

TFIP-AI – Advanced Machine Learning

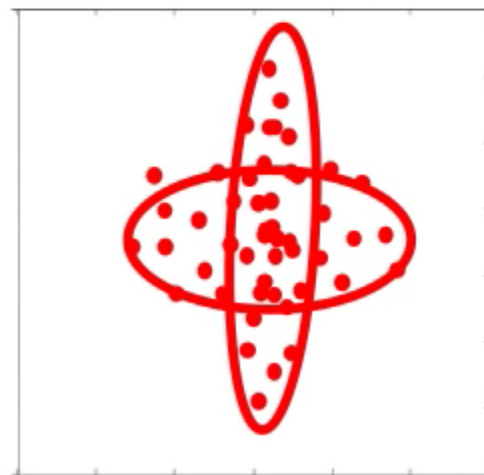
Unit 2 Clustering (Gaussian Mixture)

Iterative Schema + Gaussian →
EM

EM algorithm

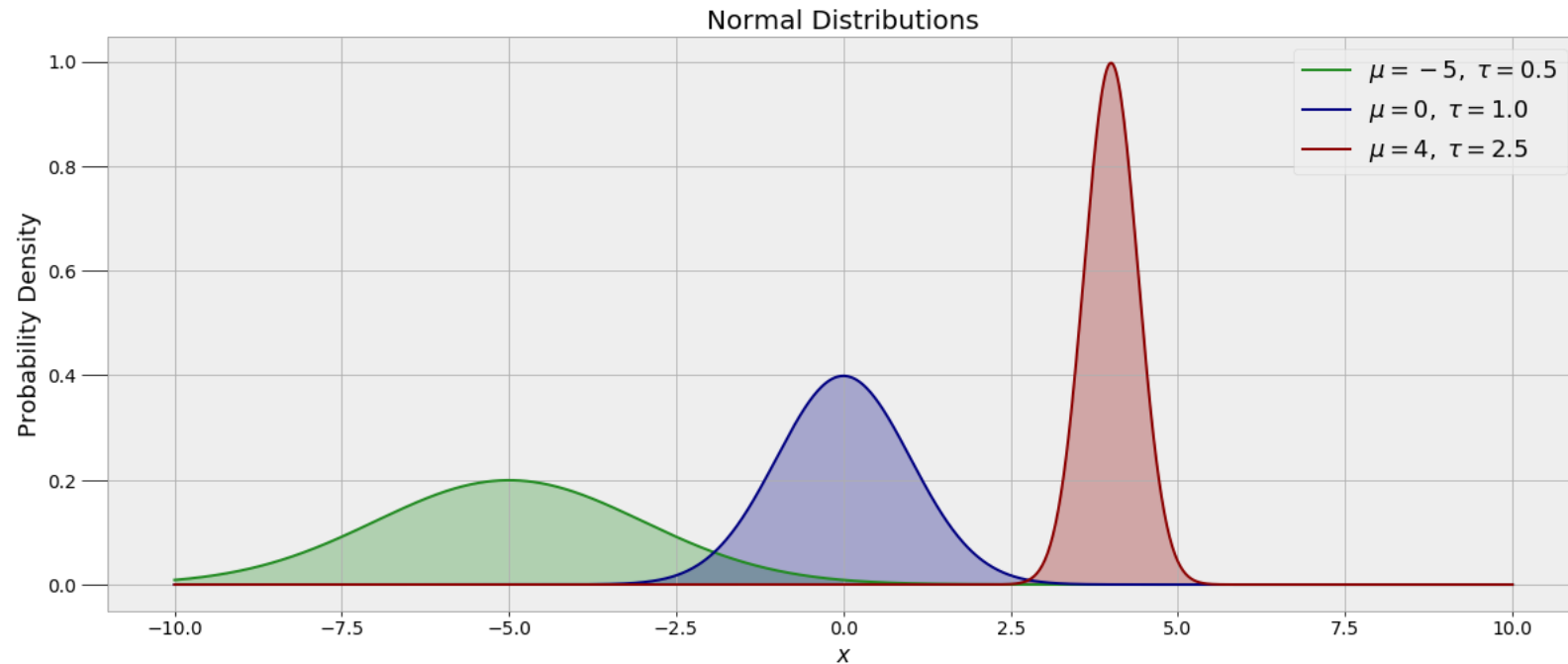
Mixtures of Gaussians

- K-means algorithm
 - Assigned each example to exactly one cluster
 - What if clusters are overlapping?
 - Hard to tell which cluster is right
 - Maybe we should try to remain uncertain
 - Used Euclidean distance
 - What if cluster has a non-circular shape?
- Gaussian mixture models
 - Clusters modeled as Gaussians
 - Not just by their mean
 - EM algorithm: assign data to cluster with some *probability*
 - Gives probability model of x ! (“generative”)



Gaussian/Normal Distribution – univariate

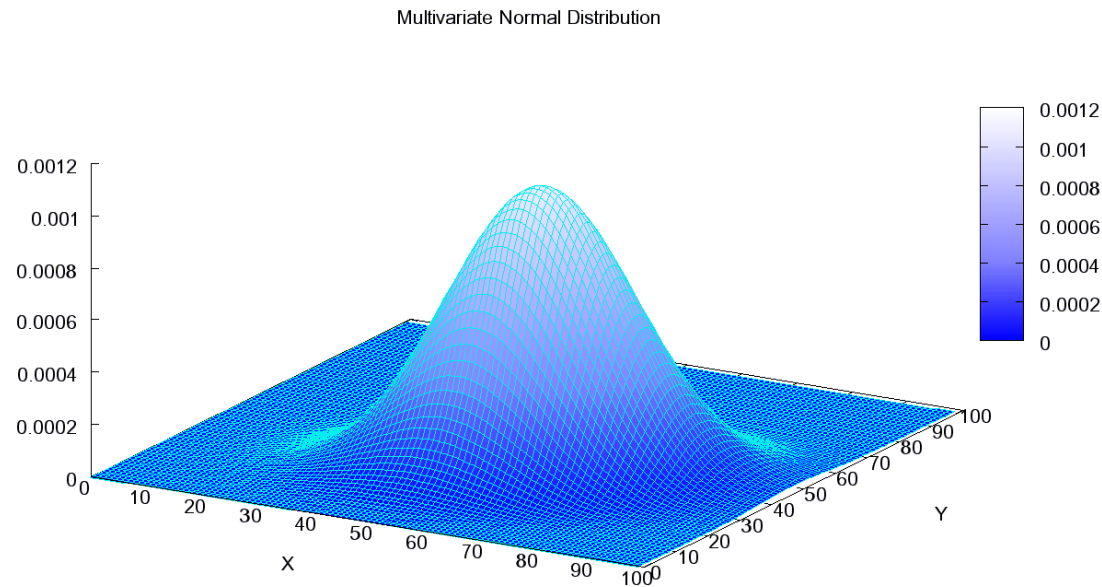
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\}$$



- Parameter: $p(XY|C = 1)$
- $p(XY|C = 1) = p(X|C = 1)p(Y|C = 1)$

Gaussian/Normal Distribution - Multivariate

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



$\boldsymbol{\mu}$: d dimensional mean vector

$\boldsymbol{\Sigma}$: k×k covariance matrix

$|\boldsymbol{\Sigma}|$: determinant of $\boldsymbol{\Sigma}$

$$\boldsymbol{\mu}: (50, 50)$$
$$\boldsymbol{\Sigma}: \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

- Parameter: $p(XY|C = 1)$

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

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

Taking log

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

Partial derivation

Maximizing it with respect to μ

Maximizing it with respect to σ^2

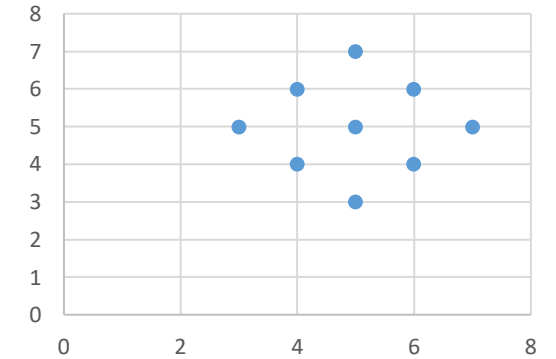
$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n$$

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$$

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

Gaussian Learning - multivariate

data set



X1	X2	X1- μ_1	X2- μ_2
6	6	1	1
3	5	-2	0
4	4	-1	-1
5	5	0	0
6	4	1	-1
7	5	2	0
4	6	-1	1
5	7	0	2
5	3	0	-2

MLE

$$\mu: (5,5) \quad \Sigma: \begin{bmatrix} 1.33 & 0 \\ 0 & 1.33 \end{bmatrix}$$

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

	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

X	Y	$p(c=1 (x,y))$	$p(c=0 (x,y))$
5	8		
4	7		
8	9		
6	8		
8	2		
7	1		
5	2		

K-Means like scheme

1. Randomly select u_k, σ_k, π_k ($k = 1..K$)
2. Calculate $p(c = k|(xy))$
3. Update u_k, σ_k, π_k
4. Repeat step 2 until convergence

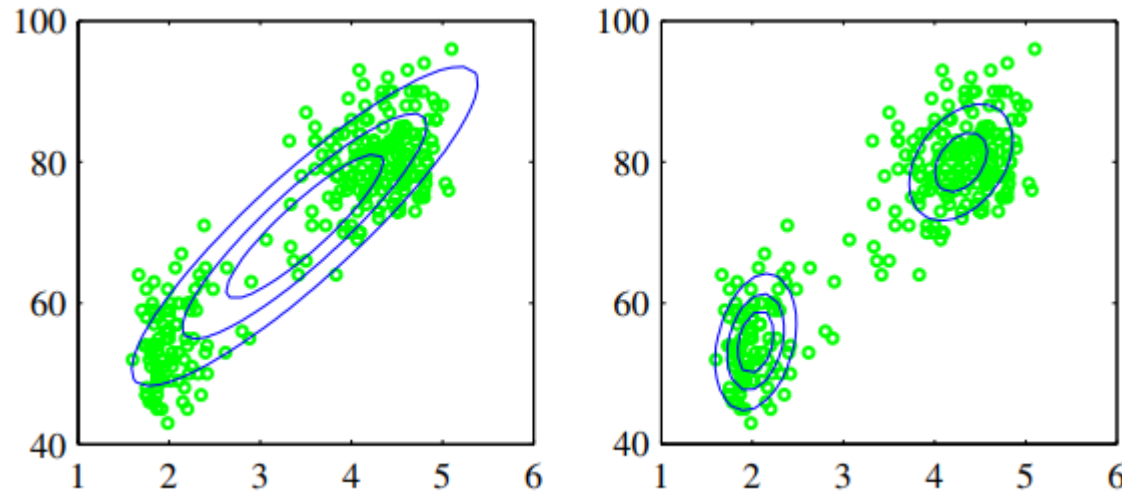
$$p(c = 1|(5,8)) \propto p((5,8)|c = 1) \times p(c = 1) = N((5,8)|u_1, \sigma_1) \times \pi_1$$

GMM \rightarrow EM

MLE

- Review MLE Method
 - Step1: likelihood function $p(\text{dataset}|\theta)$
 - Step2: taking log of $p(\text{dataset}|\theta)$
 - Step3: taking partial derivative wrt. θ , and equate to zero
- Question
 - For a supervised problem, $p(x|\theta_1)$ if x is in C_1 (the class label is known)
 - For an unsupervised problem: ?

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

Partitioning Algorithms

- **K-means**

- hard assignment:** each object belongs to only one cluster

$$\theta_i \in \{\theta_1, \dots, \theta_K\}$$

- **Mixture modeling**

- soft assignment:** probability that an object belongs to a cluster

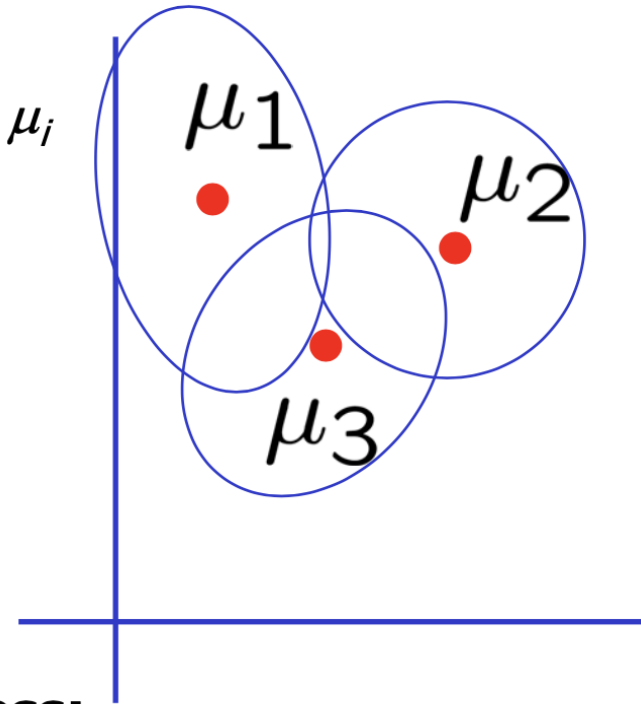
$$(\pi_1, \dots, \pi_K), \pi_i \geq 0, \sum_{i=1}^K \pi_i = 1$$

Gaussian Mixture Model

Mixture of K Gaussians distributions: (Multi-modal distribution)

- There are K components
- Component i has an associated mean vector μ_i

Component i generates data from $N(\mu_i, \Sigma_i)$



Each data point is generated using this process:

- 1) Choose component i with probability $\pi_i = P(y = i)$
- 2) Datapoint $x \sim N(\mu_i, \Sigma_i)$

Gaussian Mixture Model cont...

Mixture of K Gaussians distributions: (Multi-modal distribution)

Hidden variable

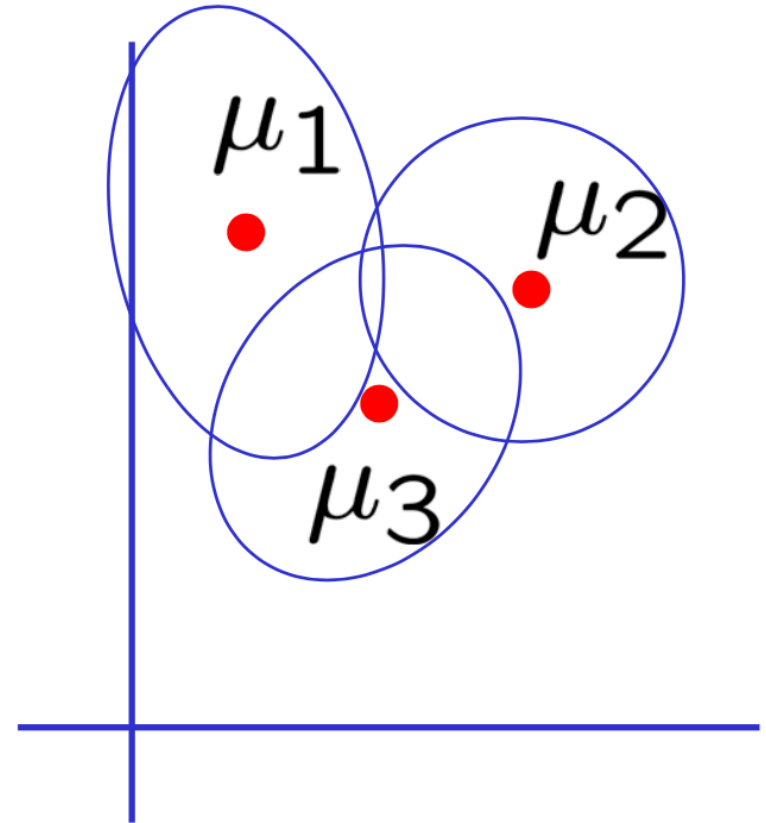
$$p(x|y = i) = N(\mu_i, \Sigma_i)$$

$$p(x) = \sum_{i=1}^K p(x|y = i)P(y = i)$$

Observed
data

Mixture
component

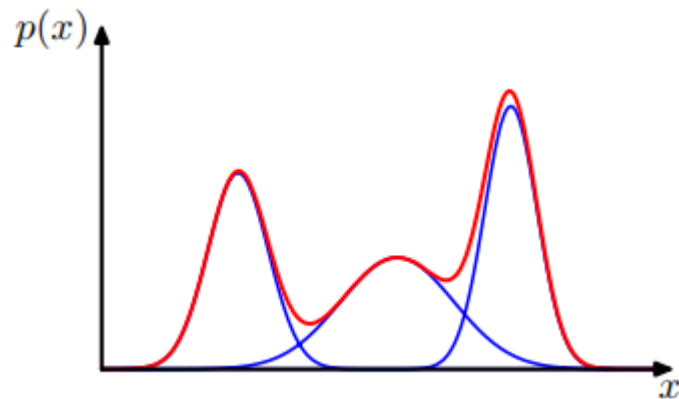
Mixture
proportion



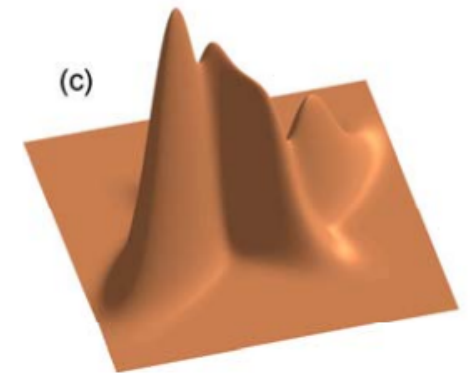
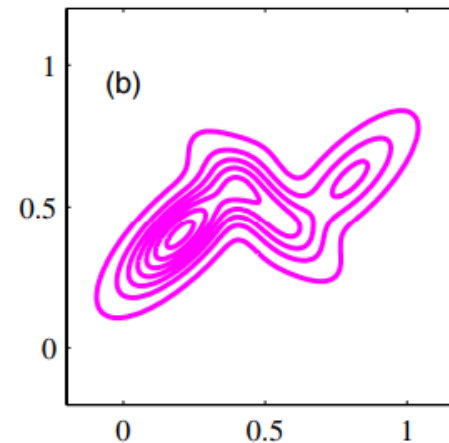
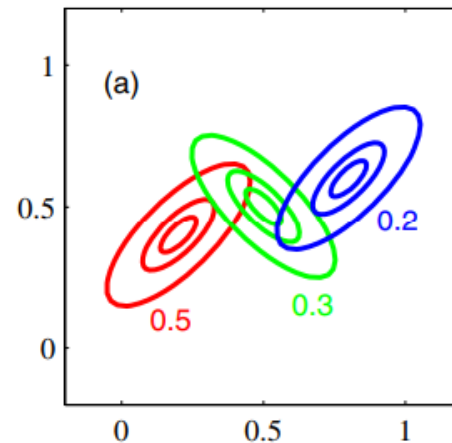
Gaussian Mixture Model

For a sample \mathbf{x} , $p(\mathbf{x}|\theta) = \sum_{k=1}^K p(\mathbf{x}|c = k)p(c = k)$

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



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



two dimension GMM
three Gaussians with coefficient

Log-likelihood

- For each sample \mathbf{x} which follow a GMM

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

- The log-likelihood for the dataset
 - K: number of classes
 - N: number of samples

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Taking derivative w.r.t. (μ, Σ, π) and equate to zero

$$\begin{aligned}\mu_k &= \frac{1}{N_k} \sum_{n=1}^N z_{nk} \mathbf{x}_n \\ \Sigma_k &= \frac{1}{N_k} \sum_{n=1}^N z_{nk} (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \\ \pi_k &= \frac{N_k}{N}\end{aligned}$$

$z_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \mu_j, \Sigma_j)}$

$P(C = k | X_n)$

$$N_k = \sum_{n=1}^N z_{nk}$$

However, it is not a closed-form solution
e.g. to calculate μ_k , we need to know the
values of all other parameters

Solution: iterative scheme

- Try to find a correct solution
 - Maximizing the log likelihood function for a GMM turns out to be a more complex problem than for the case of a single Gaussian.
 - The difficulty arises from the presence of the summation over k that appears inside the logarithm.
- Go back and observe the result of setting derivatives of the log likelihood to zero
 - The results do suggest a simple iterative scheme for finding a solution

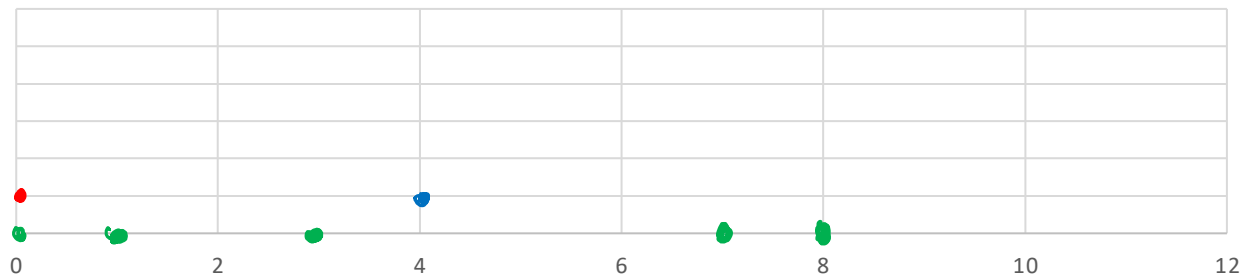
EM - 1-d Example - Estep

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

$\uparrow 3.14$

X	N(x μ1,σ1)	N(x μ2,σ2)	P(C=1 X)	P(C=2 X)
0	0.40	0.0001	1	0
1	0.24	0.0044	0.98 = $\frac{0.5 \times 0.24}{0.5 \times 0.24 + 0.5 \times 0.0044}$	0.02
3	0.004	0.242	0.02	0.98
7	0	0.0044	0	1
8	0	0.0001	0	1

$$\begin{aligned} \mu_1 &= 0, \sigma_1 = 1, \pi_1 = 0.5 \\ \mu_2 &= 4, \sigma_2 = 1, \pi_2 = 0.5 \end{aligned}$$



$$z_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

EM - 1-d Example - Mstep

X	P(C=1 X)	P(C=2 X)
0	1	0
1	0.98	0.02
3	0.02	0.98
7	0	1
8	0	1
N _k	2	3

$$\mu_1=0, \sigma_1=1, \pi_1=0.5$$

$$\mu_2=4, \sigma_2=1, \pi_2=0.5$$

$$\mu_1=1.04/2=0.52, \sigma_1=1.16/2=0.58, \pi_1=\frac{2}{5}=0.4$$
$$\mu_2=17.96/3=5.99, \sigma_2=14.32/3=4.77, \pi_2=0.6$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N z_{nk} \mathbf{x}_n \quad \pi_k = \frac{N_k}{N}$$
$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N z_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

EM - 1-d Example - Mstep

X	P(C=1 X)	P(C=2 X)
0	1	0
1	0.98	0.02
3	0.02	0.98
7	0	1
8	0	1
N _k	2	3

$$\mu_1=0, \sigma_1=1, \pi_1=0.5$$

$$\mu_2=4, \sigma_2=1, \pi_2=0.5$$

$$\pi_1 = \frac{2}{5}$$

$$\mu_1=1.04/2=0.52, \sigma_1=0.62/2=0.31, \pi_1=0.4$$

$$\mu_2=17.96/3=5.99, \sigma_2=26.16/3=8.72, \pi_2=0.6$$

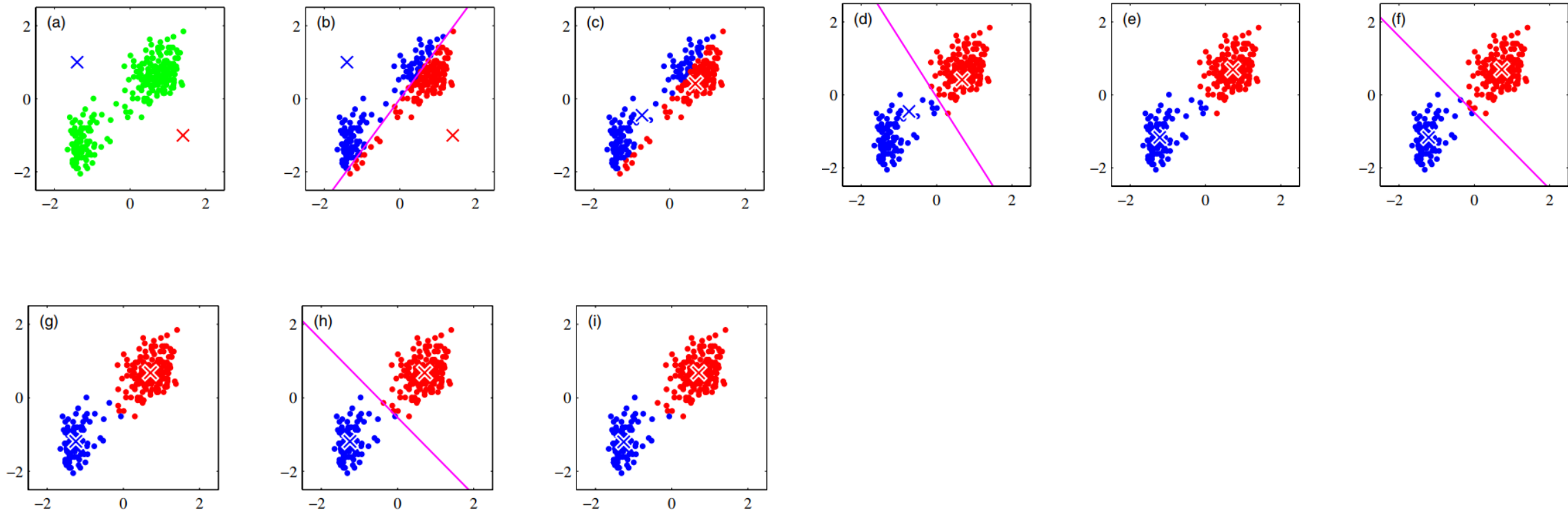
$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N z_{nk} \mathbf{x}_n \quad \pi_k = \frac{N_k}{N}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N z_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

$$u_1 = \frac{0 \times 1 + 1 \times 0.98 + 3 \times 0.02 + 7 \times 0 + 8 \times 0}{2} = 0.52$$

$$\sigma_2 = \frac{0 \times (0 - 4)^2 + 0.02 \times (1 - 4)^2 + 0.98 \times (3 - 4)^2 + 1 \times (7 - 4)^2 + 1 \times (8 - 4)^2}{3}$$

K-means



EM with K-means-like iteration

$$P(c=1|x_g) \propto \underbrace{P(x_g|c=1)}_{\cdot \pi_1} \cdot \underbrace{P(c=1)}_{\cdot \eta_1}$$

$$= \mathcal{N}(x_g|\mu, \sigma^2) \cdot \eta_1$$

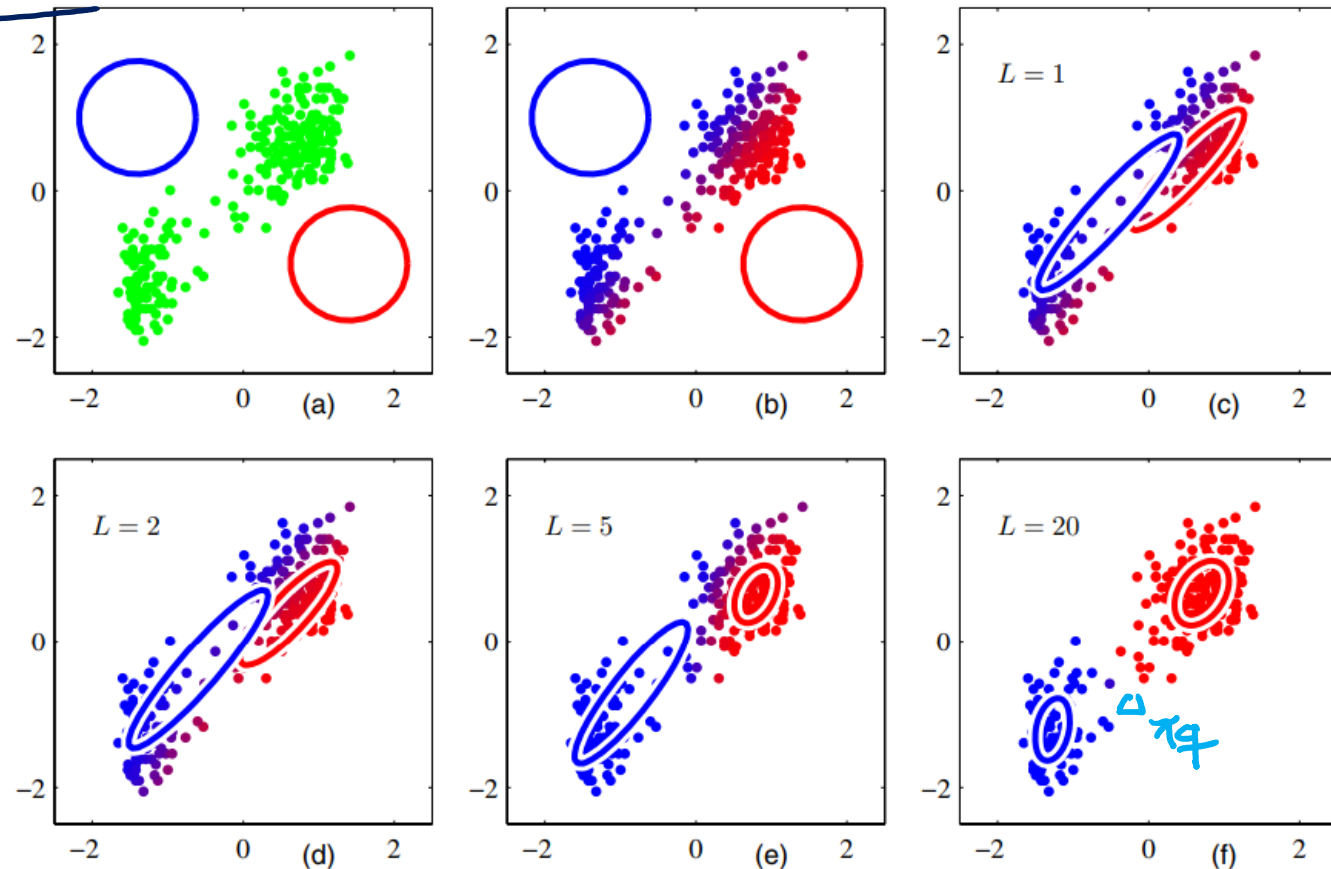
K=2

2 Gaussians

L: number of iteration

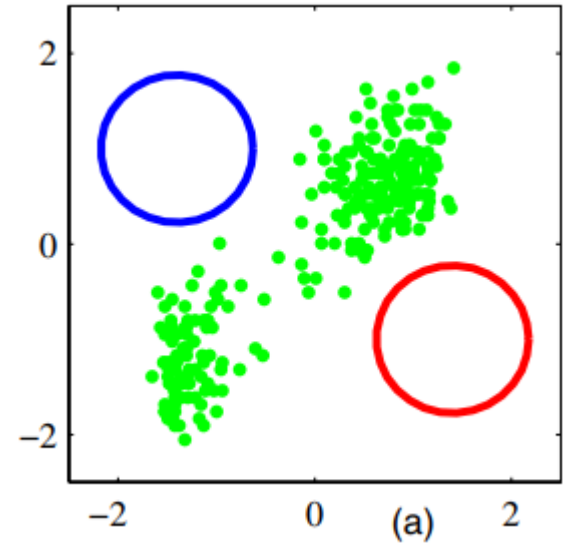
$$P(c=1|x_g) = 60\%$$

$$P(c=2|x_g) = 40\%$$



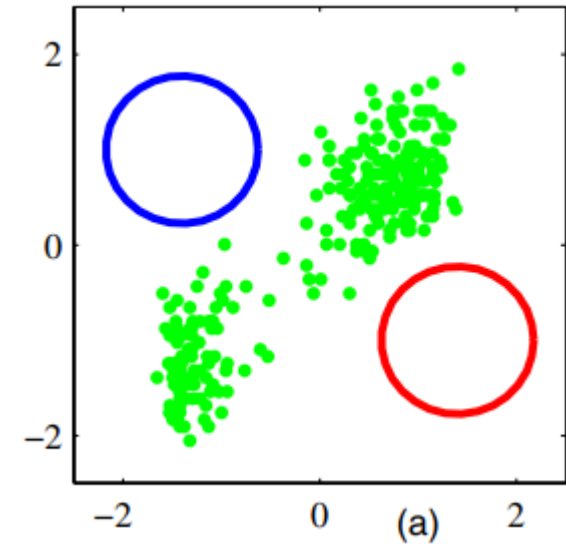
initialization

- How many parameters?
- K is set as 2 \rightarrow 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

- $P(C=\text{blue})=0.6$, then $P(C=\text{red})=0.4$
- Cluster Blue:
 - $\mu=(-1.5, 1.5)$
 - $\Sigma: \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- Cluster Red:
 - $\mu=(1.5, -1)$
 - $\Sigma: \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



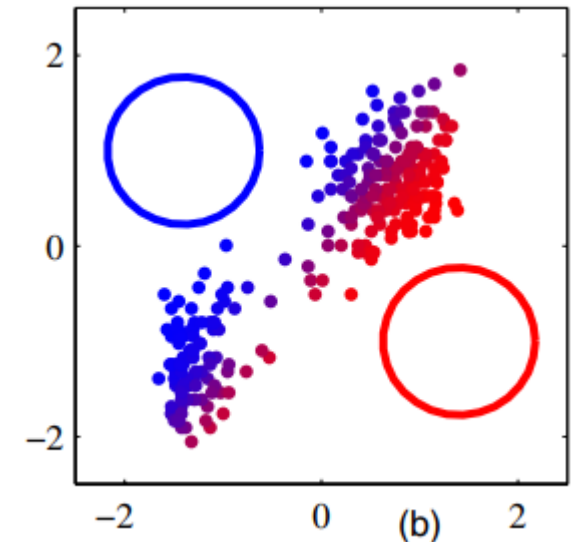
E-step

- For each object X (green dot), calculate $p(X)$ using current values for the parameters

$$p(C = \text{Blue} | X) = \frac{p(X|C=\text{Blue})p(C=\text{Blue})}{p(X|C=\text{Blue})p(C=\text{Blue}) + p(X|C=\text{Red})p(C=\text{Red})}$$

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

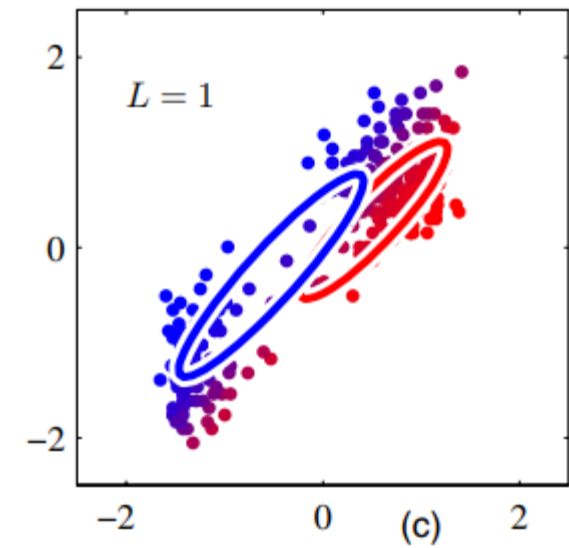
X1	X2	P(C=Blue X)	P(C=Red X)
0.6	1.6	0.8	0.2
-1.3	1.5	0.72	0.28
-0.44	0.4	0.1	0.9
1.5	-1.5	0.5	0.5
...	...		



M-step – Expectation Maximization

- Re-estimate the parameters by Expectation Maximization

- $\mu_{\text{blue}}^{\text{new}} = \frac{1}{N_{\text{blue}}} \sum_{n=1}^N p(C = \text{Blue} | X_n) X_n$
- $\Sigma_{\text{blue}}^{\text{new}} = \frac{1}{N_{\text{blue}}} \sum_{n=1}^N p(C = \text{Blue} | X_n) (X_n - \mu_{\text{blue}}) (X_n - \mu_{\text{blue}})^T$
- $p(c = \text{blue}) = \frac{N_{\text{blue}}}{N}$



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
...	...		

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

Evaluate the log likelihood

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

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

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

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

EM and missing data

- EM is a general framework for partially observed data
 - “Complete data” x_i, z_i – features and assignments
 - Assignments z_i are missing (unobserved)
- EM corresponds to
 - Computing the distribution over all z_i given the parameters
 - Maximizing the “expected complete” log likelihood
 - GMMs = plug in “soft assignments”, but not always so easy
- Alternatives: Stochastic EM, Hard EM
 - Instead of expectations, just sample the z_i or choose best (often easier)
 - Called “imputing” the values of z
 - Hard EM: similar to EM, but less “smooth”, more local minima
 - Stochastic EM: similar to EM, but with extra randomness
 - Not obvious when it has converged