Clustering: Models and Algorithms

Shikui Tu

Shanghai Jiao Tong University

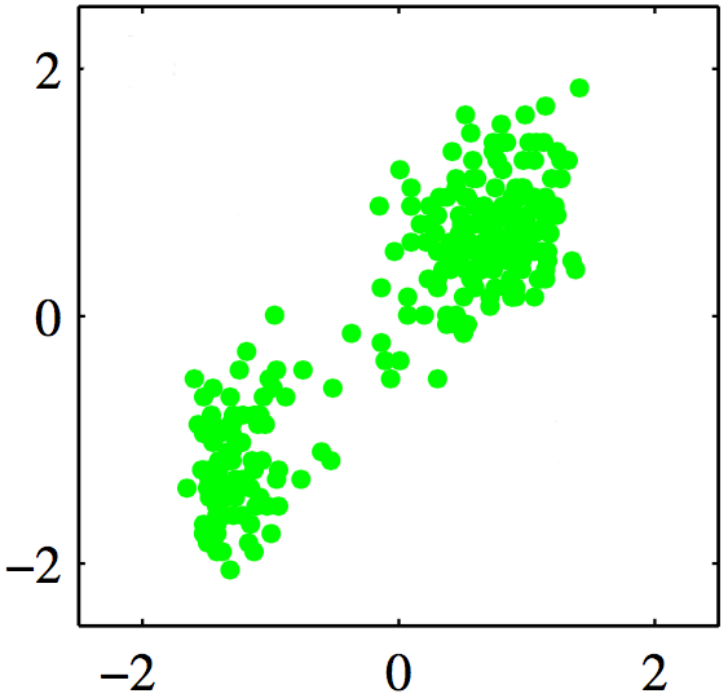
2021-03-09

Outline

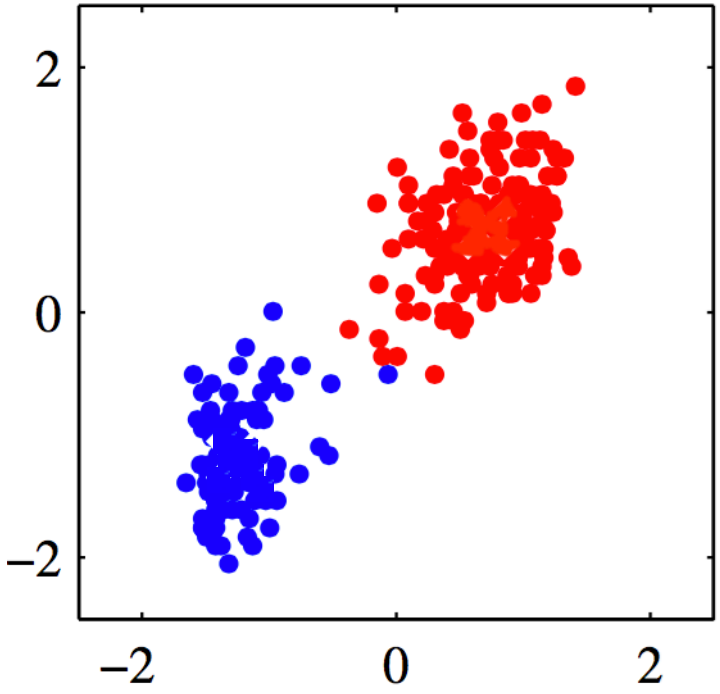
- A brief review of K-mean clustering
- Gaussian Mixture Models (GMM)
- Expectation-Maximization (EM) for maximum likelihood

Clustering the data

We have the following data:



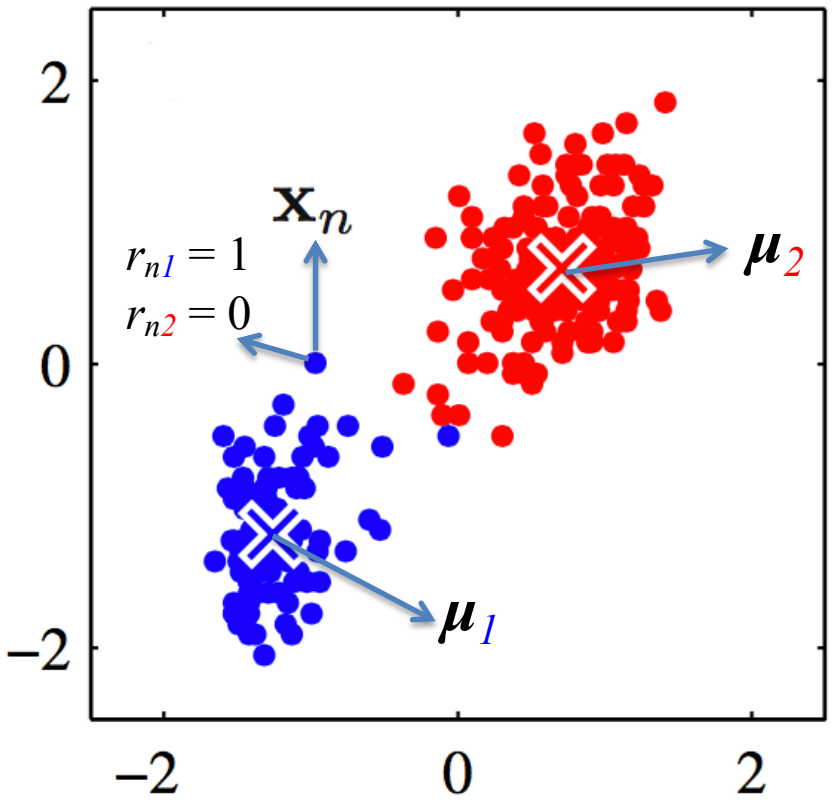
We want to cluster the data into two clusters (red and blue)



How?

Minimize the sum of square distances J

$$\text{minimize} \quad J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$



$r_{nk} = 1$ if and only if data point $\mathbf{x}_n$ is assigned to cluster k;
otherwise $r_{nk} = 0$.

$k = 1, 2$; K = 2 clusters

$n = 1, \ldots, N$;
N: the total number of points.

We need to calculate $\{ r_{nk} \}$ and $\{ \boldsymbol{\mu}_k \}$.

If we know r_{n1} , r_{n2} for all n=1,…,N

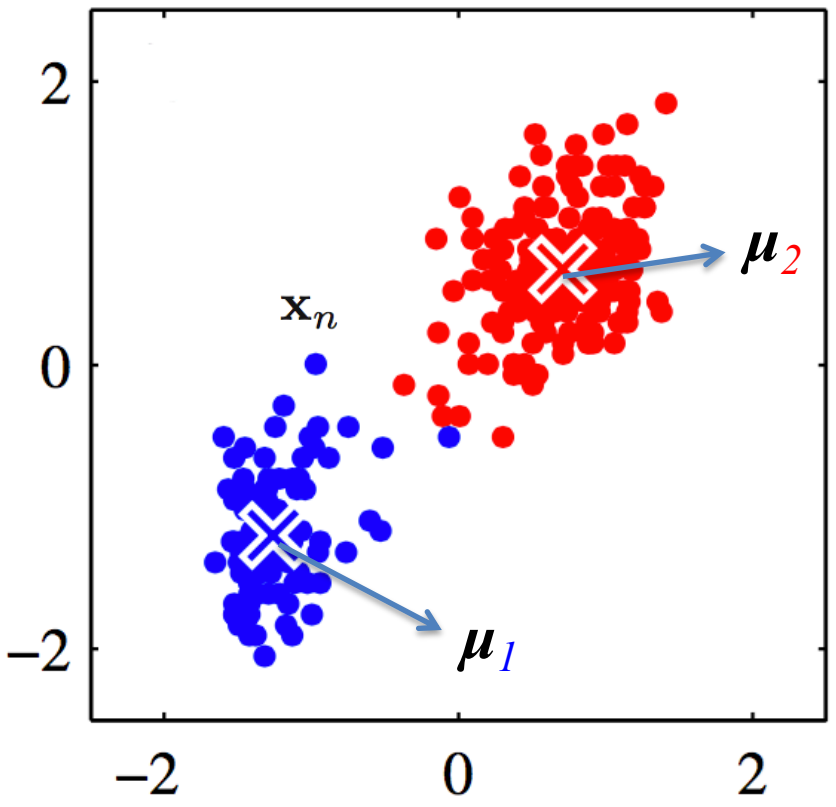
Since the points have been assigned to cluster 1 or cluster 2, we calculate

$\boldsymbol{\mu}_1$ = mean of the points in cluster 1

$\boldsymbol{\mu}_2$ = mean of the points in cluster 2

Or formally

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$



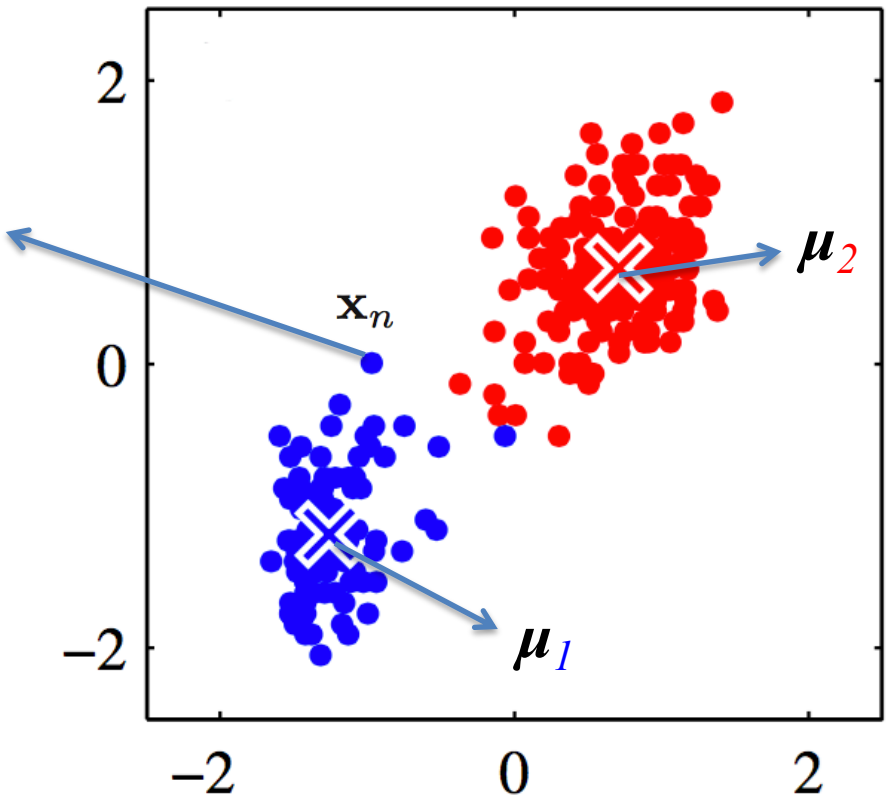
We call it the **M Step.**

If we know $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$

We should assign point $\mathbf{x}_n$ to cluster 1, because

$$\| \mathbf{x}_n - \boldsymbol{\mu}_1 \|^2 < \| \mathbf{x}_n - \boldsymbol{\mu}_2 \|^2$$

Then, $r_{n1} = 1$
$r_{n2} = 0$

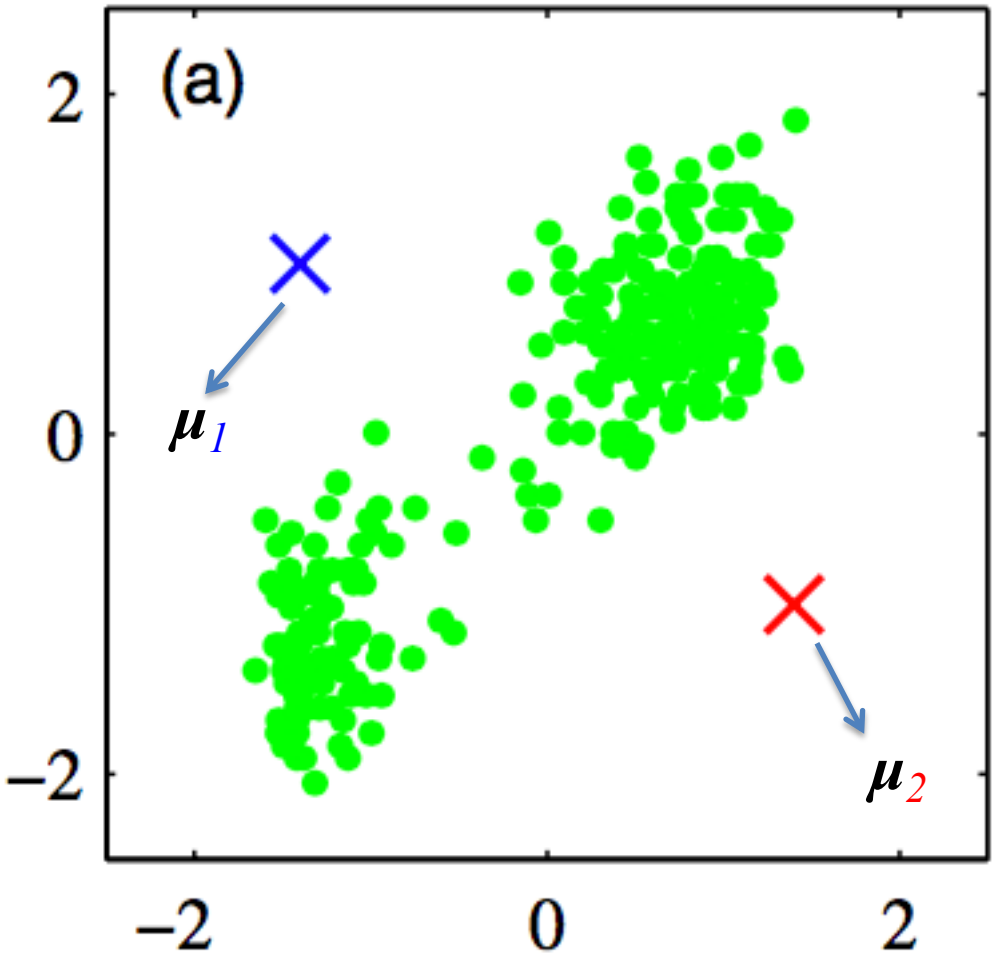


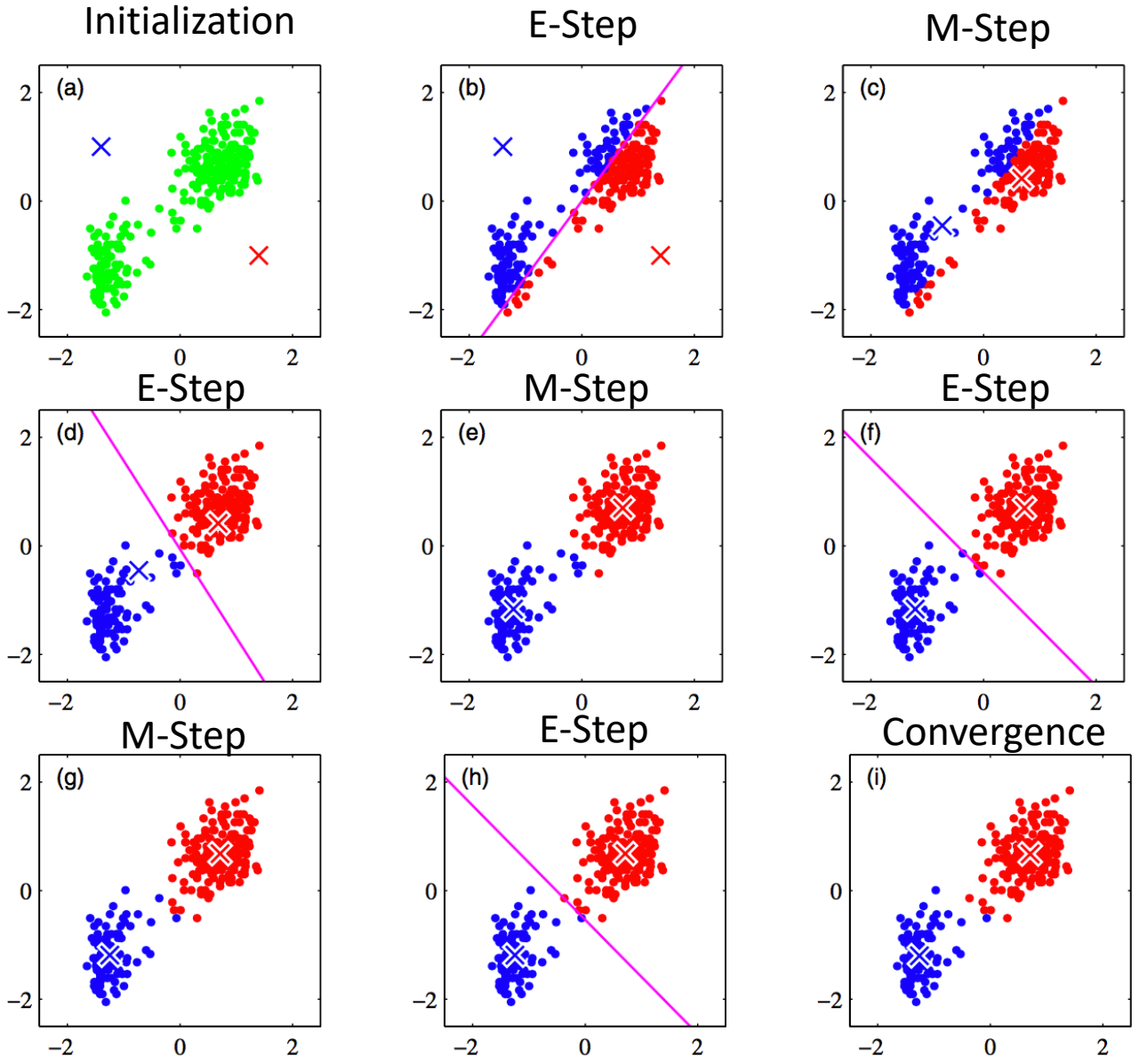
Or formally

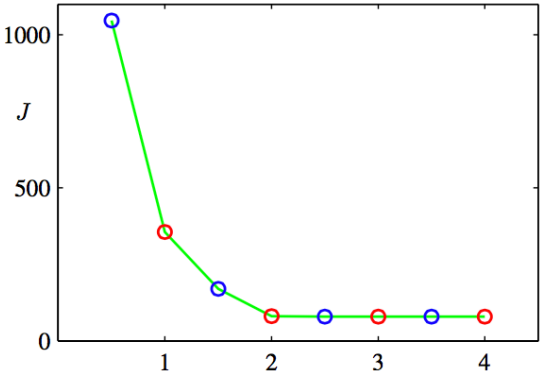
$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

We call it the **E Step**

Initialization







If J does not change, or $\{ \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \}$ do not change, then the algorithm converges.

K均值法小结

- 初始化均值点$\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k$
- 迭代如下
  - 把每个数据点按照就近原则分配给相应的$\boldsymbol{\mu}_i$
  - 把$\boldsymbol{\mu}_i$更新为所分配的数据点的均值
- 迭代停止，如果聚类分配不变

Initialize $\boldsymbol{m}_i, i = 1, \ldots, k$, for example, to k random $\boldsymbol{x}^t$

Repeat

For all $\boldsymbol{x}^t \in \mathcal{X}$

$$b_i^t \leftarrow \begin{cases} 1 & \text{if } \|\boldsymbol{x}^t - \boldsymbol{m}_i\| = \min_j \|\boldsymbol{x}^t - \boldsymbol{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$

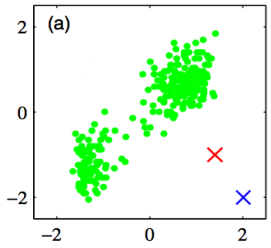
For all $\boldsymbol{m}_i, i = 1, \ldots, k$

$$\boldsymbol{m}_i \leftarrow \sum_t b_i^t \boldsymbol{x}^t / \sum_t b_i^t$$

Until $\boldsymbol{m}_i$ converge

Issues for K-mean algorithm

- Does it find the global optimum of J?
  - No, the nearest local optimum, depending on initialization



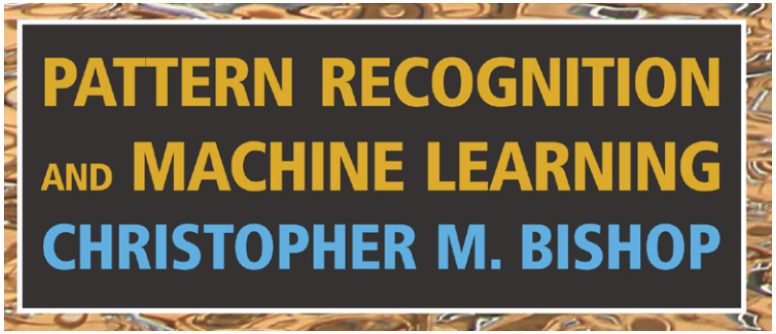
- If Euclidean distance is not good for some data, do we have other choices?
- Can we assign each data point to the clusters probabilistically?
- If K (the total number of clusters) is unknown, can we estimate it from the data?

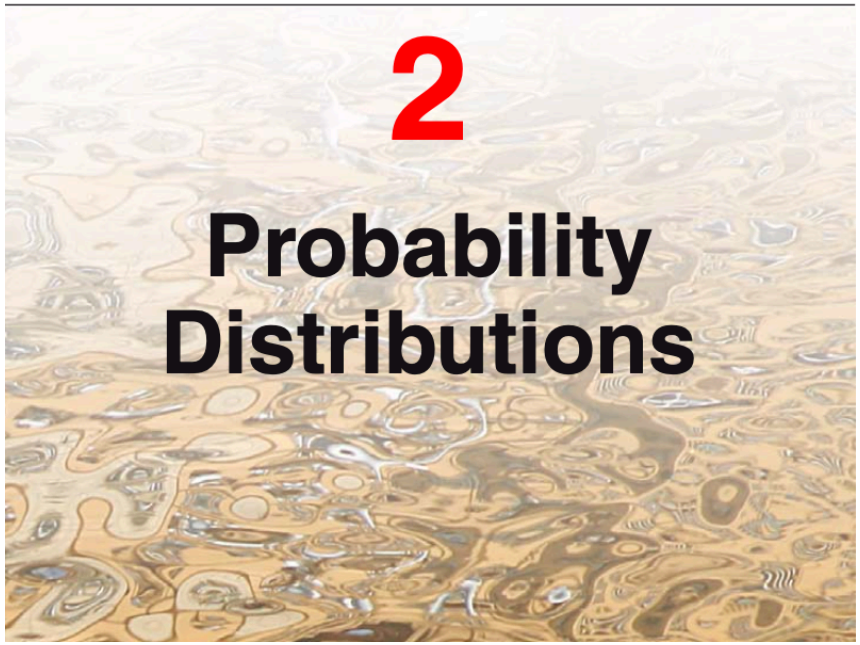
Outline

- Clustering
  - K-mean clustering, hierarchical clustering
- Adaptive learning (online learning)
  - CL, FSCL, RPCL
- Gaussian Mixture Models (GMM)
- Expectation-Maximization (EM) for maximum likelihood

Probability distributions

- Read Bishop's PRML book, Chapter 2





Binary variables

$$x \in \{0, 1\}$$

$$0 \leqslant \mu \leqslant 1$$

- Bernoulli distribution

The probability of x = 1

$$p(x = 1|\mu) = \mu$$

$$p(x = 0|\mu) = 1 - \mu$$

$$\mathrm{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

$$\mathbb{E}[x] = \mu$$

$$\mathrm{var}[x] = \mu(1 - \mu)$$

Binary variables

$x \in \{0, 1\}$

$0 \leqslant \mu \leqslant 1$

- Bernoulli distribution

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$$

Now suppose we have a data set $\mathcal{D} = \{x_1, \ldots, x_N\}$ of observed values of x. We can construct the likelihood function, which is a function of μ, on the assumption that the observations are drawn independently from $p(x|\mu)$, so that

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}. \tag{2.5}$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n) \ln(1-\mu)\}.$$

Multinomial variables

1-of-*K* scheme

- Example $\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}} \quad K = 6$ $\quad x_3 = 1$ $\quad \sum_{k=1}^{K} x_k = 1$
- Multinomial distribution

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} \quad \mu_k \geqslant 0 \quad \sum_k \mu_k = 1$$

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_M)^{\mathrm{T}} = \boldsymbol{\mu}$$

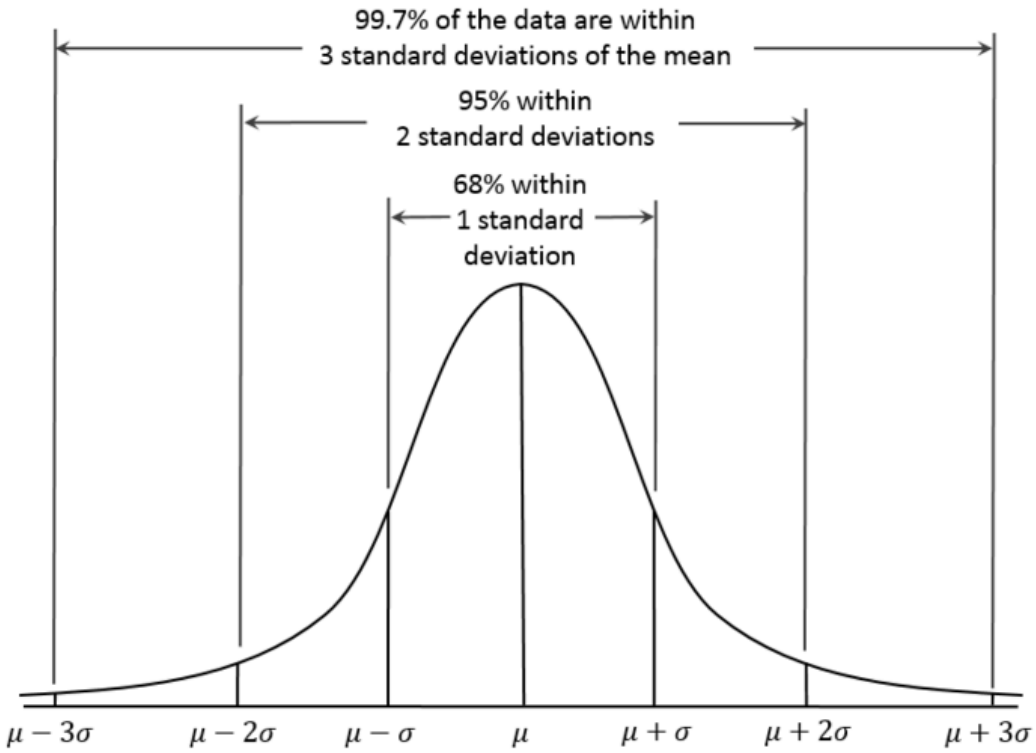
$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\left(\sum_n x_{nk}\right)} = \prod_{k=1}^{K} \mu_k^{m_k} \qquad m_k = \sum_n x_{nk}$$

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

Gaussian distribution

also known as the normal distribution

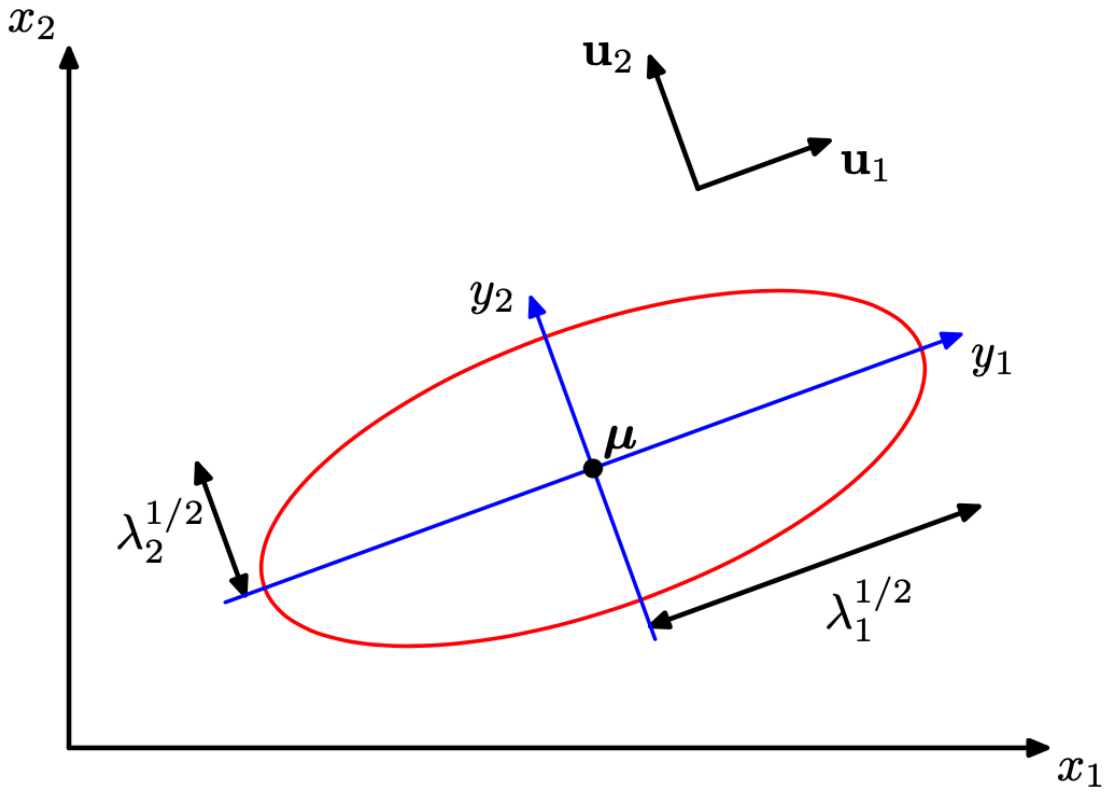
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



Multivariate Gaussian

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

The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space $\mathbf{x} = (x_1, x_2)$ on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix, with corresponding eigenvalues λ_i.



$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

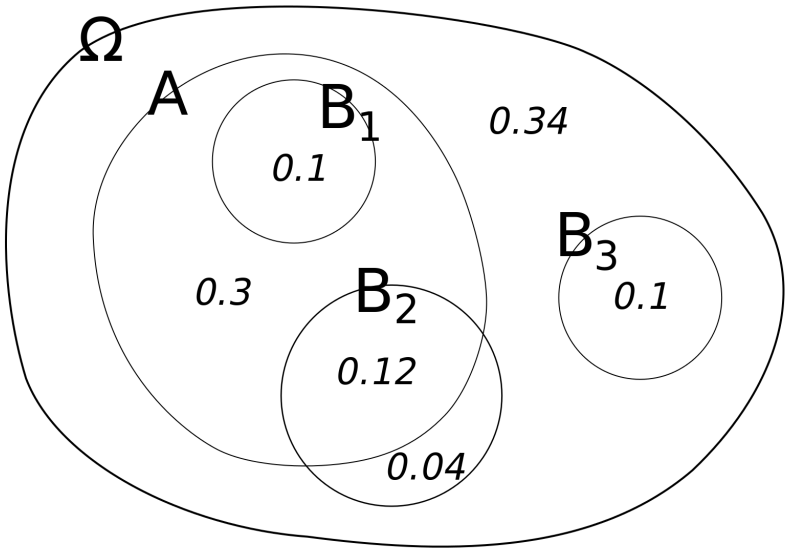
$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] = \boldsymbol{\Sigma}.$$

Conditional probability

From wiki

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},$$



The **unconditional** probability

P(A) = 0.30 + 0.10 + 0.12 = 0.52.

The **conditional** probability

$P(A|B_1) = 1,$

$P(A|B_2) = 0.12 \div (0.12 + 0.04) = 0.75,$

$P(A|B_3) = 0.$

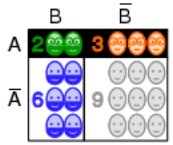
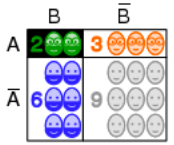
Bayes Theorem

From wiki

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

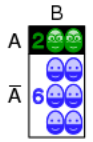
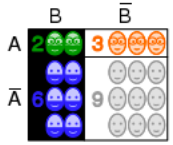
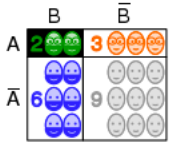
$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

Number of occurences	Beard: B	No beard: $\overline{B}$	sum
Astigmatic: A	2	3	5
Not astigmatic: $\overline{A}$	6	9	15
sum	8	12	20

$$P(B, \text{given } A) \cdot P(A) = P(B|A) \cdot P(A)$$

$$\frac{2}{2+3} \cdot \frac{2+3}{2+3+6+9} = \frac{2}{2+3+6+9}$$

$$P(A, \text{given } B) \cdot P(B) = P(A|B) \cdot P(B)$$

$$\frac{2}{2+6} \cdot \frac{2+6}{2+3+6+9} = \frac{2}{2+3+6+9}$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

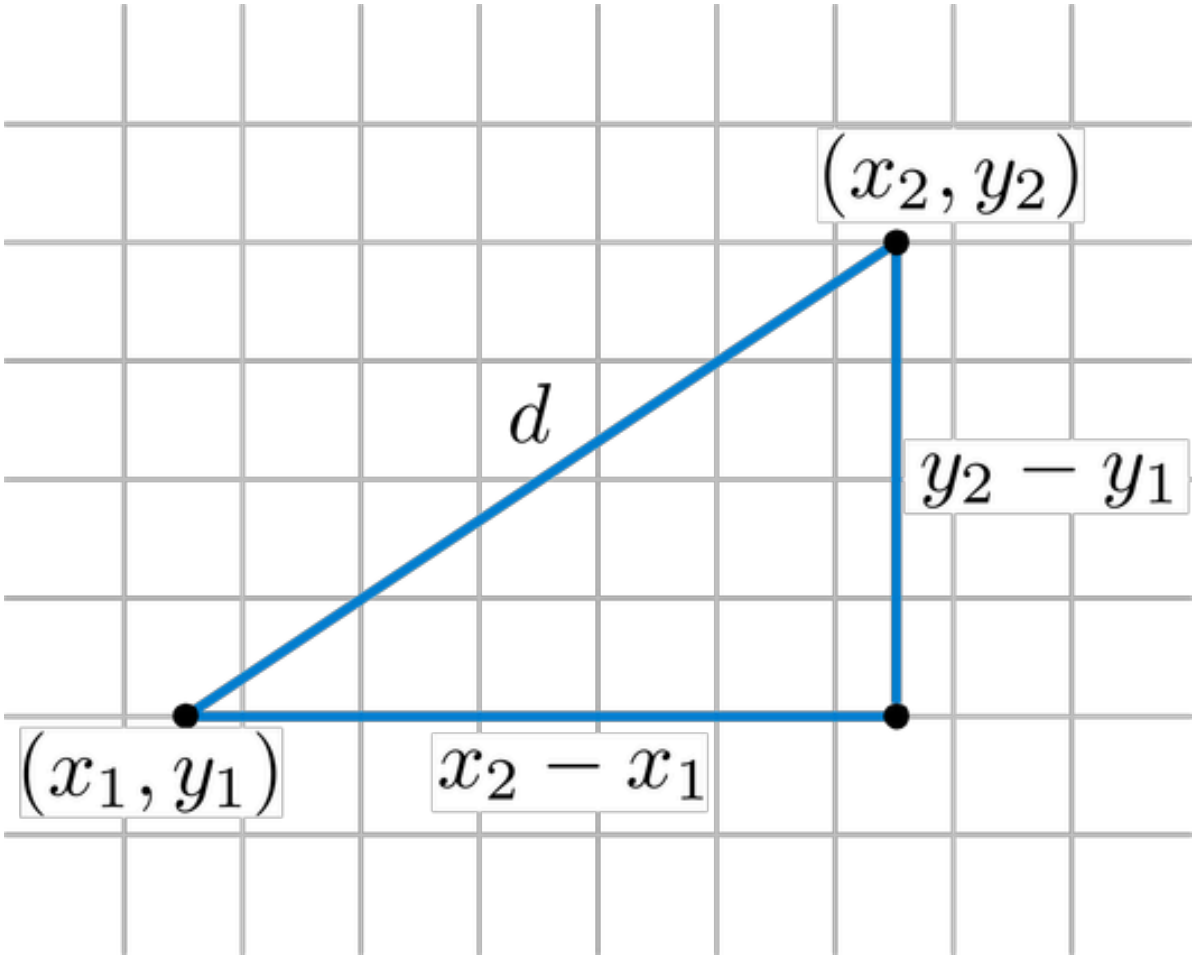
$$\therefore \quad P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

The Rules of Probability

sum rule $$p(X) = \sum_{Y} p(X, Y)$$

product rule $$p(X, Y) = p(Y|X)p(X).$$

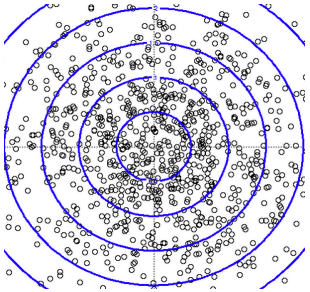
Euclidean Distance



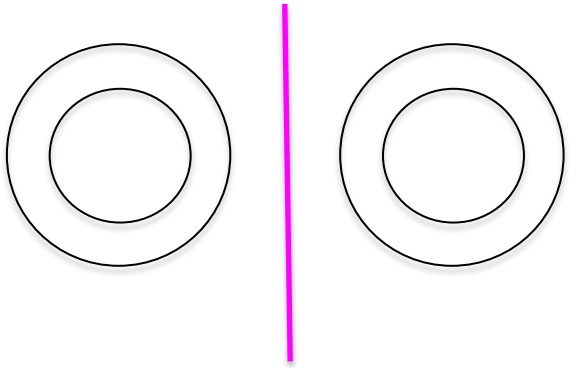
Glossary. (n.d.). Retrieved February 13, 2015, from http://rosalind.info/glossary/euclidean-distance/

Euclidian distance may not be a good measure for some data

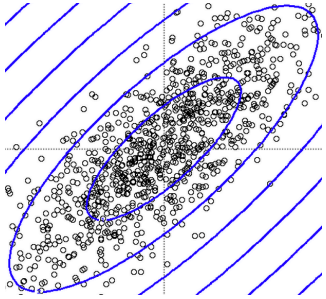
Euclidean distance

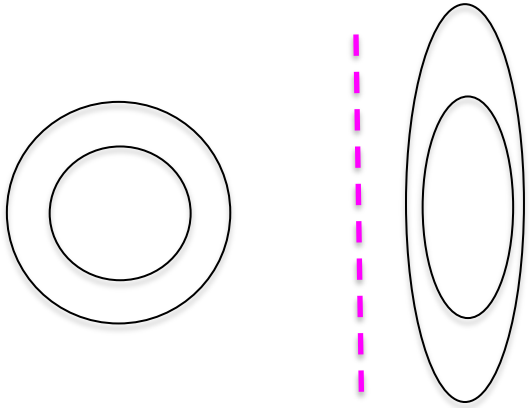


Equal distance line

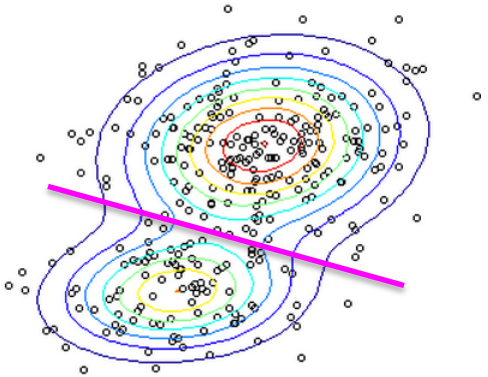


Mahalanobis distance





In general



Distances at different directions could be different!

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \Sigma^{-1} (\vec{x}-\vec{y})}.$$

Σ is the covariance matrix

More Distance Measures

Table 1 Gene expression similarity measures

Manhattan distance (city-block distance, L1 norm)	$d_{fg} = \sum_c \lvert e_{fc} - e_{gc} \rvert$
Euclidean distance (L2 norm)	$d_{fg} = \sqrt{\sum_c (e_{fc} - e_{gc})^2}$
Mahalanobis distance	$d_{fg} = (e_f - e_g)' \Sigma^{-1} (e_f - e_g)$, where Σ is the (full or within-cluster) covariance matrix of the data
Pearson correlation (centered correlation)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c (e_{fc} - \bar{e}_f)(e_{gc} - \bar{e}_g)}{\sqrt{\sum_c (e_{fc} - \bar{e}_f)^2 \sum_c (e_{gc} - \bar{e}_g)^2}}$
Uncentered correlation (angular separation, cosine angle)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c e_{fc} e_{gc}}{\sqrt{\sum_c e_{fc}^2 \sum_c e_{gc}^2}}$
Spellman rank correlation	As Pearson correlation, but replace $\boldsymbol{e_{gc}}$ with the rank of $\boldsymbol{e_{gc}}$ within the expression values of gene g across all conditions $\boldsymbol{c = 1 \ldots C}$
Absolute or squared correlation	$d_{fg} = 1 - \lvert r_{fg} \rvert$ or $d_{fg} = 1 - r_{fg}^2$

d_{fg}, distance between expression patterns for genes f and g. e_{gc}, expression level of gene g under condition c.

D'haeseleer, P. (2005). How does gene expression clustering work? Nat Biotech *23*, 1499–1501.

From distance to probability

distance → likely

$$\| x - \mu \|^2 \longrightarrow \exp\{-\lambda \| x - \mu \|^2\}$$

"The closer, the more likely."

Sum or integral to be one

Probability

$$\frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

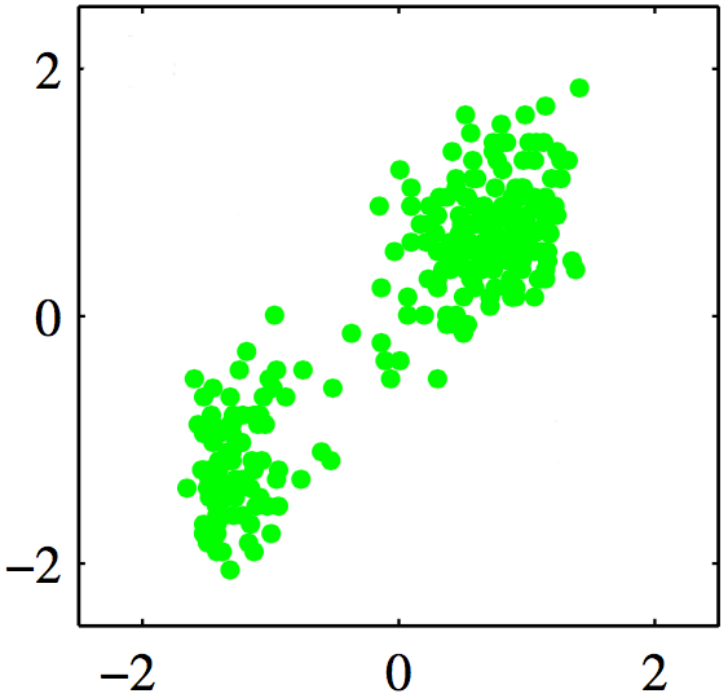
Gaussian distribution with the Mahalanobis distance

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}.$$

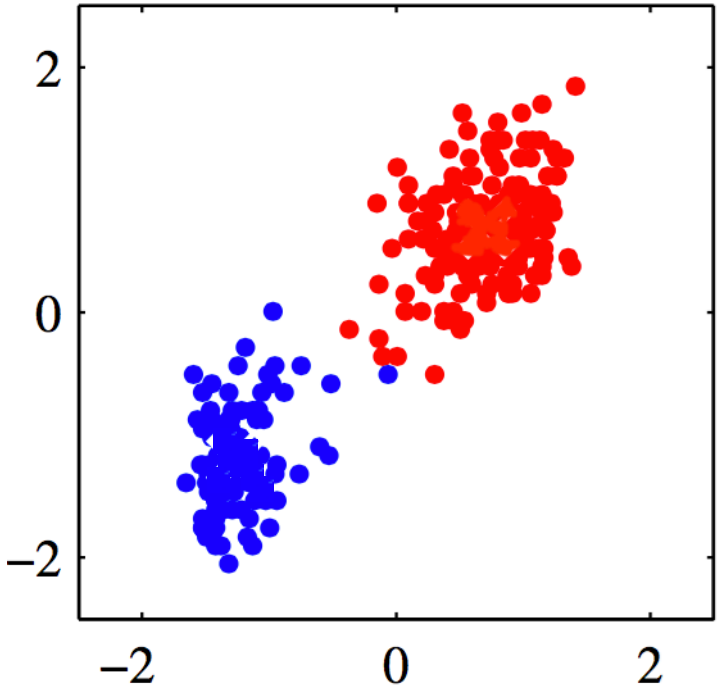
It is more powerful to consider everything in probability framework!

Review the clustering problem again

We have the following data:

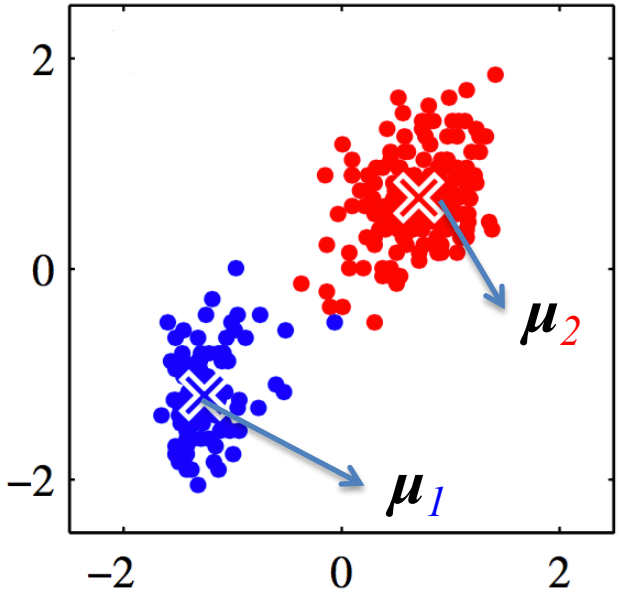


We want to cluster the data into two clusters (red and blue)

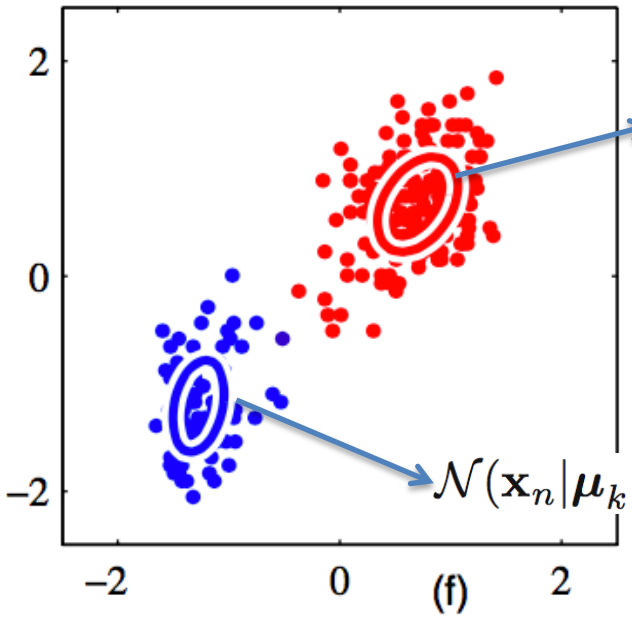


Instead if using $\{\mu_1, \mu_2\}$, each cluster is represented as a Gaussian distribution

K-means



Gaussian Mixture Model (GMM)

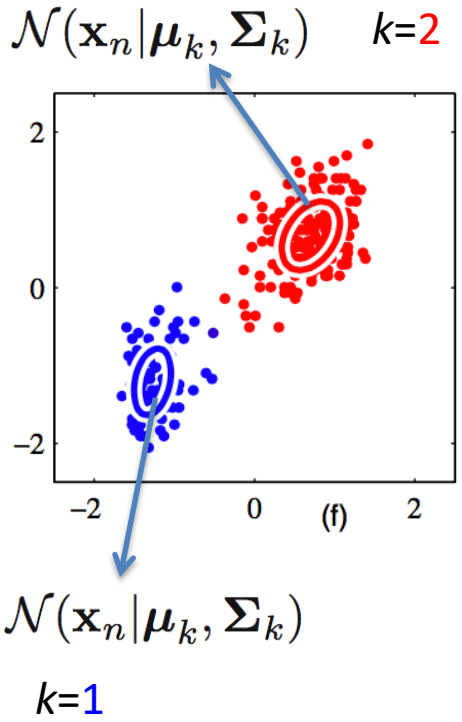


k=2

k=1

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

Gaussian Mixture Model (GMM)



We use $z_k = 1$ to indicate a point $\mathbf{x}$ belongs to cluster k

$$\mathbf{z} = (z_1, \ldots, z_K) \qquad z_k \in \{0, 1\} \qquad \sum_k z_k = 1$$

Assume the points in the same cluster follow a **Gaussian distribution**

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

A mixing weight for each cluster:

$$p(z_k = 1) = \pi_k \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

prior probability of point belonging to a cluster

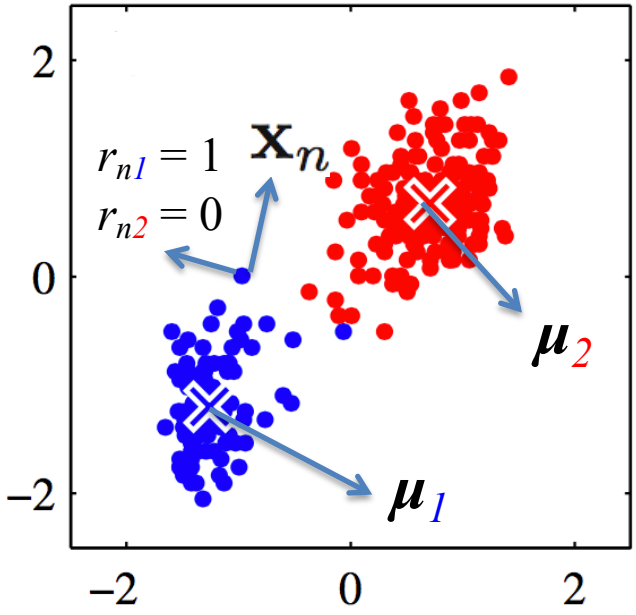
So, we get a distribution for the data point $\mathbf{x}$:

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

From minimizing sum of square distances to finding maximum likelihood

minimize

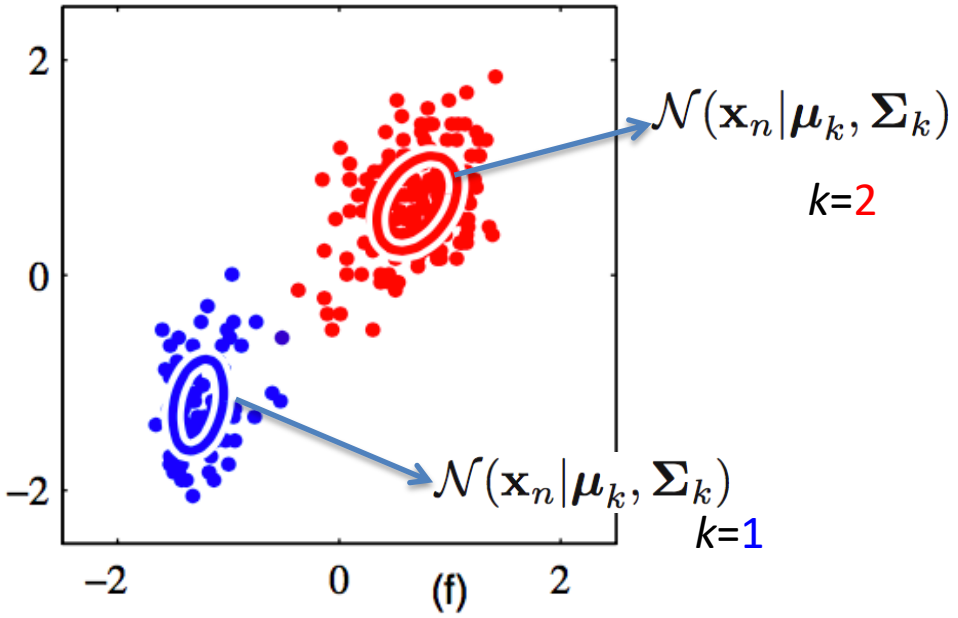
$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$



maximize likelihood

$$p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$$

$$X = \{x_1, \ldots, x_N\}$$
$$\pi = \{\pi_1, \ldots, \pi_K\}$$
$$\mu = \{\mu_1, \ldots, \mu_K\}$$
$$\Sigma = \{\Sigma_1, \ldots, \Sigma_K\}$$



Remember: **The closer the distance, the more likely the probability.**

Maximum likelihood

Given a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$ in which the observations $\{\mathbf{x}_n\}$ are assumed to be drawn independently from a multivariate Gaussian distribution, we can estimate the parameters of the distribution by maximum likelihood. The log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

Maximizing the log-likelihood function:

$$\frac{\partial}{\partial \boldsymbol{\mu}}\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0 \quad \longrightarrow \quad \boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

Similarly we get $\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$

$\boldsymbol{\mu}_{\mathrm{ML}}$ and $\boldsymbol{\Sigma}_{\mathrm{ML}}$ are the maximum likelihood estimates of the mean and the co-variance matrix.

Thank you!

Matrix-cook-book

$$
\begin{aligned}
\partial \mathbf{A} &= 0 && (\mathbf{A} \text{ is a constant}) \\
\partial(\alpha \mathbf{X}) &= \alpha \partial \mathbf{X} \\
\partial(\mathbf{X} + \mathbf{Y}) &= \partial \mathbf{X} + \partial \mathbf{Y} \\
\partial(\mathrm{Tr}(\mathbf{X})) &= \mathrm{Tr}(\partial \mathbf{X}) \\
\partial(\mathbf{X}\mathbf{Y}) &= (\partial \mathbf{X})\mathbf{Y} + \mathbf{X}(\partial \mathbf{Y}) \\
\partial(\mathbf{X} \circ \mathbf{Y}) &= (\partial \mathbf{X}) \circ \mathbf{Y} + \mathbf{X} \circ (\partial \mathbf{Y}) \\
\partial(\mathbf{X} \otimes \mathbf{Y}) &= (\partial \mathbf{X}) \otimes \mathbf{Y} + \mathbf{X} \otimes (\partial \mathbf{Y}) \\
\partial(\mathbf{X}^{-1}) &= -\mathbf{X}^{-1}(\partial \mathbf{X})\mathbf{X}^{-1} \\
\partial(\det(\mathbf{X})) &= \det(\mathbf{X})\mathrm{Tr}(\mathbf{X}^{-1}\partial \mathbf{X}) \\
\partial(\ln(\det(\mathbf{X}))) &= \mathrm{Tr}(\mathbf{X}^{-1}\partial \mathbf{X}) \\
\partial \mathbf{X}^T &= (\partial \mathbf{X})^T \\
\partial \mathbf{X}^H &= (\partial \mathbf{X})^H
\end{aligned}
$$