Lecture 11. Gaussian mixture model

Xin Chen

These slides are based on slides from Mahdi Roozbahani

Outline

- Overview
- Gaussian mixture model
- The expectation-maximization algorithm

Recap

Conditional Probabilities:

$$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$

Bayes rule:

$$p(A|B) = \frac{p(A, B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$$

$$p(A = 1) = \sum_{i=1}^{K} p(A = 1, B_i) = \sum_{i=1}^{K} p(A = 1|B_i)p(B_i)$$

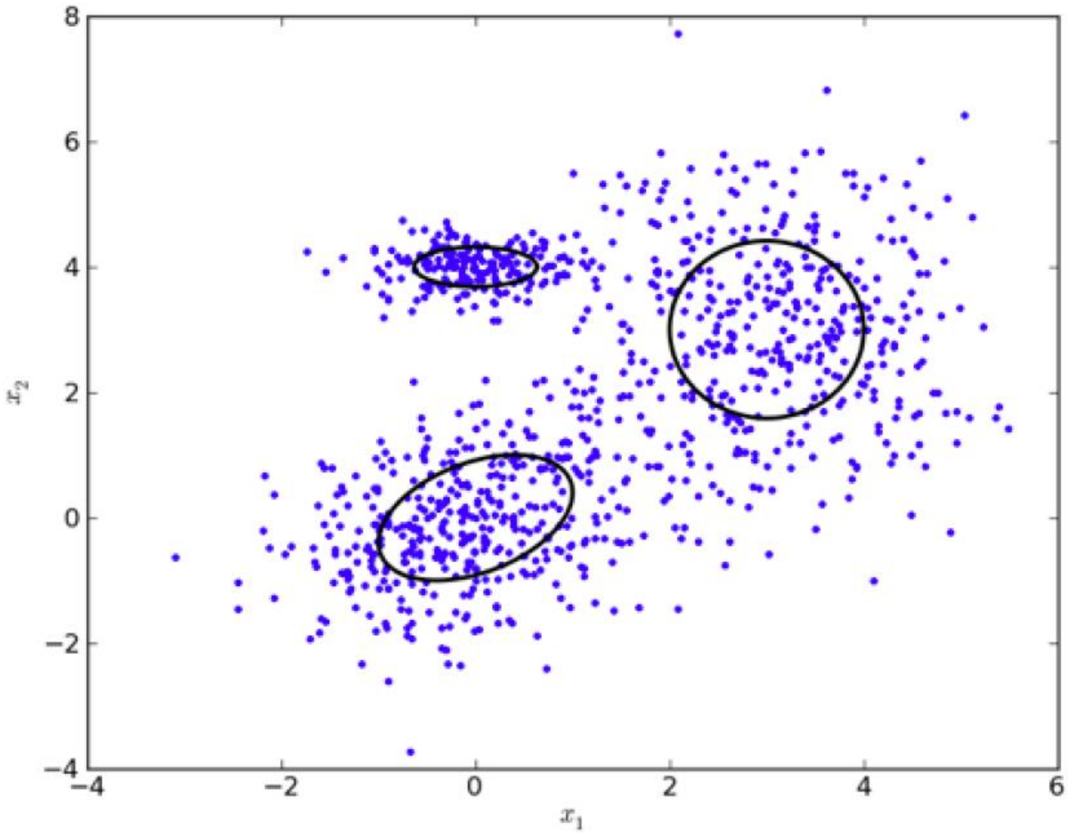
A simple example

	Tomorrow=Rainy	Tomorrow=Cold
Today=Rainy	4/9	2/9
Today=Cold	2/9	1/9
P(Tomorrow)	[4/9 + 2/9] = 2/3	[2/9 + 1/9] = 1/3

$$P(Tomorrow = Rainy) =$$

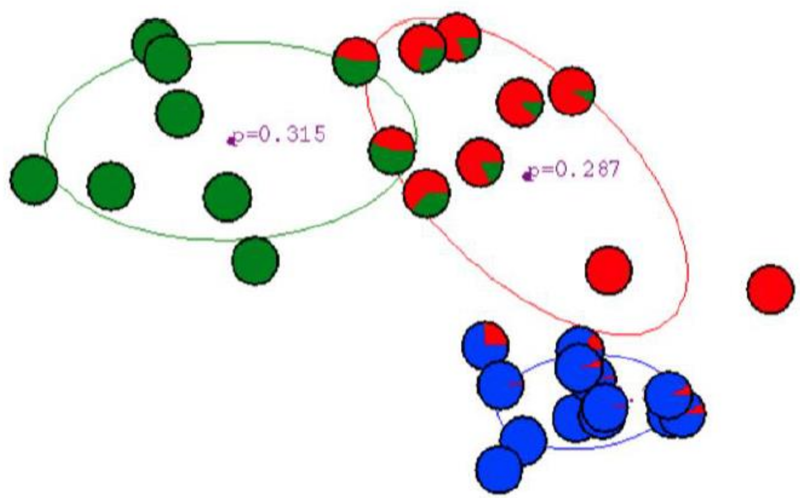
Hard clustering can be difficult

- Hard clustering: K-means, hierarchical clustering, DMSCAN



Toward soft clustering

- K-means
  - Hard assignment: each data point belongs to only one cluster
- Mixture modeling
  - Soft assignment: probability that a data point belongs to a cluster



Comparison

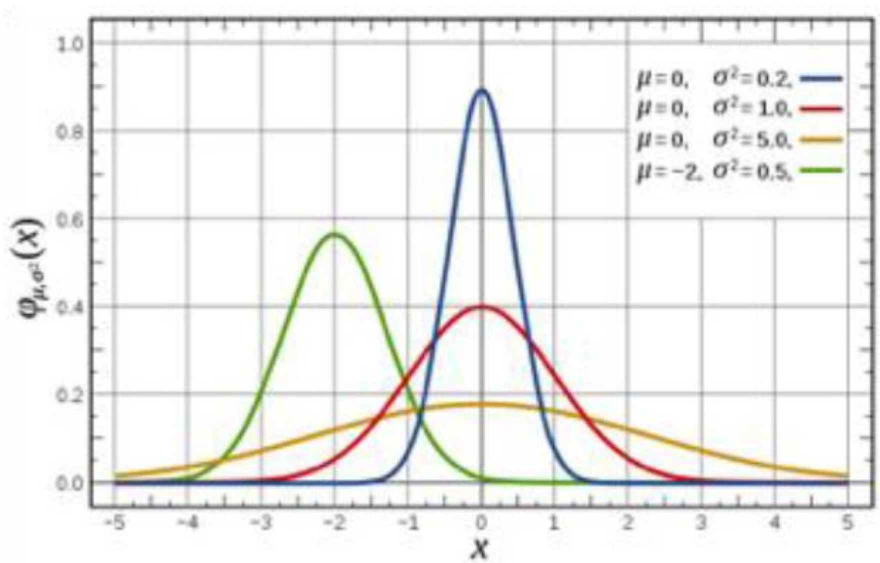
- Hard clustering
  - It is an assignment of x_n to a single cluster. It selects a mode of the conditional distribution $argmax\ p(z_n = k|x_n)$
- Soft clustering
  - It assigns a probability π_{nk} for data point x_n to each cluster k.

Outline

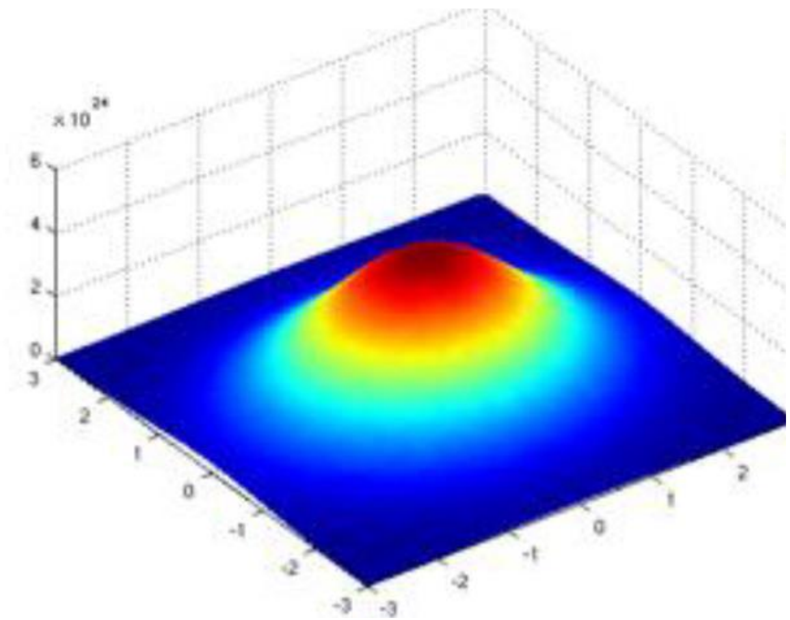
- Overview
- Gaussian mixture model
- The expectation-maximization algorithm

Gaussian Distribution

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



1-D Gaussian



2-D Gaussian

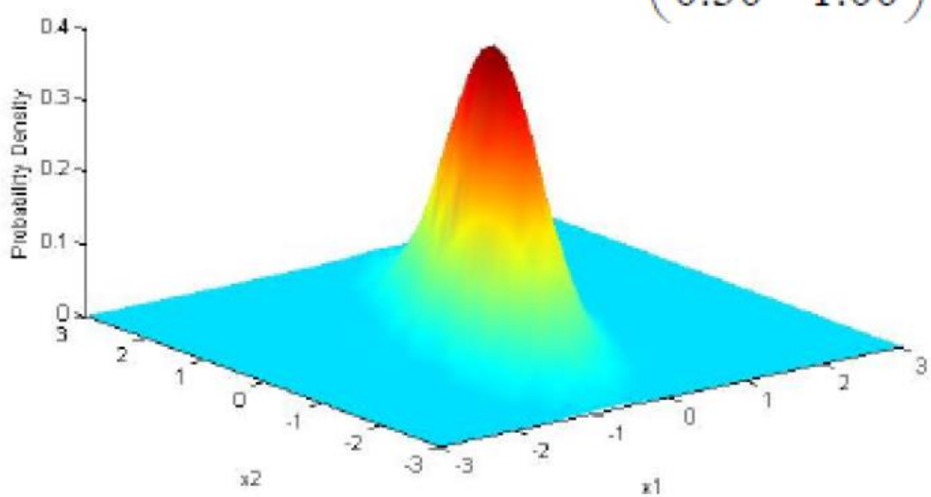
What is Gaussian?

- For d dimensions, the Gaussian distribution of a vector $x = (x^1, x^2, \ldots, x^n)^T$ is defined by

  $$N(x|\mu, \Sigma) = \frac{1}{2\pi^{d/2}\sqrt{|\Sigma|}} \exp(\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)),$$

  where μ is the mean, Σ is the covariance matrix of the Gaussian.

Example: $\mu = (0,0)^T$ $\Sigma = \begin{pmatrix} 0.25 & 0.30 \\ 0.30 & 1.00 \end{pmatrix}$



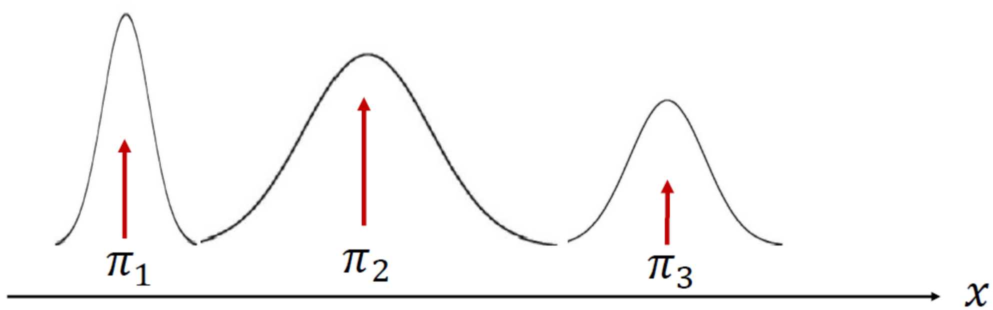
Gaussian Mixture Models (GMM)

- Formally a mixture model is the weighted sum of a number of probability density function (pdf), where the weights are determined by a distribution, π

$$p(x) = \pi_0 f_0(x) + \pi_1 f_1(x) + \cdots + \pi_k f_k(x),$$

where $\sum_{i=0}^{k} \pi_i = 1$

$$p(x) = \sum_{i=0}^{k} \pi_i f_i(x)$$



π_i is the unknown probability of selecting component i

Some notes

- Is summation of a bunch of Gaussians a Gaussian itself?
  - $p(x)$ is a Probability density function or it is also called a marginal distribution function
  - $p(x)$=the density of selecting a data point from the pdf which is created from a mixture model. Also, we know that the area under a density function is equal to 1.

Mixture models are generative

- Generative simply means dealing with joint probability $p(x, z)$
- Let's say $f(.)$ is a Gaussian distribution

$$p(x) = \sum_{k=0}^{K} \pi_k f_k(x)$$

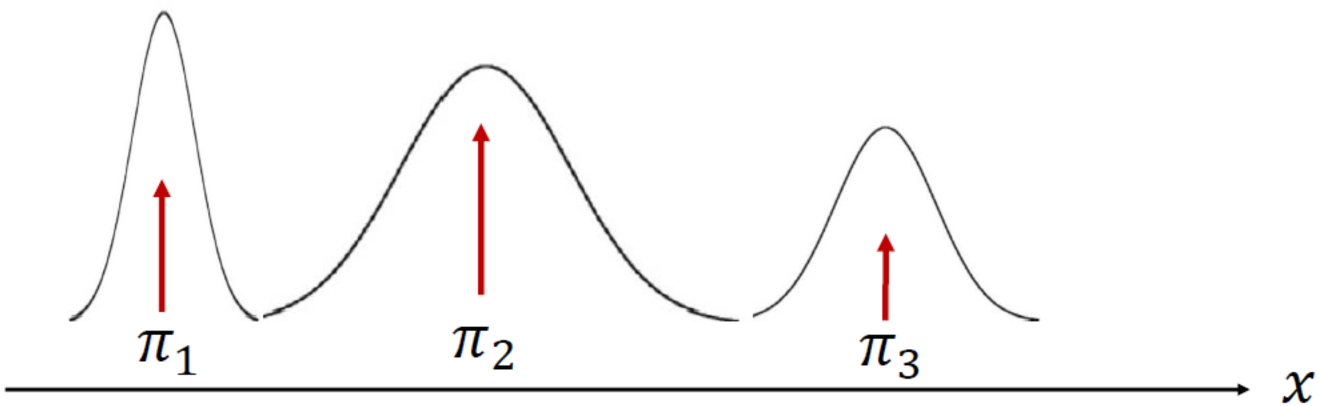
$$p(x) = \pi_0 N(x|u_0, \sigma_0) + \pi_1 N(x|u_1, \sigma_1) + \cdots + \pi_k N(x|u_k, \sigma_k)$$

$$\boxed{\begin{aligned} p(x) &= \sum_{k=0}^{K} \pi_k N(x|\, u_k, \sigma_k) \\ p(x) &= \sum_{k=0}^{K} p(z_k) p(x|z_k) \end{aligned}}$$

z_k is component k

$$p(x) = \sum_{k=0}^{K} p(x, z_k)$$

What is soft assignment?



- What is the probability of a data point x in each component?
- How many components we have here?
- How many probability?
- What is the sum of the 3 probabilities for each data point?

How to calculate the probability of data points in the first component?

$$p(x) = \pi_0 N(x|u_0, \sigma_0) + \pi_1 N(x|u_1, \sigma_1) + \pi_2 N(x|u_2, \sigma_2)$$

Let's calculate the responsibility of the first component among the rest

Let's call that τ_0

$$\tau_0 = \frac{N(X|\mu_0, \sigma_0)\pi_0}{N(X|\mu_0, \sigma_0)\pi_0 + N(X|\mu_1, \sigma_1)\pi_1 + N(X|\mu_2, \sigma_2)\pi_2}$$

$$\tau_0 = \frac{p(x|z_0)p(z_0)}{p(x|z_0)p(z_0) + p(x|z_1)p(z_1) + p(x|z_1)p(z_1)}$$

$$\tau_0 = \frac{p(x, z_0)}{\sum_{k=0}^{k=2} p(x, z_k)} = \frac{p(x, z_0)}{p(x)} = p(z_0|x)$$

Inferring cluster membership

- We have representations of the joint $p(x, z_{nk}|\theta)$ and the marginal, $p(x|\theta)$
- The conditional of $p(x, z_{nk}|\theta)$ can be derived using Bayes rule.
  – The responsibility that a mixture component takes for explaining an observation x.

z_{nk} represents the latent component indicator or latent cluster k for data point x_n

$$p(x|z_{nk}) = N(x|u_k, \sigma_k)$$

$$\tau(z_{nk}) = p(z_{nk}|x) = \frac{p(z_{nk})p(x|z_{nk})}{\sum_{j=1}^{K} p\big(z_{ij}\big)p(x|z_{ij})} = \frac{\pi_k N(x|u_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|u_j, \sigma_j)}$$

Mixtures of Gaussians

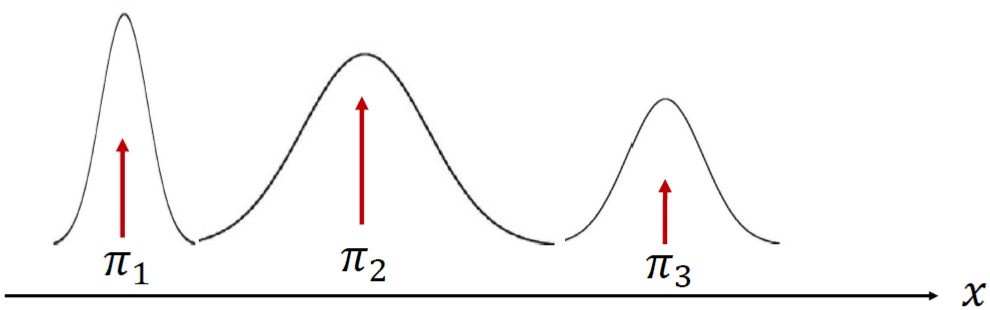
- What is the probability of picking a mixture component (Gaussian model)? $p(z_k) = \pi_k$

- What is the probability of picking data from that specific mixture component? $p(x|z_k)$

Note z_k is a **latent variable**. We only observe x, but z_k is hidden

$$p(x, z_k) = p(x|z_k)p(z_k) \qquad p(x, z_k) = \pi_k N(x|u_k, \sigma_k)$$

Generative model, because of joint distribution

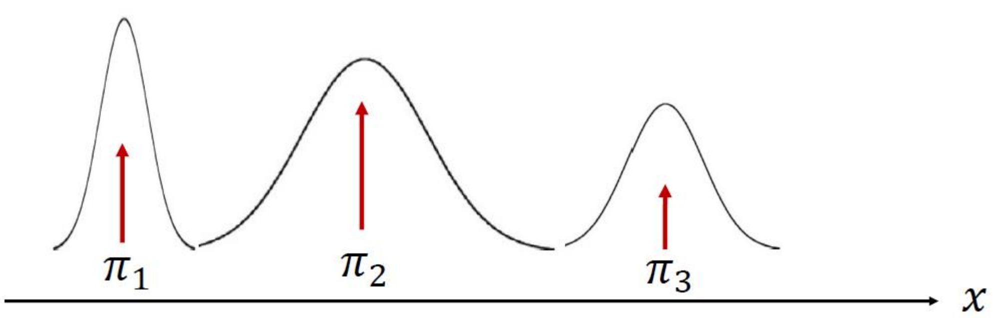


What are GMM parameters?

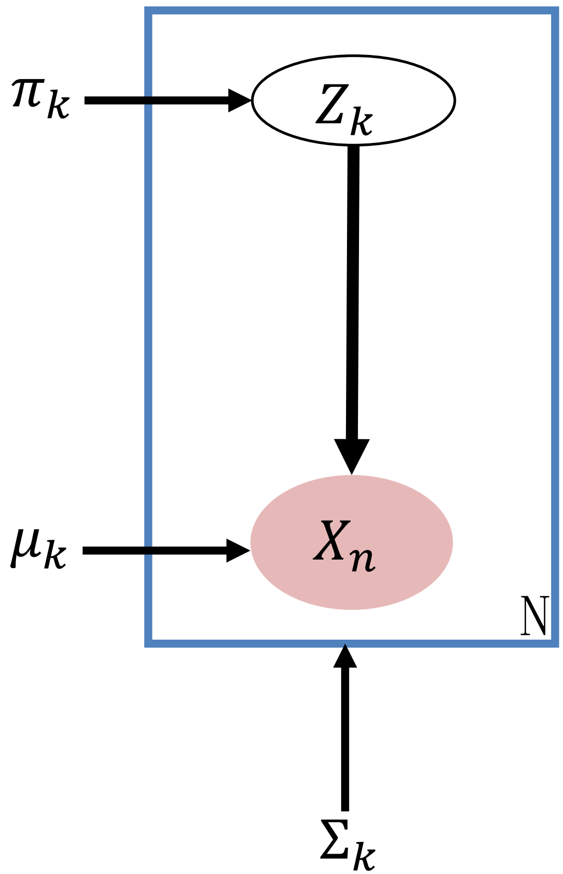
Mean u_k, Variance: σ_k, Proportion: π_k

$$p(x, z_k) = p(x|z_k)p(z_k) = \pi_k N(x|u_k, \sigma_k)$$

- $p(z_k|\theta) = \pi_k$ select a mixture component with probability π_k
- $p(x|z_k) = N(x|u_k, \sigma_k)$ sample from the component's Gaussian.



GMM with graphical model concept

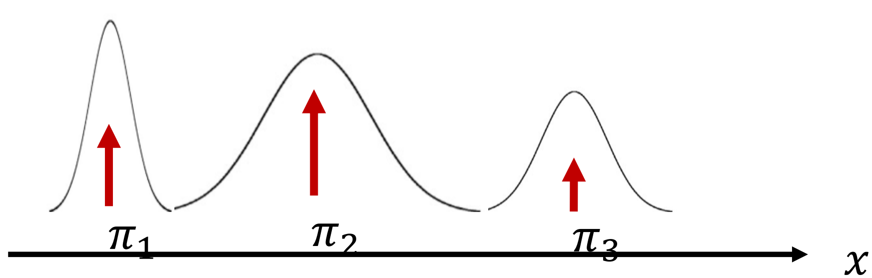


$$p(z_{nk}|\pi_k) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

Z_k is the latent variable
1-of-K representation

Given $z, \pi, \mu,$ and Σ, what is the probability of x in component k

$$p(x|z_{nk}, \overbrace{\pi, \mu, \Sigma}^{\theta}) = \prod_{k=1}^{K} \left(N(x|\mu_k, \Sigma_k)\right)^{z_{nk}}$$



Why having “Latent variable”

- A variable can be unobserved (latent) because:
  - it is an imaginary quantity meant to provide some simplified and abstractive view of the data generation process.
    - e.g., speech recognition models, mixture models (soft clustering)…
  - it is a real-world object and/or phenomena, but difficult or impossible to measure
    - e.g., the temperature of a star, causes of a disease, evolutionary ancestors …
  - it is a real-world object and/or phenomena, but sometimes wasn’t measured, because of faulty sensors, etc.

- Discrete latent variables can be used to partition/cluster data into sub-groups.
- Continuous latent variables (factors) can be used for dimensionality reduction (factor analysis, etc).

Latent variable representation

$$\mathrm{p}(\mathrm{x}|\theta) = \sum\nolimits_k p(x, z_{nk}|\theta) = \sum\nolimits_k p(z_{nk}|\theta) p(x|z_{nk}, \theta) = \sum_{k=0}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

$$p(z_{nk}|\theta) = \prod\nolimits_{k=1}^{K} \pi_k^{z_{nk}}$$

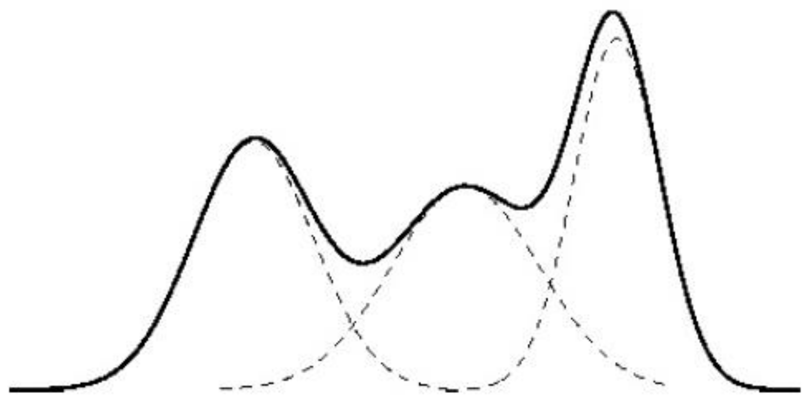
$$p(x|z_{nk}, \theta) = \prod\nolimits_{k=1}^{K} \left(N(x|\mu_k, \Sigma_k)\right)^{z_{nk}}$$

Why having the latent variable?

The distribution that we can model using a mixture of Gaussian components is much more expressive than what we could have modeled using a single component.

Define latent variable

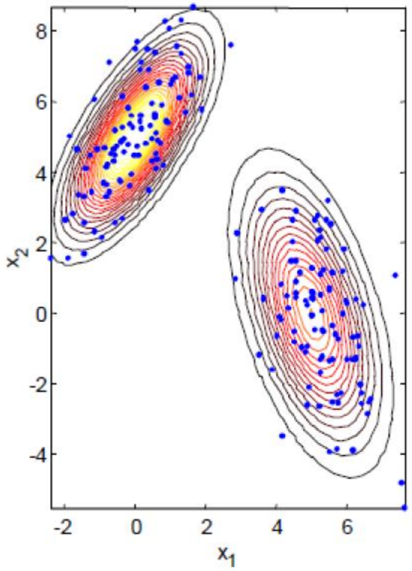
For a point x_i, let the cluster to which that point belongs be labeled z_i. z_i is a latent variable, which is unobserved.

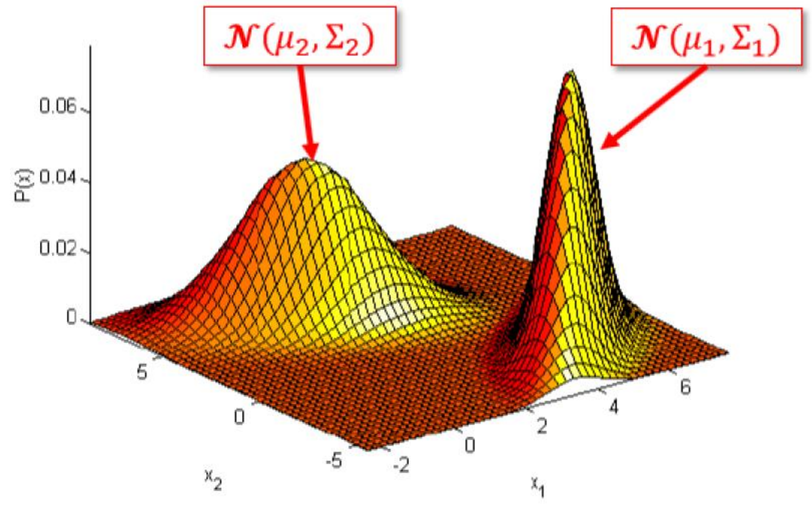


The density of a univariate Gaussian Mixture Model with three Gaussian mixture components, each with their own mean and variance terms (K = 3, d = 1). [Source: http://prateekvjoshi.com]

Multimodal distribution

- What if we know the data consists of a few Gaussians.
- What if we want to fit parametric models?





Gaussian mixture model

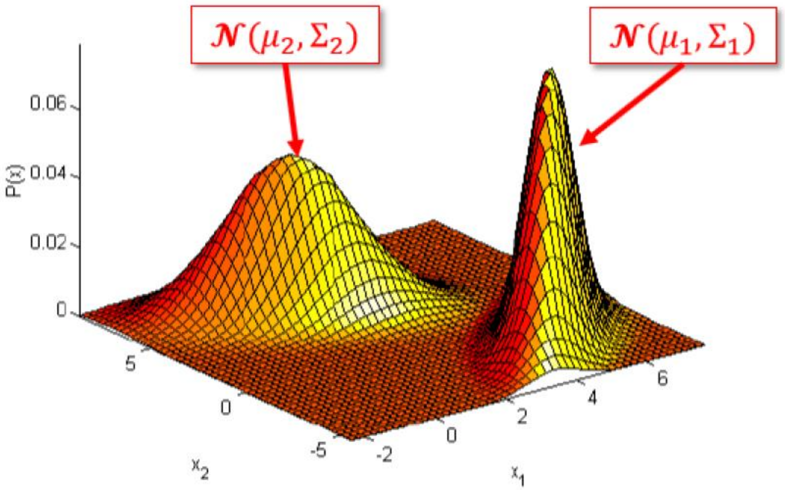
- A density model $p(x)$ may be multi-modal: model it as a mixture of uni-modal distribution (e.g. Gaussians).

- Consider a mixture of K Gaussians

$$p(x) = \sum^{K} \pi_k N(x|\, u_k, \sigma_k)$$

mixing proportion

mixture Component



Learn mean u_k, Variance: σ_k, Proportion: π_k

Learning GMM parameters

- Maximum likelihood estimation

$$argmax\ p(x|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k N(x_i|u_k, \sigma_k)$$

$$\log\ (p(x|\theta) = \sum_{i=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k N(x_i|u_k, \sigma_k)\ \}$$

The fundamental difficulty is that the parameters are coupled.

$$\log(p(x|\theta) = \sum_{i=1}^{N} \log\{\sum_{k=1}^{K} p(x_i|z_k)|p(z_k)\ \}$$

$\boldsymbol{z_{nk}}$ **Latent variable**

Now we assume that $\tau(z_{nk}) = p(z_{nk}|x)$ is known.

Estimate the mean in GMM

$$\ln p(x|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n|\mu_k,\Sigma_k) \right\}$$

$$\begin{aligned}
\frac{\partial \ln p(x|\pi,\mu,\Sigma)}{\partial \mu_k} &= \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)} \Sigma_k^{-1}(x_k - \mu_k) = 0 \\
&= \sum_{n=1}^{N} \tau(z_{nk}) \Sigma_k^{-1}(x_k - \mu_k) = 0
\end{aligned}$$

$$\boxed{\mu_k = \frac{\sum_{n=1}^{N} \tau(z_{nk}) x_n}{\sum_{n=1}^{N} \tau(z_{nk})}}$$

Estimate the variance in GMM

$$\ln p(x|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k) \right\}$$

$$\boxed{\Sigma_k \quad = \quad \frac{1}{\sum_{n=1}^{N} \tau(z_{nk})} \sum_{n=1}^{N} \tau(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T}$$

Estimate the mixing term in GMM

$$\ln p(x|\pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$$0 = \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j, \Sigma_j)} + \lambda$$

$$\boxed{\pi_k = \frac{\sum_{n=1}^{N} \tau(z_{nk})}{N}}$$

Parameter results

Means: $$\mu_k = \frac{\sum_{n=1}^{N} \tau(z_{nk}) x_n}{\sum_{n=1}^{N} \tau(z_{nk})}$$

Variance: $$\Sigma_k = \frac{1}{\sum_{n=1}^{N} \tau(z_{nk})} \sum_{n=1}^{N} \tau(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T$$

Mixing term: $$\pi_k = \frac{\sum_{n=1}^{N} \tau(z_{nk})}{N}$$

Note that all these based on the assumption that $\tau(z_{nk})$ is known, which is our guess. How to guess?

Outline

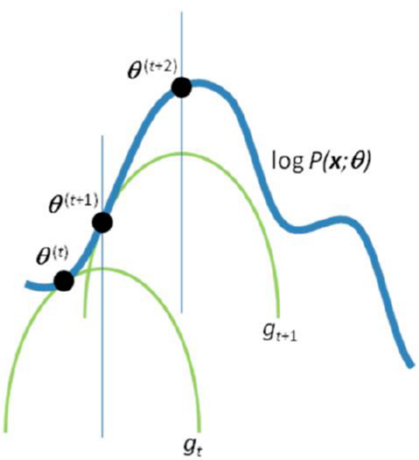
- Overview
- Gaussian mixture model
- The expectation-maximization algorithm

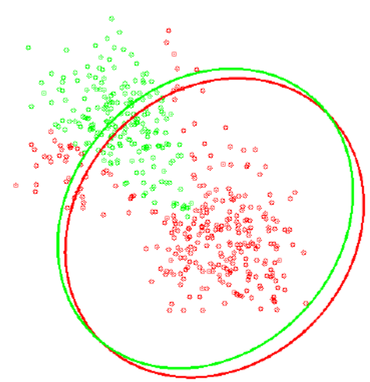
Estimating GMM parameters with Expectation-Maximization (EM)

- EM is a general algorithm to deal with hidden variable.

- Two steps:
  - E-step: Fill in hidden values using inference
  - M-step: Apply standard MLE method to estimate parameters

- EM always converges to a local minimum of the likelihood.





E-step for GMM

We assume that $\theta^t = (\pi_k, u_k, \sigma_k)$ are known, and then take the expectation of the latent variable with the current values of our parameters.

Posterior expectation $E(z_{nk}) \propto \pi_k N(x_n|u_k, \sigma_k)$, posterior probability of data x_n belonging to cluster k

$$E(z_{nk}) = \frac{\pi_k N(x_n|u_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j N\left(x_n \middle| u_j, \sigma_j\right)}$$

$$\boxed{\tau(z_{nk}) \quad = \quad \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j, \Sigma_j)}}$$

M-step for GMM

$$Q(\theta^t) = \sum_{i=1}^{n}\sum_{k=1}^{K} E[z_{nk}]log\pi_k + E[z_{nk}]logN(x_n|u_k,\sigma_k)$$

Based on the assumption that $\theta^t = (\pi_k, u_k, \sigma_k)$, we need to update θ^t with $\theta^{t+1} = argmax(Q(\theta^t))$.

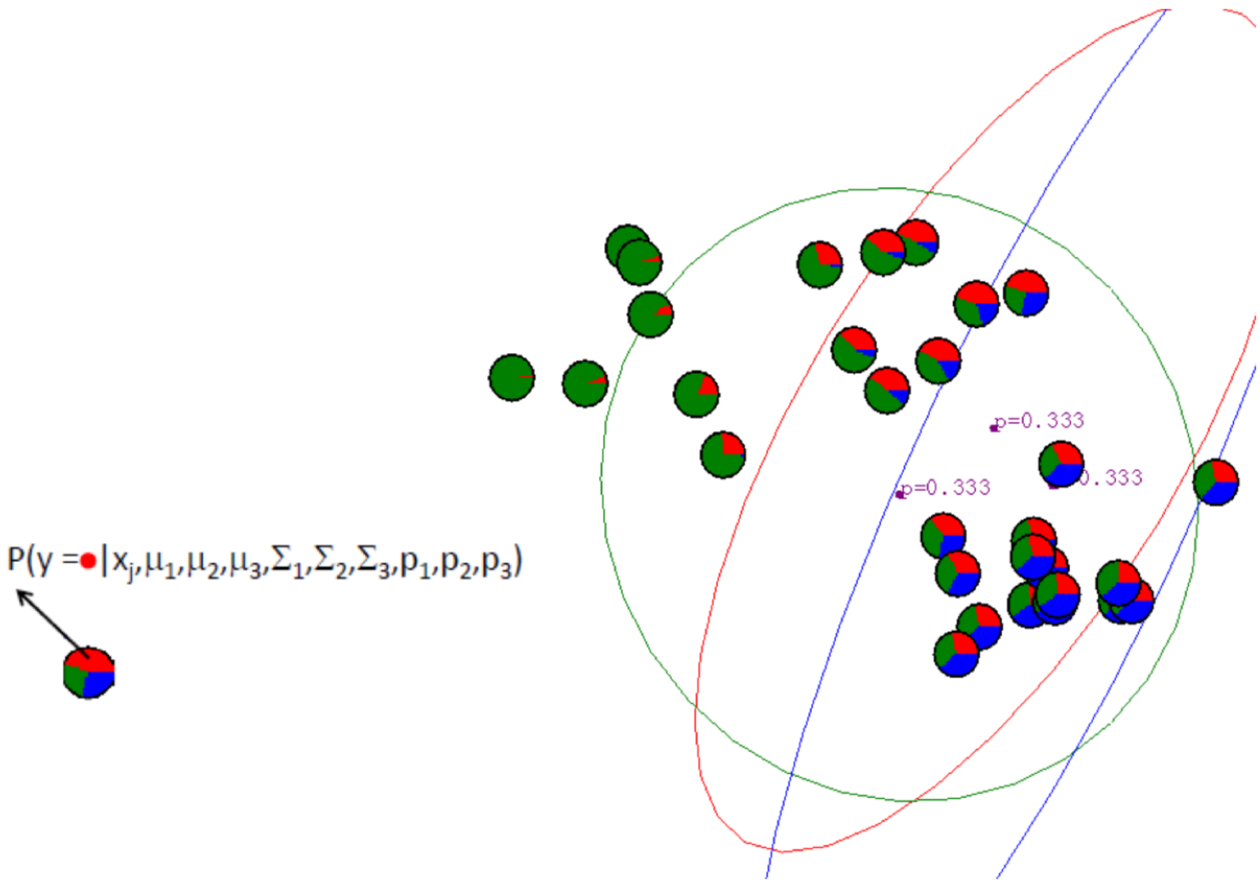
- ${u_k}^{t+1} = \frac{\sum_{n=1}^{N} \tau(z_{nk})^t x_n}{\sum_{n=1}^{N} \tau(z_{nk})^t}$
- ${\Sigma_k}^{t+1} = \frac{1}{\sum_{n=1}^{N} \tau(z_{nk})^t} \sum_{n=1}^{N} \tau(z_{nk})^t \, (x_n - {u_k}^t)(x_n - {u_k}^t)^T$
- ${\pi_k}^{t+1} = \frac{\sum_{n=1}^{N} \tau(z_{nk})^t}{N}$

Expectation-Maximization for GMMs

- Initialize π_k, u_k, σ_k arbitrarily.

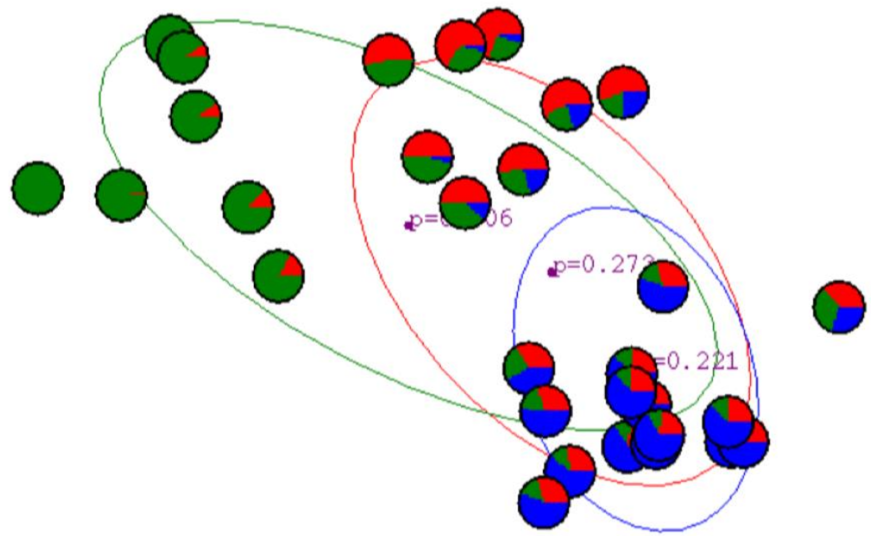
- Alternate until convergence
  - (E-step) Expectation step: compute soft class membership, with the current parameters: $\tau_{nk} = \tau(z_{nk}) = p(z_{nk}|x, \pi_k, (u_k,\ \sigma_k))$
  - (M-step) Maximization step: Update parameters by plugging in τ_{nk} (our guess)

EM for GMM example



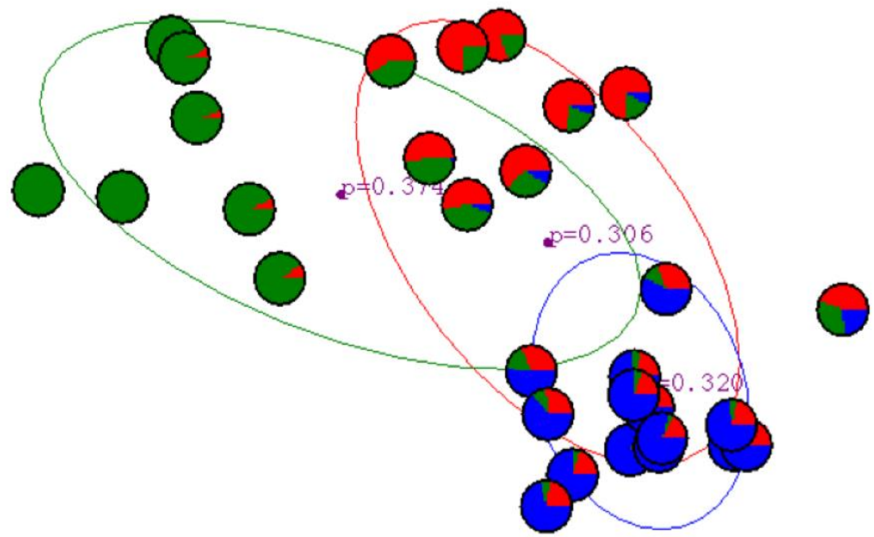
EM for GMM example

After 1^{st} iteration



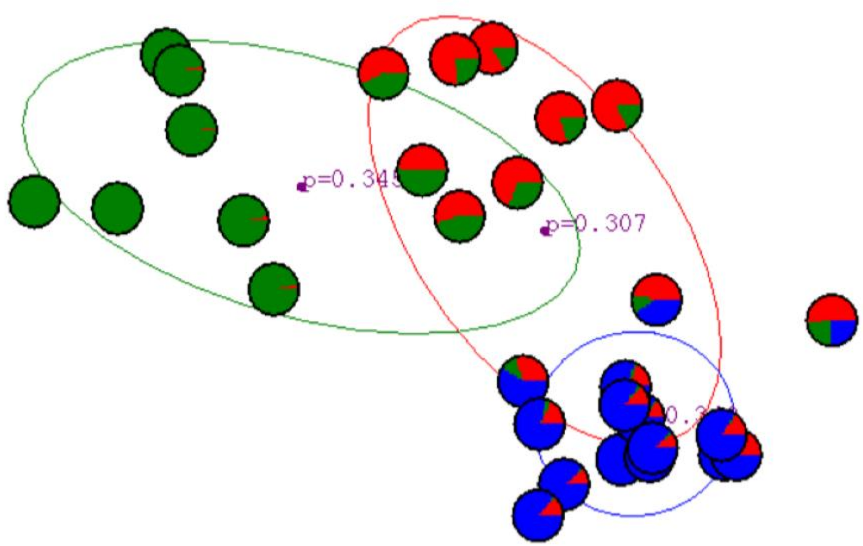
EM for GMM example

After 2nd iteration



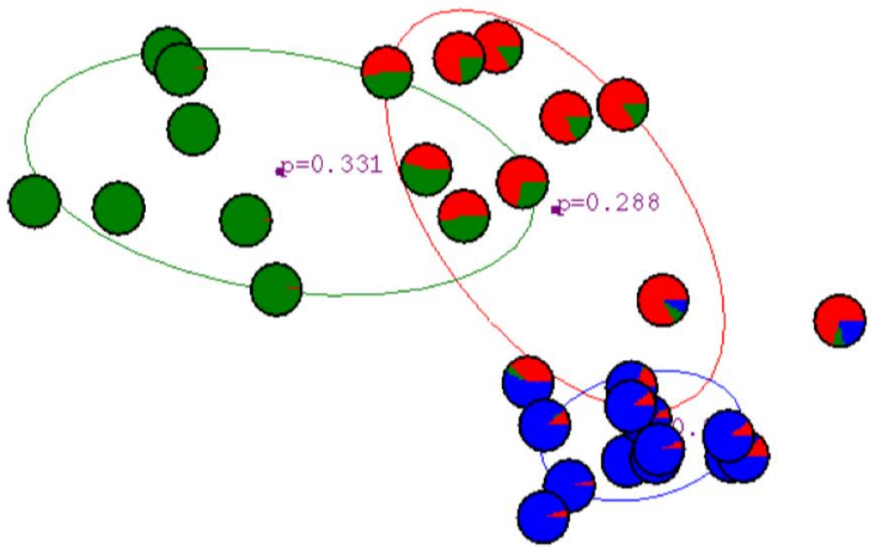
EM for GMM example

After 3rd iteration



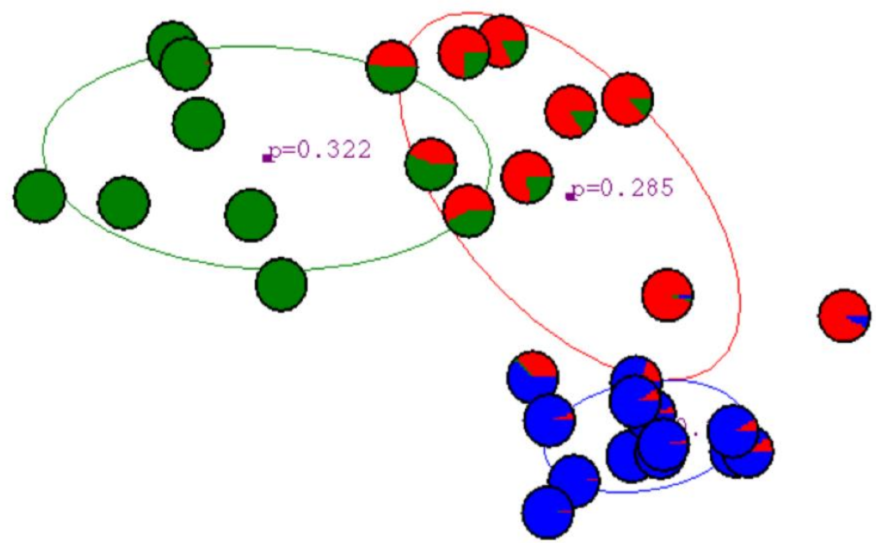
EM for GMM example

After 4^{th} iteration



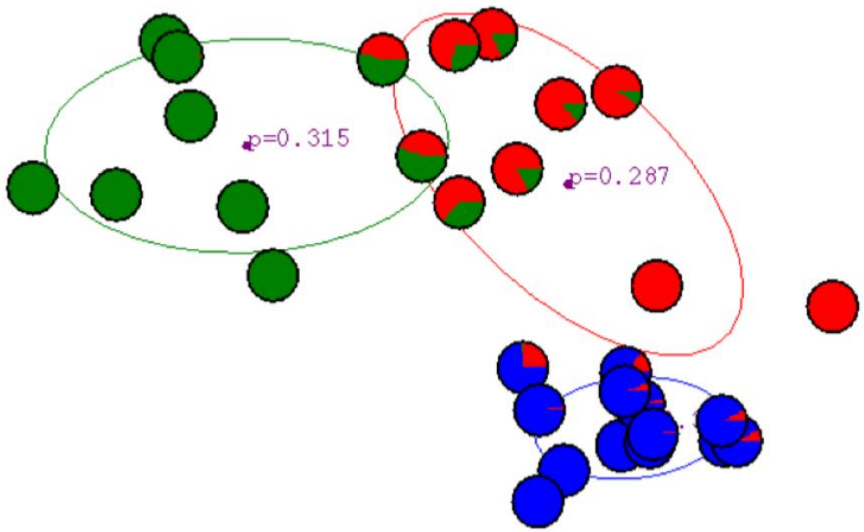
EM for GMM example

After 5th iteration



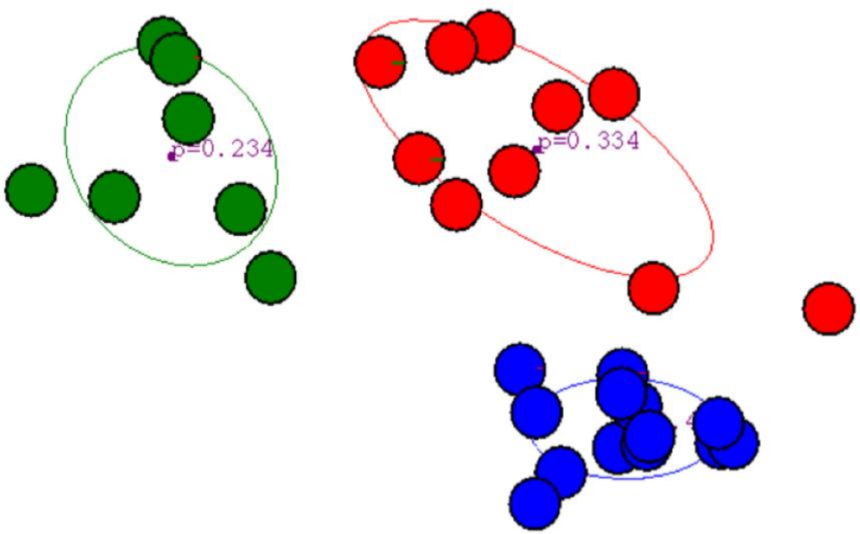
EM for GMM example

After 6^{th} iteration



EM for GMM example

After 20th iteration



General form of EM

- Given a joint distribution over observed variables and latent variables: $p(x, z|\theta)$
- Want to maximize: $p(x|\theta)$

1. Initialize parameters: θ^{old}
2. E-step, evaluate $p\left(z\middle|x, \theta^{old}\right)$
3. M-step, re-estimate parameters (based on expectation of complete-data log likelihood):

$$\theta^{new} = argmax_{\theta} \sum p(z|x, \theta^{old}) \ln p(x, z|\theta)$$

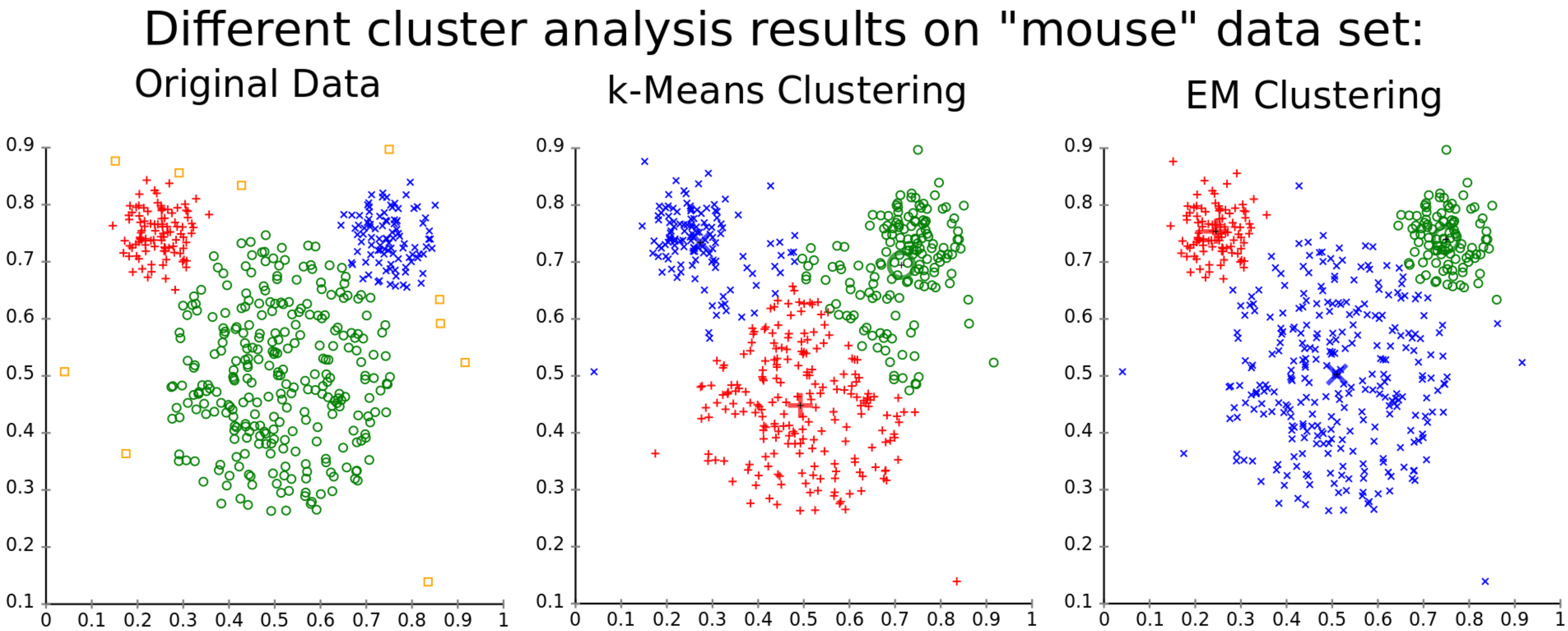
Comparison between GMM and K-Means

- Soft clustering and hard clustering
  - K-means assigns data point to a single cluster, while GMM assigns probability of observations belonging to each cluster.

- GMM assumes Gaussian model with joint probability, while K-means has no underlying probability model.

- Relationship between GMM and K-Means
  - K-means, unlike GMM, learns equal-sized cluster, where $\pi_k = \frac{1}{K}$
  - In GMM, we set $\pi_k = \frac{1}{K}$ and set the largest probability to 1 and the rest to 0. Then GMM is equivalent to K-means.

An example of comparing K-means with EM



https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm