

# Mixture Models and EM

**Mohsen Afsharchi**

# Introduction

- Mixture models provide:
  - Framework for building complex probability distributions
  - A method for clustering data
- Viewed statistically as follows:
  - Complex distribution expressed in terms of more tractable joint distribution of observed and latent variables
  - Distribution of observed variables alone is obtained by marginalization

# Plan of discussion

- Using  $K$ -means algorithm for finding clusters in a set of data points
- Latent variable view of mixture distributions
  - Assigning data points to specific components of mixture
- General technique for finding m.l. estimators in latent variable models
  - Expectation Maximization (EM) algorithm
- EM Algorithm
  - Gaussian mixture models motivates EM
  - Latent variable viewpoint
  - $K$ -means seen as non-probabilistic limit of EM applied to mixture of Gaussians
  - EM in generality

# *K*-means Clustering

- Given data set  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  in  $D$ -dimensional Euclidean space
- Partition into  $K$  clusters, which is given
- One of  $K$  coding
- Indicator variable  $r_{nk} \in \{0, 1\}$  where  $k = 1, \dots, K$ 
  - Describes which of  $K$  clusters data point  $\mathbf{x}_n$  is assigned to

# Distortion measure

- Sum of squared errors

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

- Goal is to find values for  $\{r_{nk}\}$  and the  $\{\mu_k\}$  so as to minimize  $J$ 
  - Can be done by iterative procedure
  - Each iteration has two steps
    - Successive optimization w.r.t.  $r_{nk}$  and  $\mu_k$

# Two Updating Stages

- First choose initial values for  $\mu_k$
- First phase:
  - minimize  $J$  w.r.t.  $r_{nk}$  keeping  $\mu_k$  fixed
- Second phase:
  - minimize  $J$  w.r.t.  $\mu_k$  keeping  $r_{nk}$  fixed
- Two stages correspond to E (expectation) and M (maximization) of EM algorithm

# E: Determination of Indicator $r_{nk}$

- Because  $J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|x_n - \mu_k\|^2$  is a linear function of  $r_{nk}$  this optimization is performed easily (closed form solution)
- Terms involving different  $n$  are independent
  - Therefore can optimize for each  $n$  separately
  - Choosing  $r_{nk}$  to be 1 for whichever value of  $k$  gives minimum value of  $\|x_n - \mu_k\|^2$
- Thus 
$$r_{nk} = \begin{cases} 1 & \text{if } k = \operatorname{argmin}_j \|x_n - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$
- Interpretation:
  - Assign  $x_n$  to cluster whose mean is closest

# M: Optimization of $\mu_k$

- Hold  $r_{nk}$  fixed
- Objective function  $J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \| \mathbf{x}_n - \mu_k \|^2$  is a quadratic function of  $\mu_k$
- Minimized by setting derivative w.r.t.  $\mu_k$  to zero

- Thus 
$$2 \sum_{n=1}^N r_{nk} (x_n - \mu_k) = 0$$

- Which is solved to give

- Interpretation:

- Set  $\mu_k$  to mean of all data points  $\mathbf{x}_n$  assigned to cluster  $k$

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

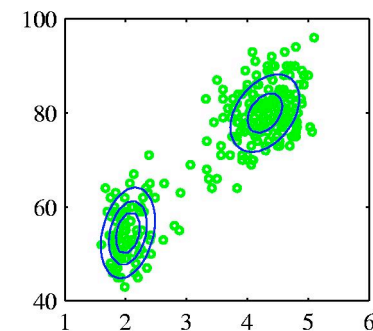
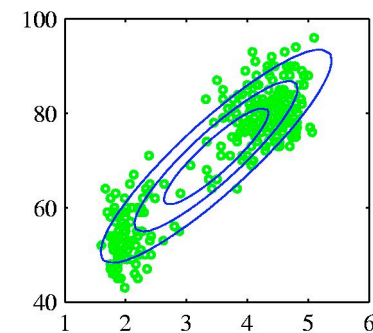
Equal to no of points assigned to cluster  $k$

# Termination of K-Means

- Two phases
  - re-assigning data points to clusters
  - Re-computing means of clusters
- Done repeatedly until no further change in assignments
- Since each phase reduces  $J$  convergence is assured
- May converge to local minimum of  $J$

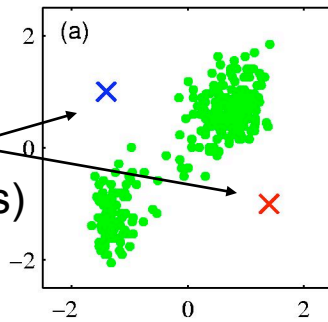
# Illustration of K-means

- Old Faithful dataset
- Single Gaussian is a poor fit
- We choose  $K = 2$
- Data set is standardized so each variable has zero mean and unit standard deviation
- Assignment of each data point to nearest cluster center is equivalent to
  - which side of the perpendicular bisector of line joining cluster centers

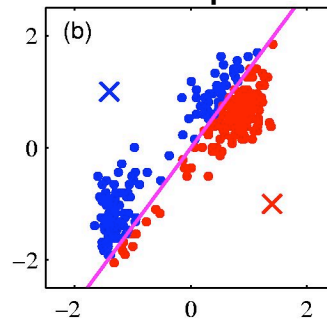


# K-means iterations

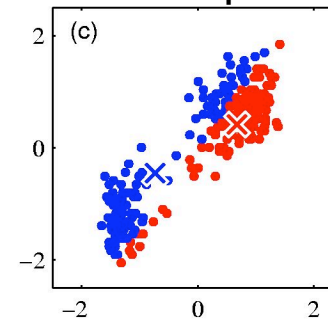
Initial  
Choice of  
Means  
(Parameters)



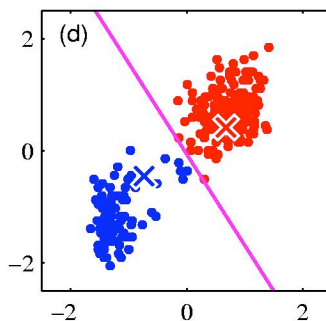
E step



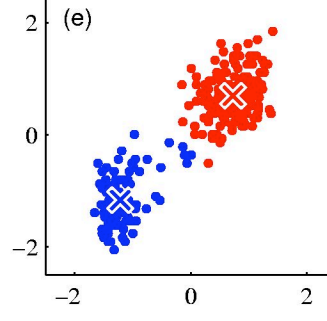
M step



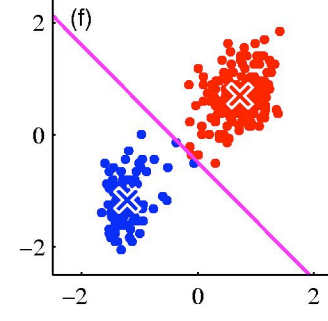
E step



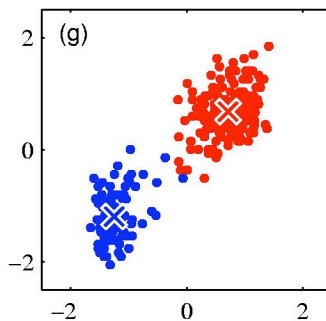
M step



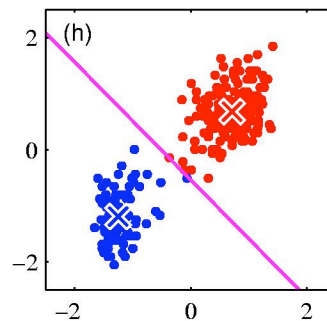
E step



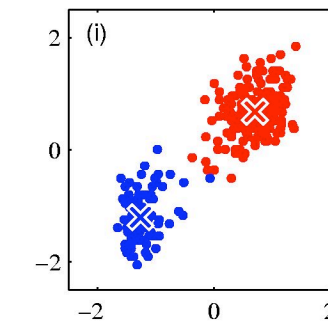
M step



E step



M step

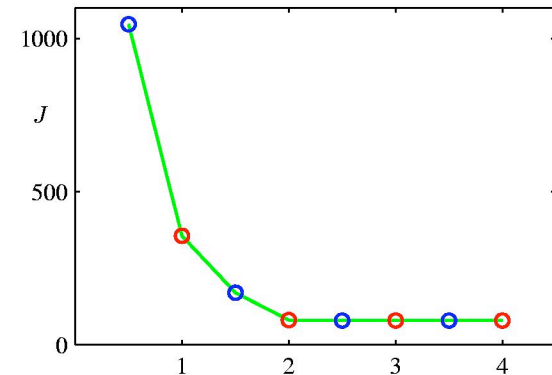


E step:  
parameters  
are fixed  
Distributions  
are  
optimized

M step:  
distributions  
are fixed  
Parameters  
are  
optimized

← Final  
Clusters  
And  
Means

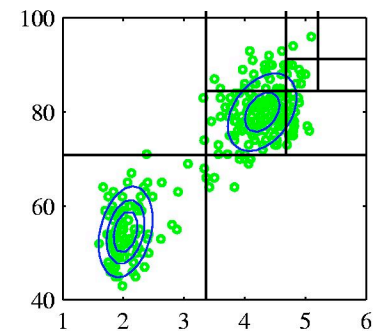
# Cost Function after Iteration



- $J$  for Old Faithful Data
- Poor initial value chosen for cluster centers
  - Several steps needed for convergence
  - Better choice is to assign  $\mu_k$  to random subset of  $k$  data points
- K-means is itself used to initialize parameters for Gaussian mixture model before applying EM

# Implementation of K-means

- Direct implementation can be slow
  - In E step Euclidean distances are computed between every mean and every data point
    - $\|x_n - \mu_k\|^2$  is computed for  $n=1, \dots, N$  and  $k=1, \dots, K$
- Faster implementations exist
  - Precomputing trees where nearby points are on same sub-tree
  - Use of triangle inequality to avoid unnecessary distance calculation



# On-line Stochastic Version

- Instead of batch processing entire data set
- Apply Robbins-Monro procedure
  - To finding roots of the regression function given by the derivative of  $J$  w.r.t  $\mu_k$

$$\mu_k^{new} = \mu_k^{old} + \eta_n (x_n - \mu_k^{old})$$

- *where*  $\eta_n$  is a learning rate parameter made to decrease monotonically as more samples are observed

# Dissimilarity Measure

- Euclidean distance has limitations
  - Inappropriate for categorical labels
  - Cluster means are non-robust to outliers
- Use more general dissimilarity measure  $v(\mathbf{x}, \mathbf{x}')$  and distortion measure

$$\tilde{J} = \sum_{n=1}^N \sum_{k=1}^K r_{nk} v(\mathbf{x}_n, \mathbf{u}_k)$$

- Which gives the *k-medoids* algorithm
- M-step is potentially more complex than for k-means

# Limitation of K-means

- Every data point is assigned uniquely to one and only one cluster
- A point may be equidistant from two cluster centers
- A probabilistic approach will have a 'soft' assignment of data points reflecting the level of uncertainty

# Image Segmentation and Compression

- Goal: partition image into regions
  - each of which has homogeneous visual appearance
  - or corresponds to objects
  - or parts of objects
- Each pixel is a point in R\_G\_B space
- K-means clustering is used with a palette of K colors
- Method does not take into account proximity of different pixels

# $K$ -means in Image Segmentation



Two examples where 2, 3, and 10 colors are chosen to encode a color image

# Data Compression

- Lossless data compression
  - Able to reconstruct data exactly from compressed representation
- Lossy data compression
  - Accept some error in return for greater compression
- K-means for Lossy compression
  - For each of  $N$  data points store only identity  $k$  of cluster center to which it is assigned
  - Store values of cluster centers  $\mu_k$  where  $K \ll N$
  - Vectors  $\mu_k$  are called *code-book vectors*
  - Method is called *Vector Quantization*
  - Data compression achieved
    - Original image needs  $24N$  bits (R,G,B need 8 bits each)
    - Compressed image needs  $24K + N \log_2 K$  bits
    - For  $K=2,3$  and  $10$ , compression ratios are 4%,8% and 17%