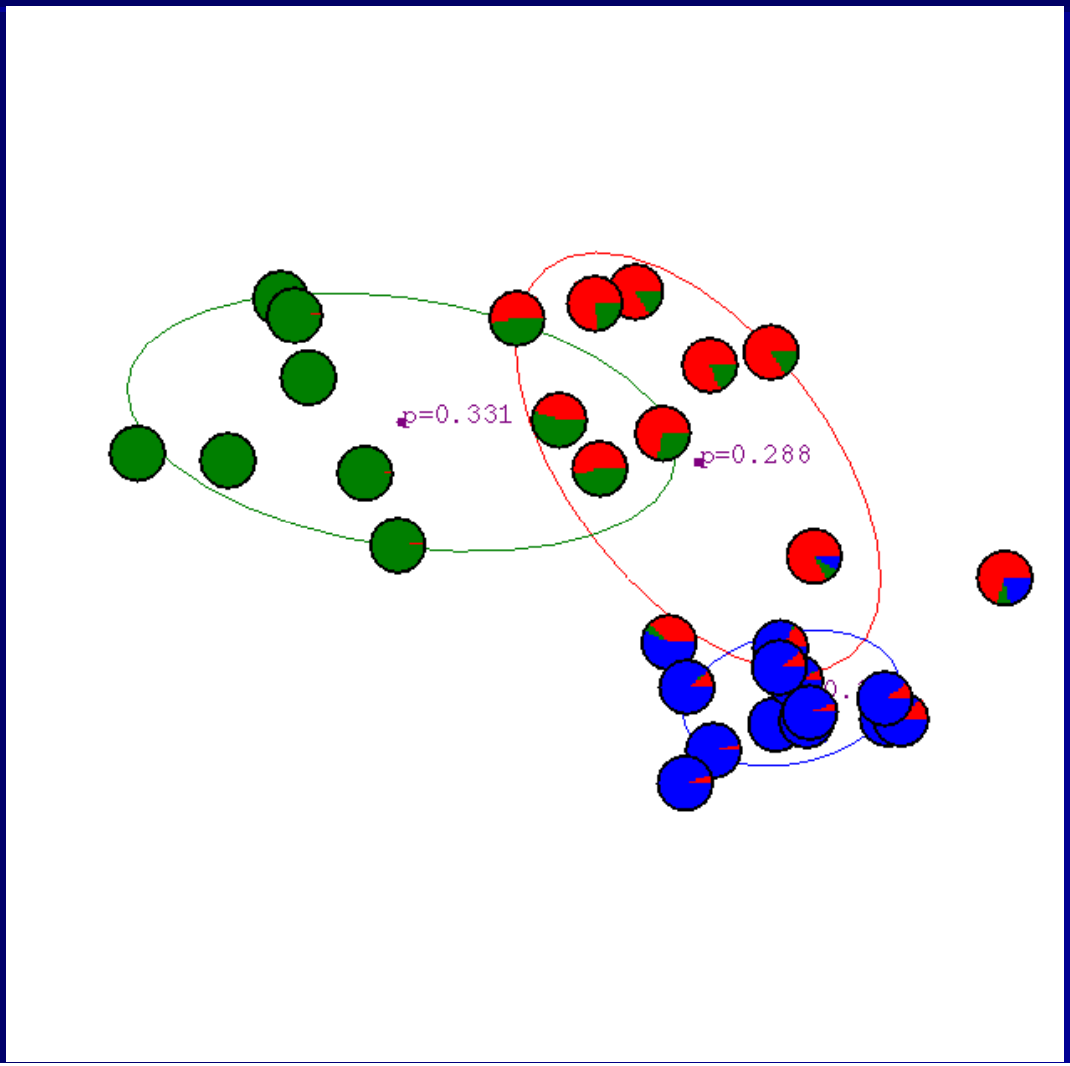
After ٢nd Iteration



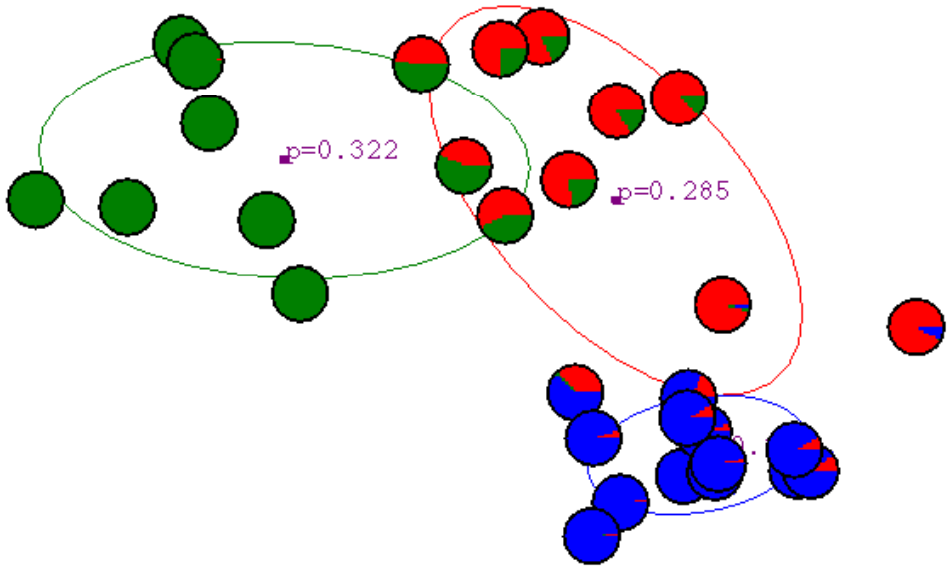
After ٣rd Iteration



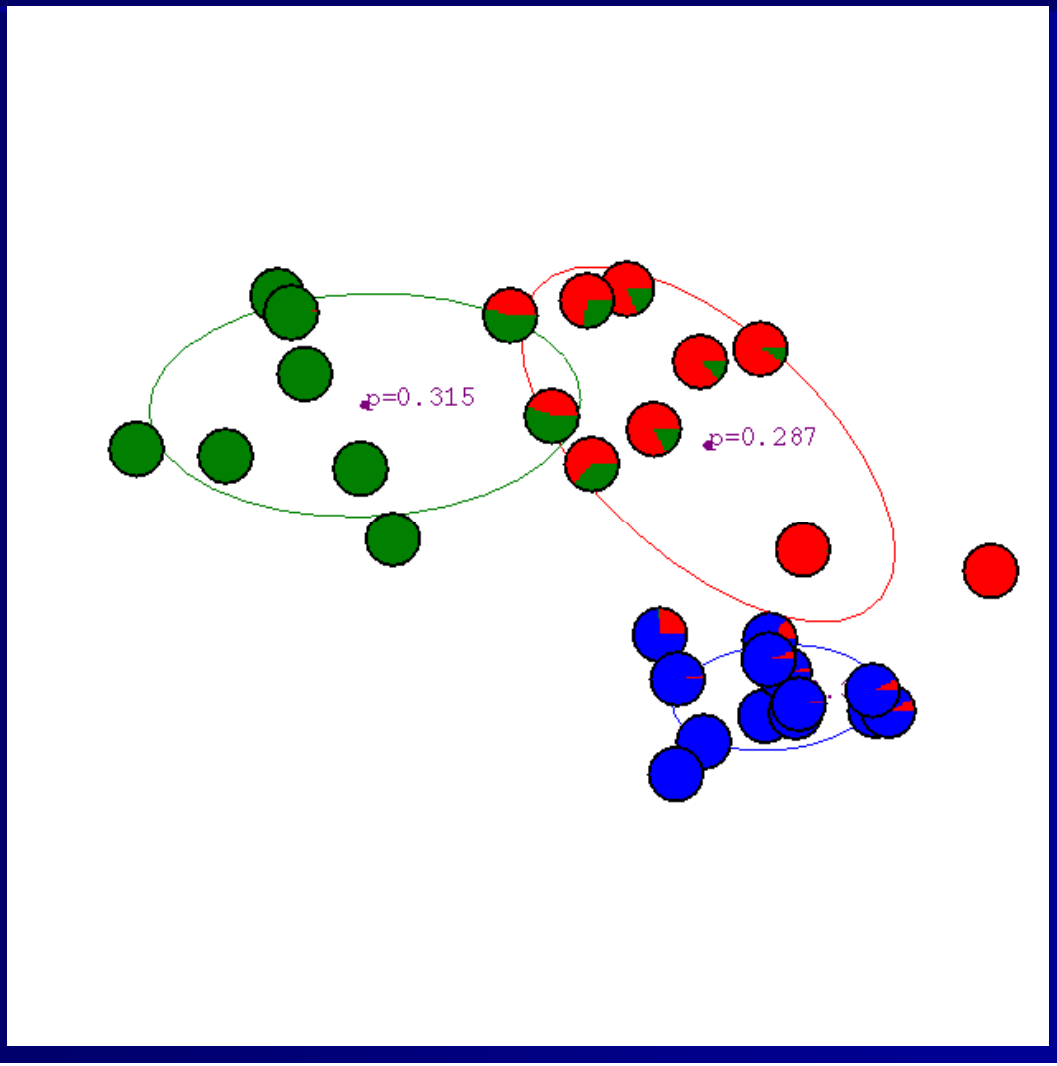
After ۴th Iteration



After ٥th Iteration



After ۶th Iteration



After ٢٠th Iteration