Expectation Maximization Algorithm

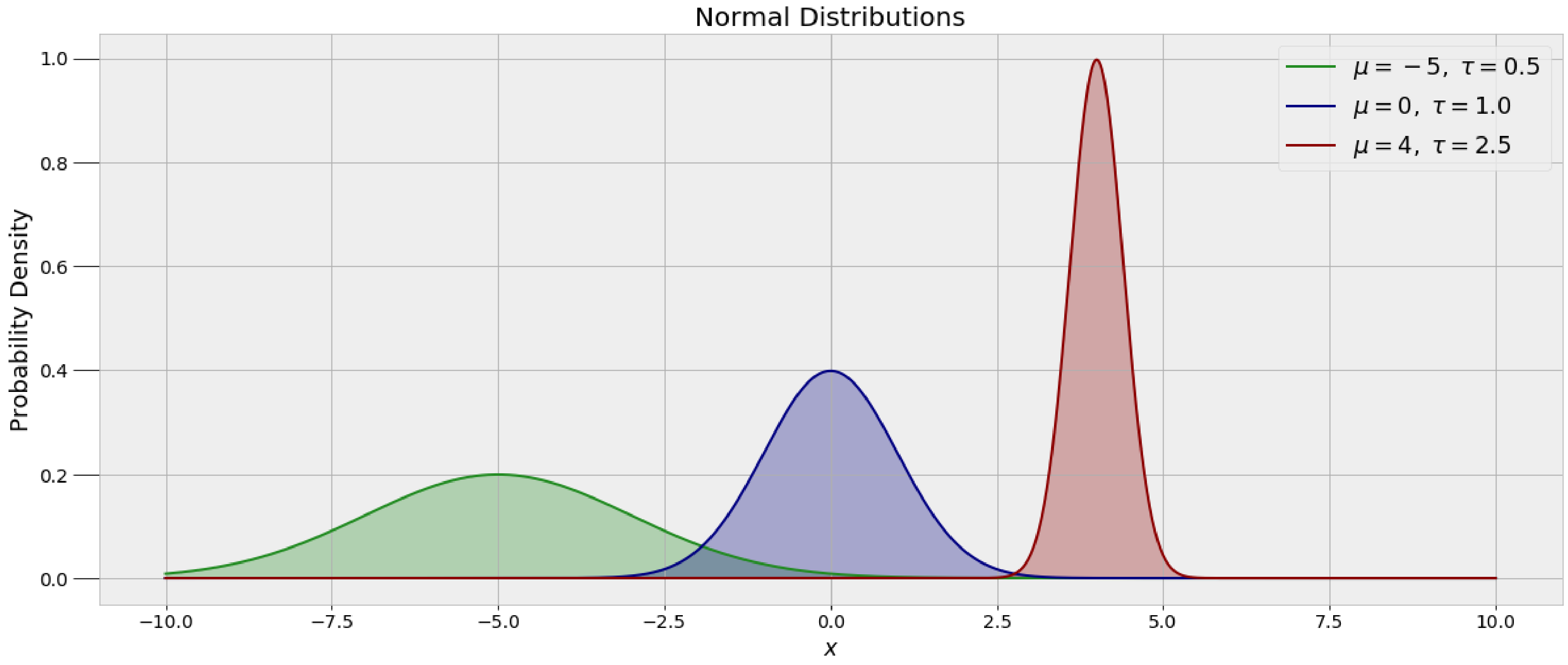
CS385 Machine Learning - Clustering

Outline

- Gaussian Distribution & Gaussian Learning
- Gaussian Mixture Model
- Expectation Maximization Algorithm (with GMM)
- Relation to K-means

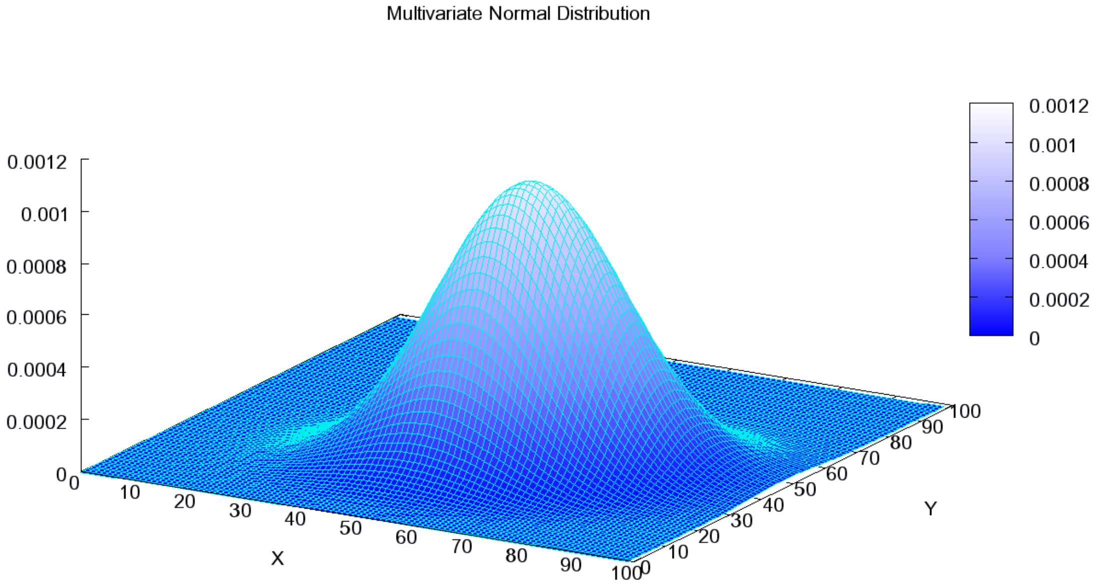
Gaussian/Normal Distribution – univariate

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



Gaussian/Normal Distribution - Multivariate

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



μ: (50,50)

Σ: $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

μ: d dimensional mean vector

Σ: k×k covariance matrix

|Σ|: determinant of Σ

Gaussian Learning - univariate

X
1
3
4
5
6
7
9

Assuming that the dataset follow a normal distribution,

Dataset is described by the normal distribution PDF

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Objective of Learning: estimate parameters (μ, σ)

ML Estimation method

X
1
3
4
5
6
7
9

data set X is i.i.d

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

Taking log

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

Partial derivation

Maximizing it with respect to μ

Maximizing it with respect to σ^2

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

$(\mu, \sigma^2) =$

Gaussian Learning - multivariate

X1	X2	X1-μ1	X2-μ2
6	6		
3	5		
4	4		
5	5		
6	4		
7	5		
4	6		
5	7		
5	3		

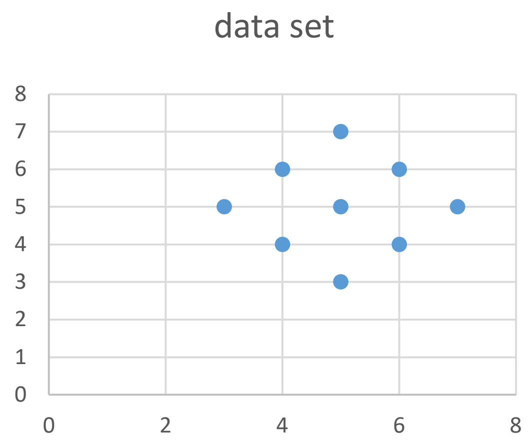
MLE

μ: (5,5) Σ: $\begin{bmatrix} 1.33 & 0 \\ 0 & 1.33 \end{bmatrix}$

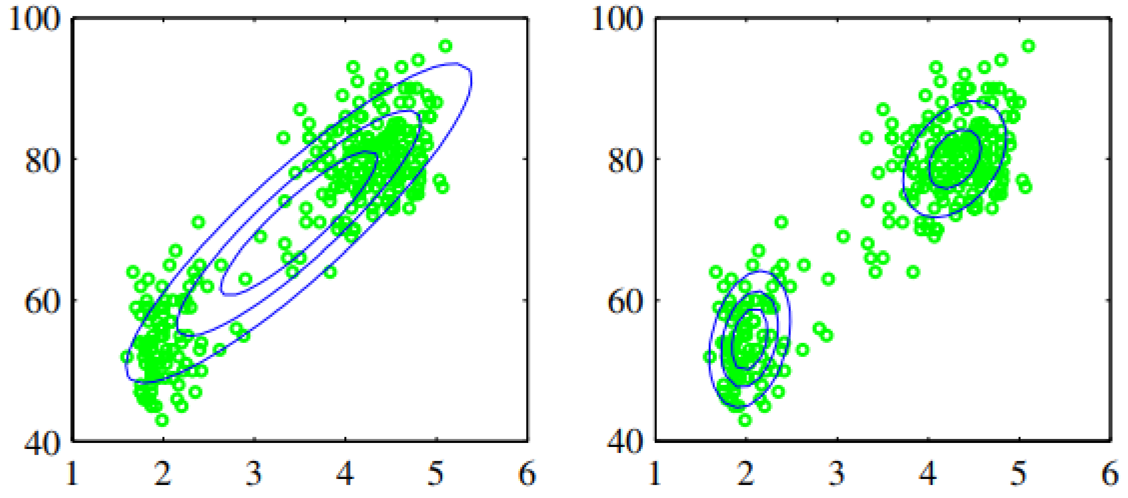
covariance: $\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))$

	X1	X2
X1	cov(X1,X1)	cov(X1,X2)
X2	cov(X2,X1)	cov(X2,X2)

	X1	X2
X1	1.33	0
X2	0	1.33



Gaussian Mixture Model – Motivation



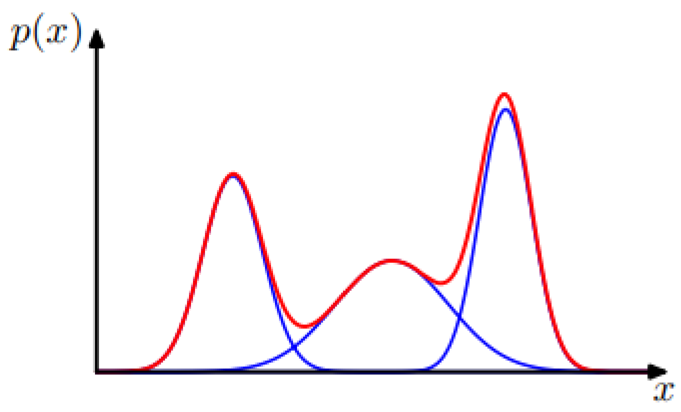
Single Gaussian distribution which has been fitted to (learnt from) the data using maximum likelihood.

fails to capture the two clusters in the data

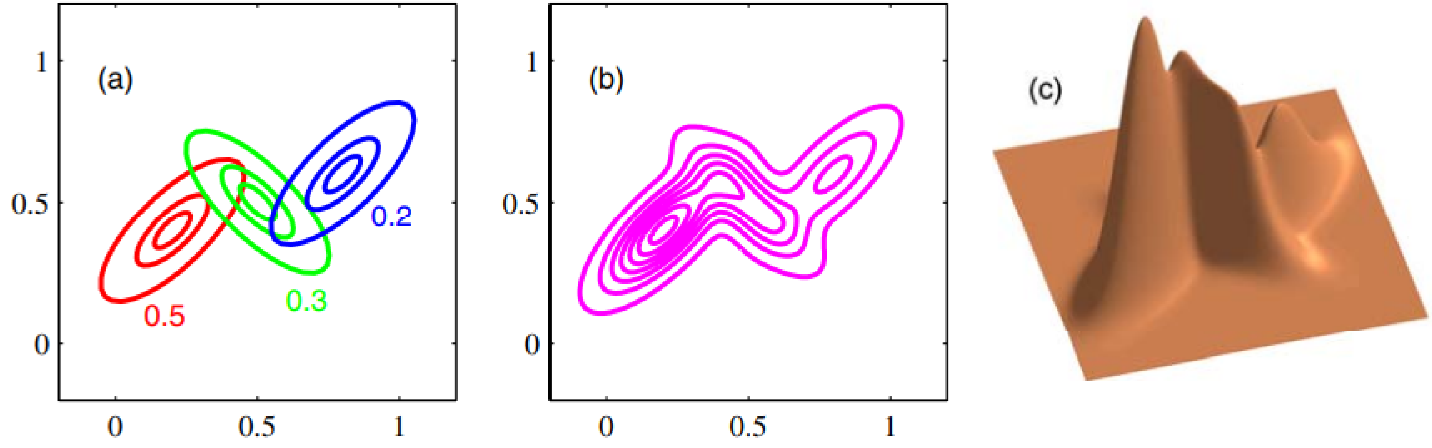
The distribution is given by a linear combination of two Gaussians

Gaussian Mixture Model

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad \sum_{k=1}^{K} \pi_k = 1$$



one dimension GMM
three Gaussians (each scaled by a coefficient) in blue and their sum in red



two dimension GMM
three Gaussians with coefficient

Gaussian Mixture Model – probability view

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad \sum_{k=1}^{K} \pi_k = 1$$

$$\Big\downarrow \text{ The sum and product rule}$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k) p(\mathbf{x}|k)$$

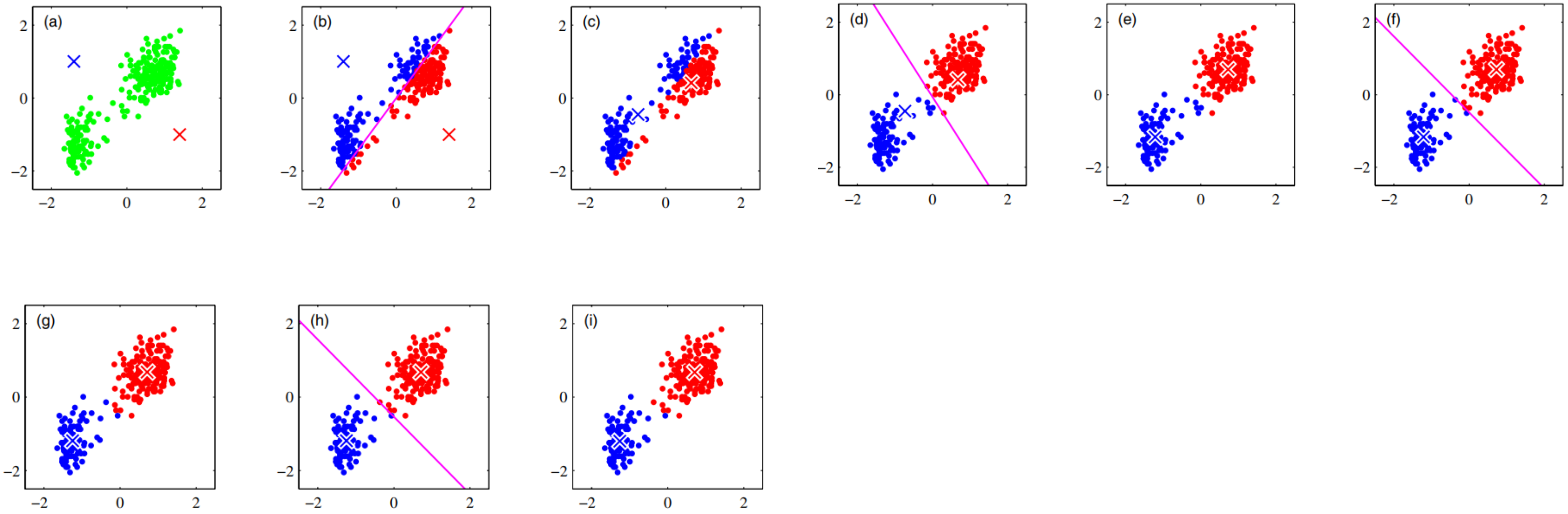
Gaussian Mixture Model – probability view

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad \sum_{k=1}^{K} \pi_k = 1$$

$$\Big\downarrow \text{ The sum and product rule}$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k) p(\mathbf{x}|k)$$

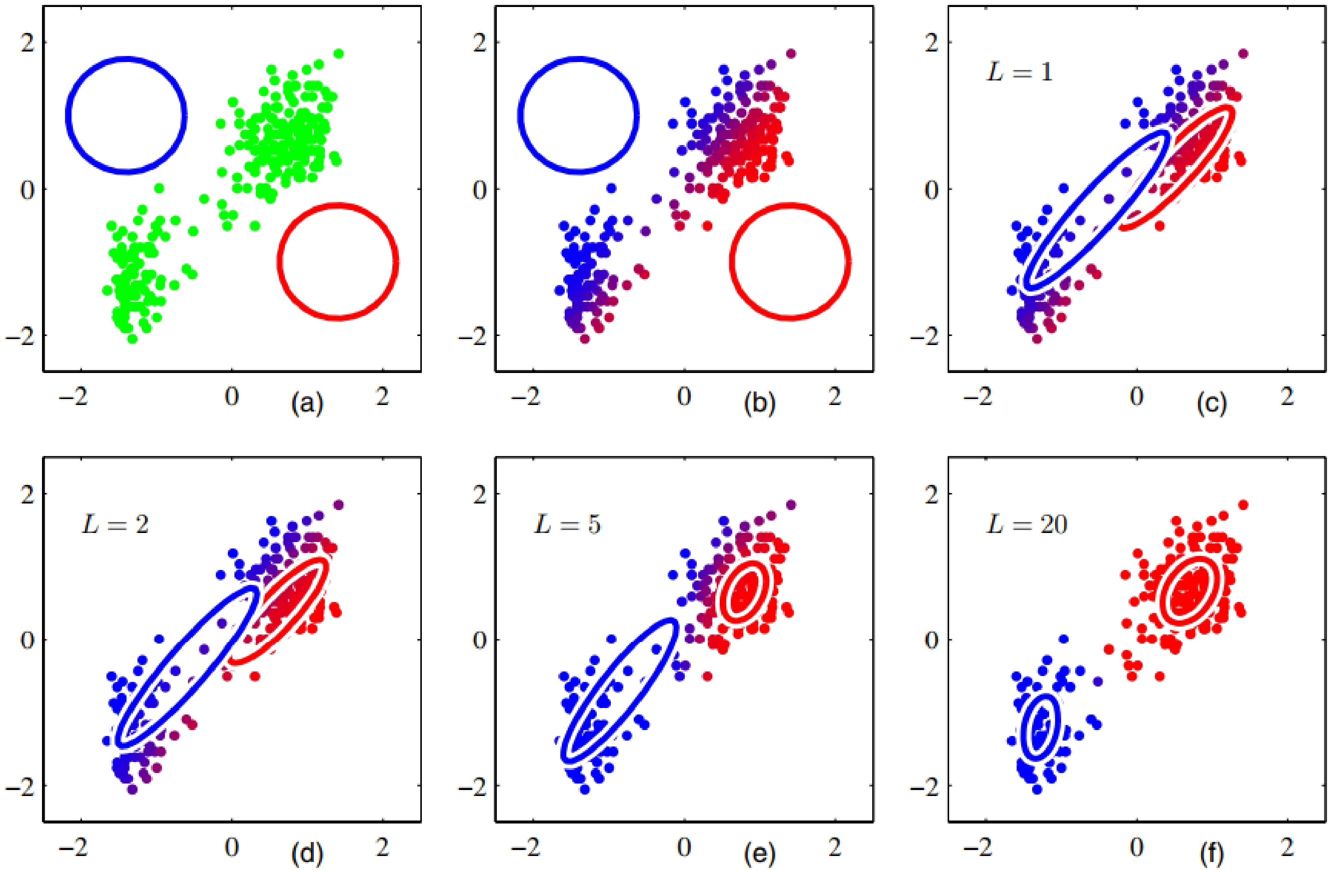
K-means review



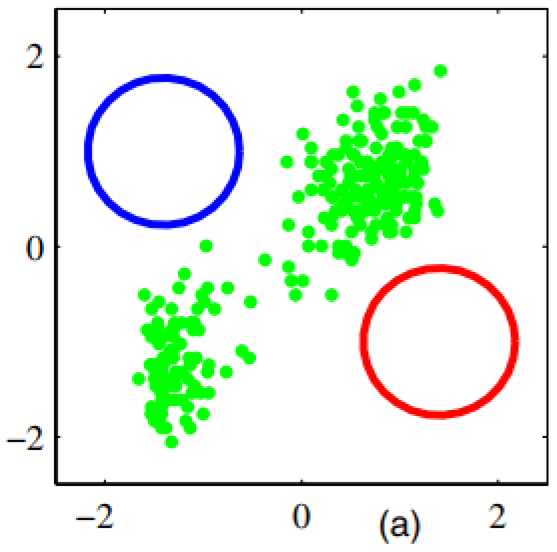
EM with K-means-like iteration

K=2

2 Gaussians



initialization

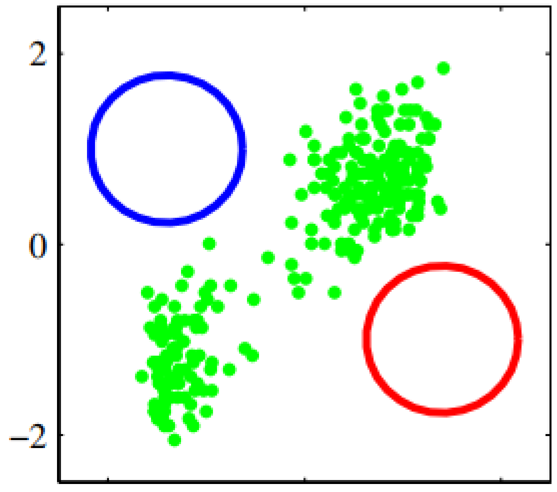


- How many parameters?
- K is set as 2 → 2 clusters
- Each cluster described by a single Gaussian
- Each Gaussian has two parameters μ and Σ
- Each object described by a GMM, and the prior of cluster is needed.

- 6 or 5 parameters

Initialization - example

- P(C=blue)=0.6, then P(C=red)=0.4
- Cluster Blue:
  - μ=(-1.5,1.5) /* all the values are made up */
  - Σ: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- Cluster Red:
  - μ=(1.5,-1)
  - Σ: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



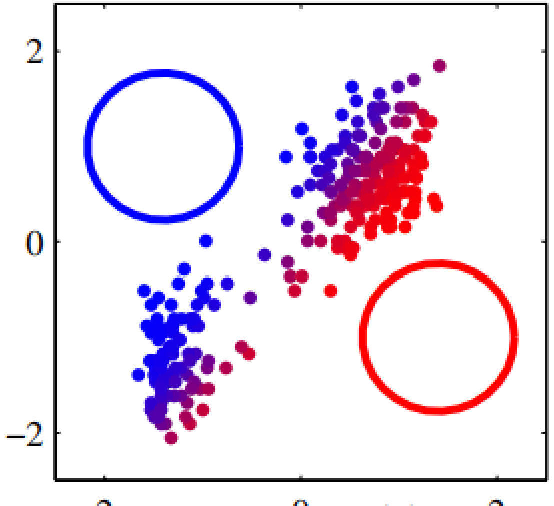
(a)

E-step

- For each object X(green dot), calculate p(X) using current values for the parameters
  - $p(C = Blue|X) = \frac{p(X|C=Blue)p(C=Blue)}{p(X|C=Blue)p(C=Blue)+p(X|C=Red)p(C=Red)}$

$$= \frac{\mathrm{N}(X|\mu_{blue},\Sigma_{blue})p(C=Blue)}{\mathrm{N}(X|\mu_{blue},\Sigma_{blue})p(C=Blue)+\mathrm{N}(X|\mu_{red},\Sigma_{red})p(C=Red)}$$

X1	X2	P(C=Blue|X)	P(C=Red|X)
0.6	1.6	0.8	0.2
-1.3	1.5	0.72	0.28
-0.44	0.4	0.1	0.9
1.5	-1.5	0.5	0.5
…	…		



(b)

M-step – Expectation Maximization

- Re-estimate the parameters by Expectation Maximization

$$p(\mathbf{x}) = \sum_{k=1}^{K} \underline{p(k)p(\mathbf{x}|k)}$$

$p(x, z|\theta) = p(z|\theta) \cdot p(x|\theta, k)$

$= p(z) \cdot p(x|k, \theta)$

θ: μ, Σ, $\pi \to p(k)$

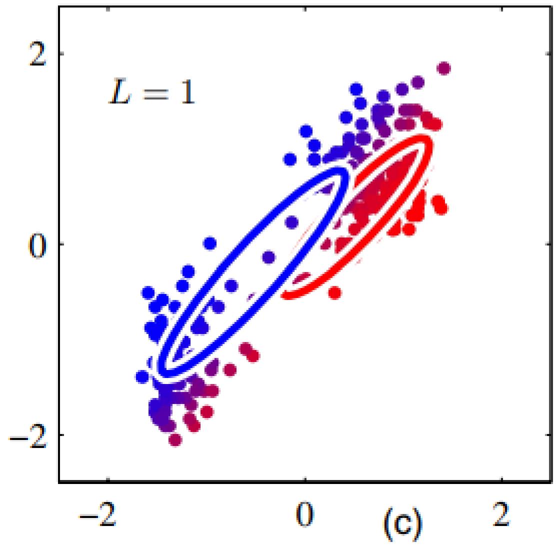
z: red or blue

- The expectation (log)

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} \underbrace{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})}_{\text{Expectation}} \ln \underline{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}$$

- Expectation Maximization $\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$

M-step – Expectation Maximization



- Re-estimate the parameters by Expectation Maximization
  - $\mu_{\text{blue}}^{\text{new}} = \frac{1}{\mathrm{N}_{blue}} \sum_{n=1}^{N} p(C = Blue|X_n)X_n$
  - $\Sigma_{\text{blue}}^{\text{new}} = \frac{1}{\mathrm{N}_{blue}} \sum_{n=1}^{N} p(C = Blue|X_n)(X_n - \mu_{\text{blue}}^{\text{new}})\,(X_n - \mu_{\text{blue}}^{\text{new}})^{\mathrm{T}}$
  - $p(c = blue) = \frac{N_{blue}}{N}$

X1	X2	P(C=Blue|X)	P(C=Red|X)
0.6	1.6	0.8	0.2
-1.3	1.5	0.72	0.28
-0.44	0.4	0.1	0.9
1.5	-1.5	0.5	0.5
…	…		

$$N_{blue} = \sum_{n=1}^{N} p(C = Blue|X_n)$$

Evaluate the log likelihood

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k) p(\mathbf{x}|k)$$

GMM

$$\sum_{n=1}^{N} \ln\{\underbrace{\textstyle\sum_{k=1}^{K} p(C=k)p(X_n|C=k)}\}$$

$$\sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} p(C=k) N(X_n|\mu_k, \Sigma_k)\}$$

X1	X2	P(C=Blue|X)	P(C=Red|X)
0.6	1.6	0.8	0.2
-1.3	1.5	0.72	0.28
-0.44	0.4	0.1	0.9
1.5	-1.5	0.5	0.5
…	…		

Termination: Check for convergence of either the parameters or the log likelihood

If the convergence criterion is not satisfied, go to E-step