Ch 9. Mixture Models and EM

Pattern Recognition and Machine Learning,
C. M. Bishop, 2006.

Maria Helena Mejia S
Cs645 Statistical modeling & inference
University of Arizona

Contents

- K-means Clustering
- Mixtures of Gaussians
- Maximum likelihood
- EM for Gaussian mixtures
- Summary

K-means Clustering

- Problem of identifying groups, or clusters, of data points in a multidimensional space

ALGORITHM

- Assume the data Euclidean space.
- Assume we want k classes.
- Assume we start with randomly located cluster centers

The algorithm alternates between two steps:

1) Assignment step: Assign each datapoint to the closest cluster.

2) Refitting step: Move each cluster center to the center of gravity of the data assigned to it.

K-means Clustering

- Goal: an assignment of data points to clusters such that the sum of the squares of the distances to each data point to its closest vector (the center of the cluster) is a minimum

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} || \boldsymbol{x}_n - \mu_k ||^2$$

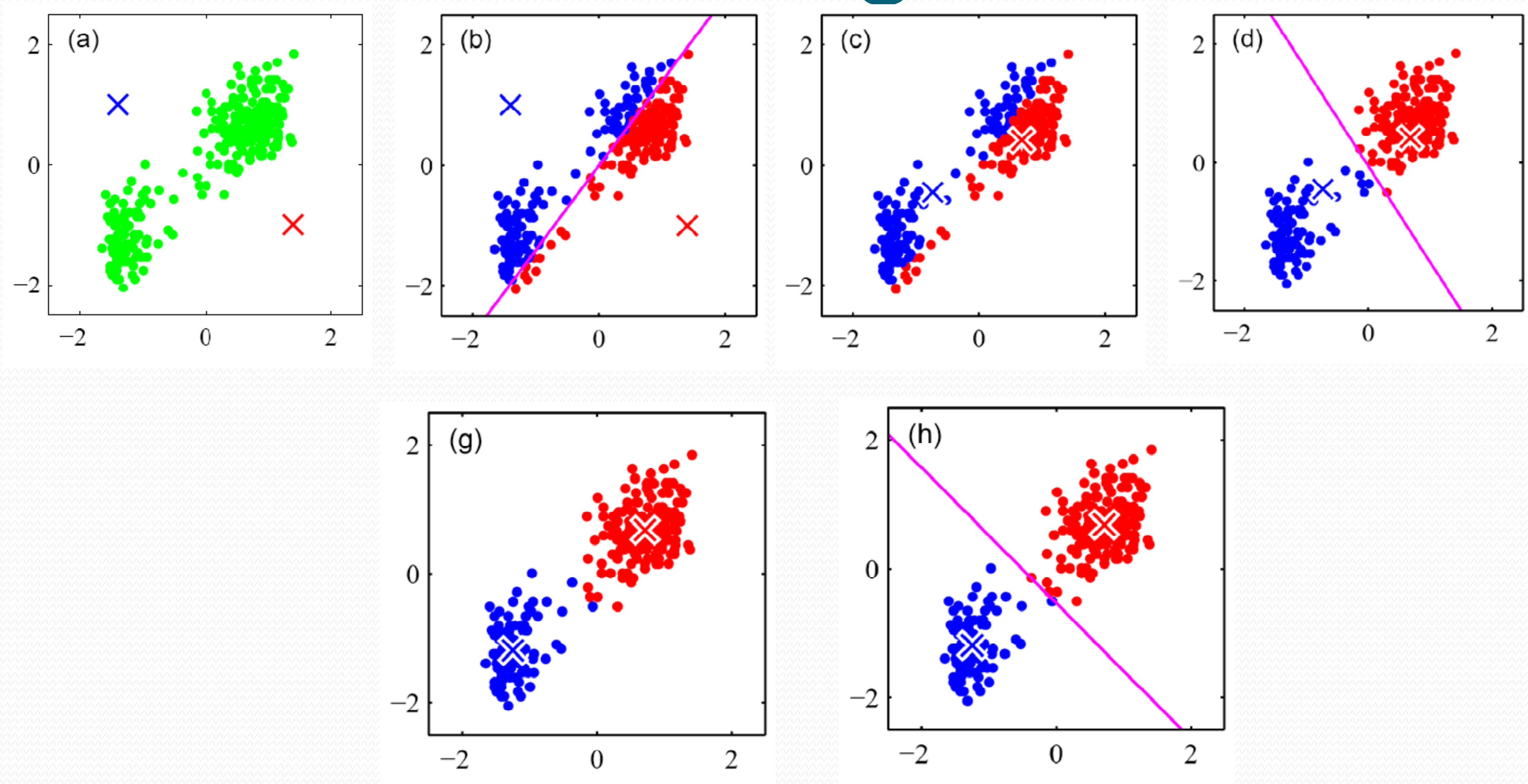
- Two-stage optimization (Repeat until converge)
- 1) Minimizing J with respect to the r_{nk}, keeping the μ_k fixed

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg min_j || \boldsymbol{x}_n - \mu_j ||^2 \\ 0 & otherwise \end{cases}$$

- 2) Minimizing J with respect to the μ_k, keeping r_{nk} fixed

$$\mu_k = \frac{\sum_n r_{nk} \boldsymbol{x}_n}{\sum_n r_{nk}}$$

K-means Clustering



Mixture Model

- We can model arbitrary distributions with density mixtures.

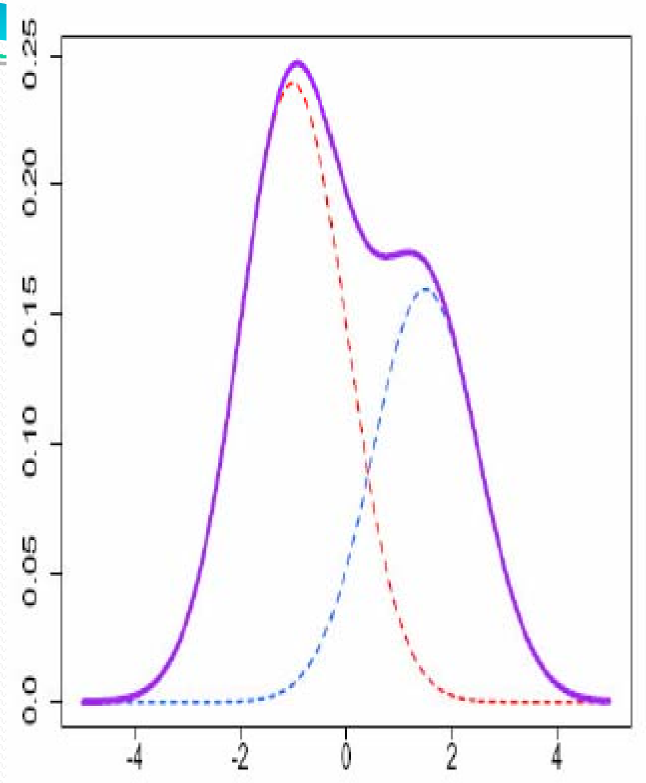
- A formalism for modeling a probability (density) function as a sum of parameterized functions.

$$f(x) = \pi_1 f_1(x) + \dots\dots\dots + \pi_K f_K(x) \qquad \pi_1 + \dots + \pi_K = 1 \quad 0 <= \pi_K <= 1$$

Mixture model with k components where each component is a probability function (Poisson, Binomial) or probability density function(Normal, Exponential, Gamma etc.)

- Probability mixture modeling as a missing data problem

Mixtures of Gaussians



- Gaussian mixture distribution can be written as a linear superposition of Gaussian

$$p(x) = \sum_{i=1}^{K} \pi_i \frac{\exp(-(x-\mu_i)^2 / 2\sigma_i^2)}{\sigma_i \sqrt{2\pi}}$$

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{x} | \mu_k, \Sigma_k)$$

Mixtures of Gaussians

- Formulation of the Gaussian mixture involving an explicit latent variable
  - Latent=hidden
  - Mixing coefficient is the latent variable in GMM.
  - In a generative model we can Interpret the mixing coefficients as prior probabilities

$$p(\boldsymbol{x}) = \sum_z p(z)\,p(\boldsymbol{x}|\,z) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\mu_k, \Sigma_k)$$

For every observed data point $\mathbf{x}_n$, there is a corresponding latent variable $\mathbf{z}_n$

Conditional probability of Z given X and model parameters. - Posterior probabilities (responsabilities)

$$\gamma(z_k) \equiv p(z_k = 1 \mid \boldsymbol{x}) = \frac{p(z_k = 1)\,p(\boldsymbol{x} \mid z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)\,p(\boldsymbol{x} \mid z_j = 1)}$$

$$= \frac{\pi_k N(\boldsymbol{x} \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x} \mid \mu_j, \Sigma_j)}$$

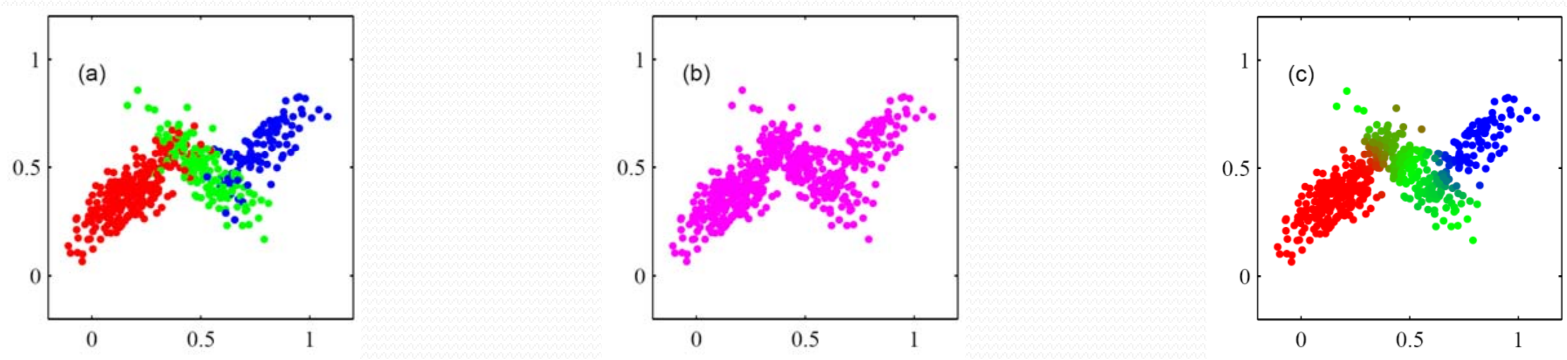
Mixtures of Gaussians

- The conditional probability of Z given X and model parameter, $\gamma(z_k)$ can also be viewed as the responsibility that component *k* takes for explaining' the observation **x. The probability that a point is generated by a particular Gaussian.**

- The values of the latent variables are unknown . However, for given parameter values it is possible compute the expected values of the latent variables.

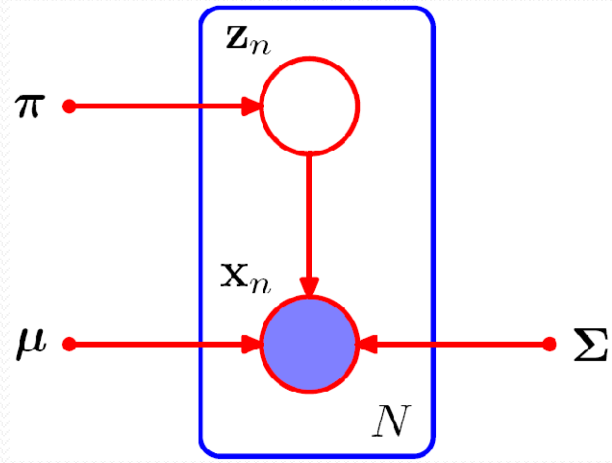
Mixtures of Gaussians

- Generating random samples distributed according to the Gaussian mixture model



Mixtures of Gaussians

- Graphical representation of a Gaussian mixture model for a set of N i.i.d. data points $\{x_n\}$, with corresponding latent (hidden) points $\{z_n\}$



- Goal: Learn model parameters from data
- A refinement of this goal is Maximum likelihood estimation

Maximum likelihood estimation

MLE= statistical method used for fitting a mathematical model to given data.

Observation of data -> Estimation of parameters

MLE for m-dimensional Gaussian

- Suppose you have x1, x2, … xR ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MLE: For which θ =(μ,Σ) is x1, x2, … xR most likely**?**

Maximum likelihood estimation

MLE for Gaussian Mixture Model

- MLE: For which θ =(μ,Σ) is x1, x2, … xN and unkown π1.. πK most likely?

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

Maximun Likelihoods Estimation

- MLE is an optimization problem

- Ln of sum is hard to solve it by an analytical way

Most common solutions

- Expectation Maximization (EM)

- MCMC sampling

EM for Gaussian mixtures

I. Assign some initial values for the means, covariances, and mixing coefficients

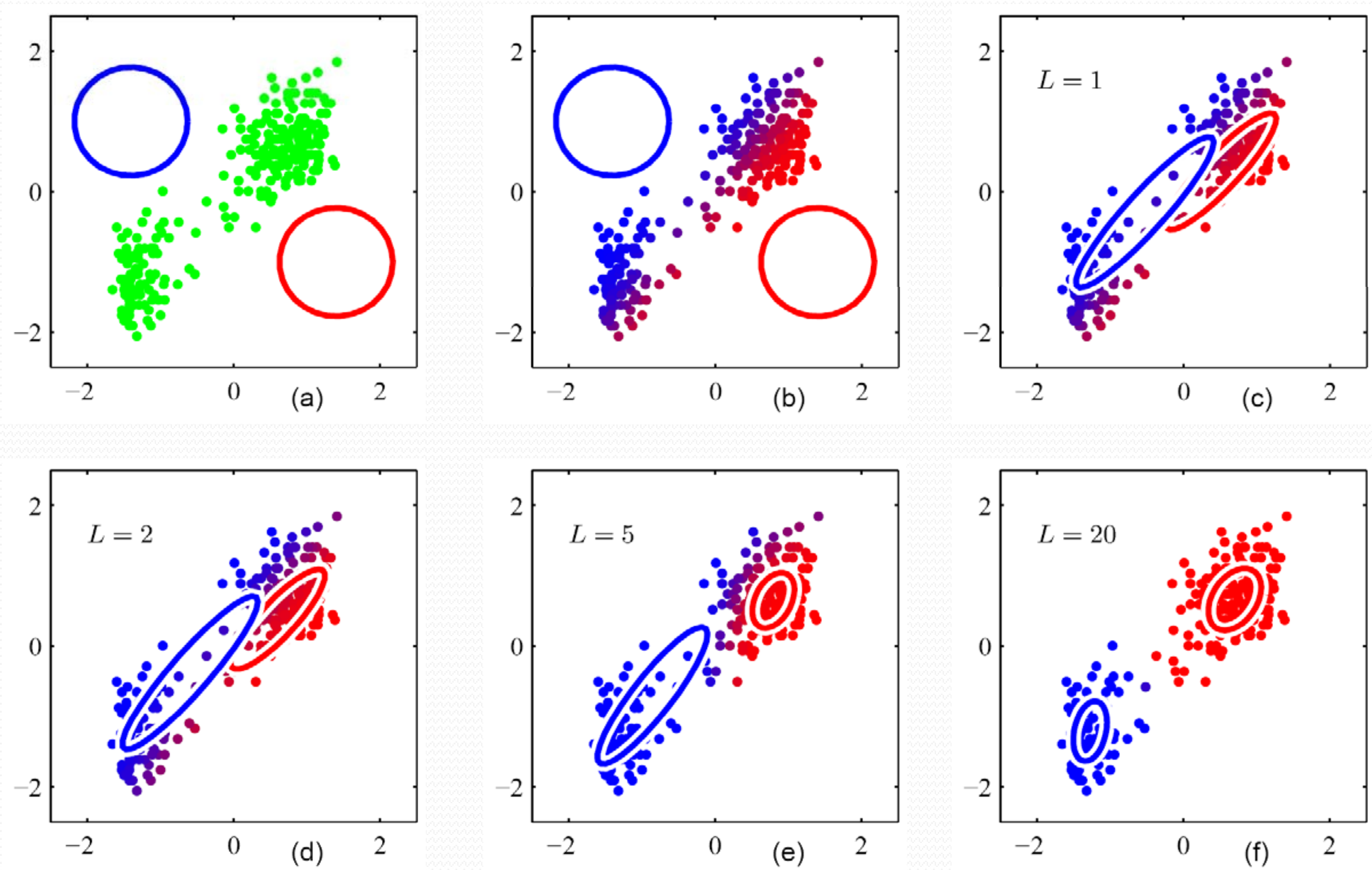
II. *Expectation* or E step

- Using the current value for the parameters to evaluate **responsibilities or the posterior probabilities** that each Gaussian generates each datapoint.

III. *Maximization* or M step

- Using the result of E step **to re-estimate the means, covariances, and mixing coefficients**
- Assuming that the data really was generated this way, change the parameters of each Gaussian to maximize the probability that it would generate the data it is currently responsible for.

EM for Gaussian mixtures



EM for Gaussian mixtures

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters.

Steps: Initialization, E, M, compare with likelyhood

Updating each Gaussian definitely improves the probability of generating the data if we generate it from the same Gaussians after the parameter updates.

General EM

- Goal:
  - Maximizing the log likelihood function

$$\ln p(\boldsymbol{X}|\theta) = \ln\{\Sigma_Z P(\boldsymbol{X},\boldsymbol{Z}|\theta)\}$$

  - Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\Theta)$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters Θ

1. Choose an initial setting for the parameters Θ^{old}
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X},\Theta^{old})$
3. **M step** Evaluate Θ^{new} given by
   $\Theta^{new} = \mathrm{argmax}_{\Theta} Q(\Theta, \Theta^{old})$
   $Q(\Theta, \Theta^{old}) = \Sigma_Z\, p(\mathbf{Z}|\mathbf{X}, \Theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\Theta)$
4. It the convergence criterion is not satisfied,
   then let $\Theta^{old} \leftarrow \Theta^{new}$

General application - Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

SUMMARY

- Mixture Models (MM) for modeling using different or the same kind of various probability function.
- Gaussian MM (GMM) used in two ways:
  - Model a distribution
  - CLUSTERING
- EM for finding MLE of parameters in probabilistic models where the model depends on latent variables.
- EM naturally applicable to training probabilistic models. It is useful in models where some data are missed.

THANKS !

ADDITIONAL REFERENCES

Gaussian Mixtures Models – MLE - EM

http://www.cs.cmu.edu/~awm/tutorials

http://www.csie.ntu.edu.tw/~mhyang/course/u0030/lectures/Bishop-ECCV-04-tutorial-B.pdf