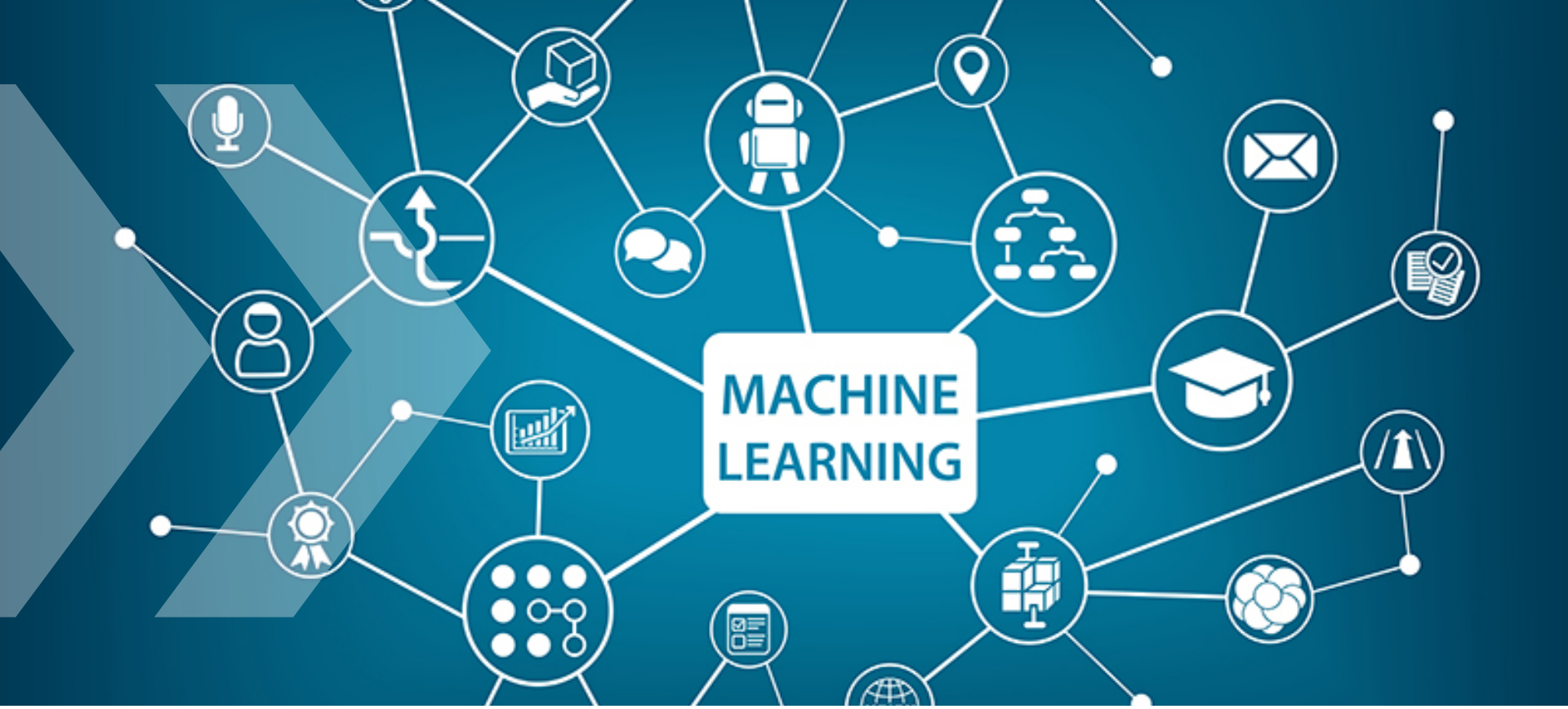
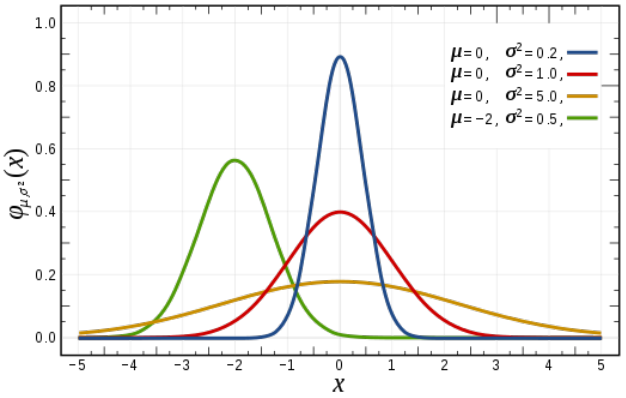


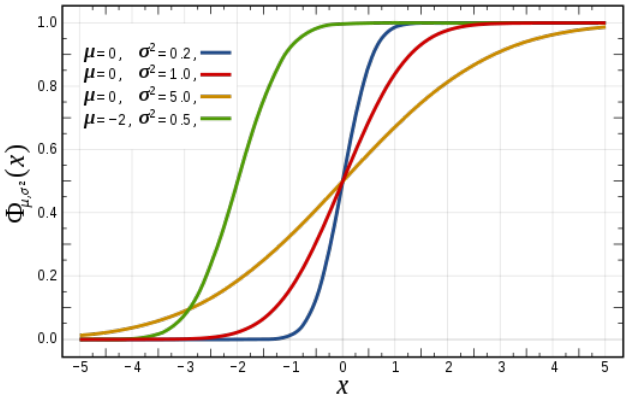
CS489/698: Intro to ML

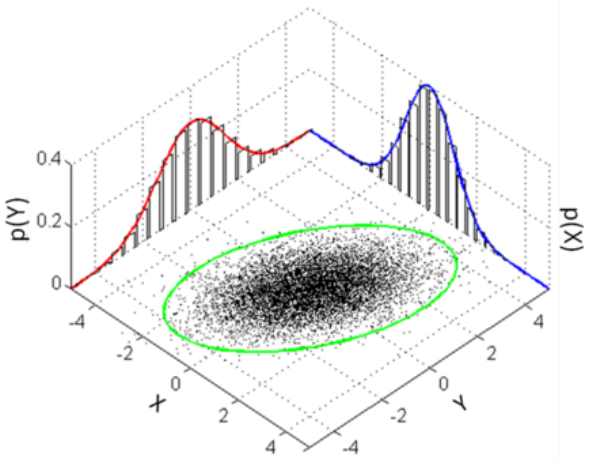
Lecture 10: Mixture of Gaussians

UNIVERSITY OF WATERLOO

Recap: Gaussian distribution



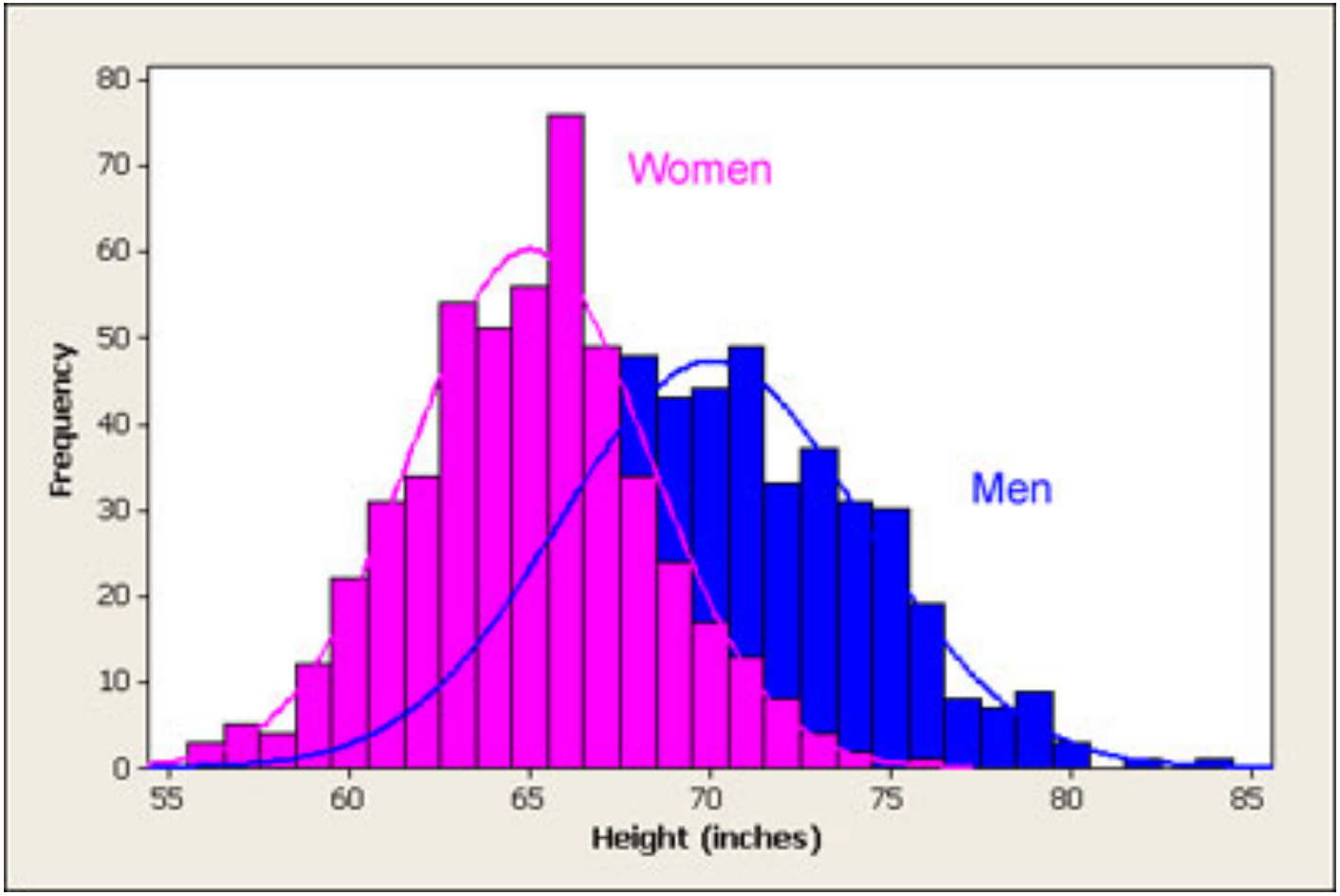




$$p(\mathbf{x}) = (2\pi)^{-d/2}|S|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

UNIVERSITY OF WATERLOO

Multi-modality



UNIVERSITY OF WATERLOO

Mixture models

$p(x,z) = p(z)\ p(x|z)$

K → # of components

$$p(x|\theta) = \sum_{k=1} p(z=k)\, p(x|z=k,\theta)$$

π_k : mixing distr.

$p_k(x|\theta)$: k-th component distr.

θ : parameters

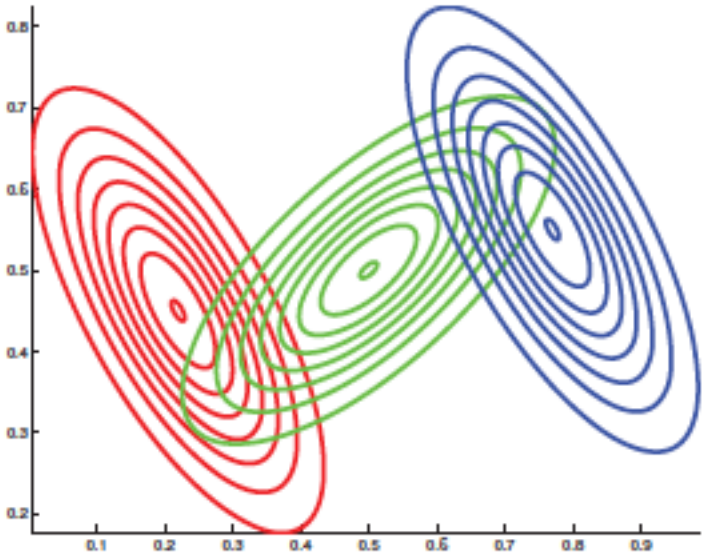
$$\pi_k \geq 0, \sum_k \pi_k = 1$$

- Where did we see a similar idea?

Example: Gaussian Mixture Models

$(\pi, \mu_1, S_1, \mu_2, S_2, \cdots, \mu_K, S_K)$

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



(a)

(b)

Universality

Theorem. GMM with sufficiently many components can approximate any probability density function on R^d.

- How many is many?
- Nothing special about Gaussian here, except computationally (later).

Example: Mixture of Experts

$$\frac{\exp(\mathbf{x}^\top \mathbf{v}_k)}{\sum_{c=1}^{K} \exp(\mathbf{x}^\top \mathbf{v}_c)}$$

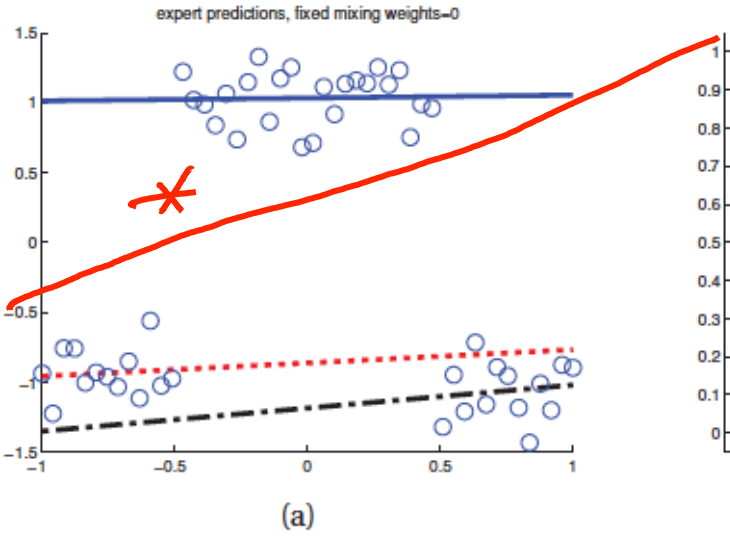
Where did we see a similar idea?

$$\mathcal{N}(y|\mathbf{w}_k^\top \mathbf{x}, \sigma_k^2)$$

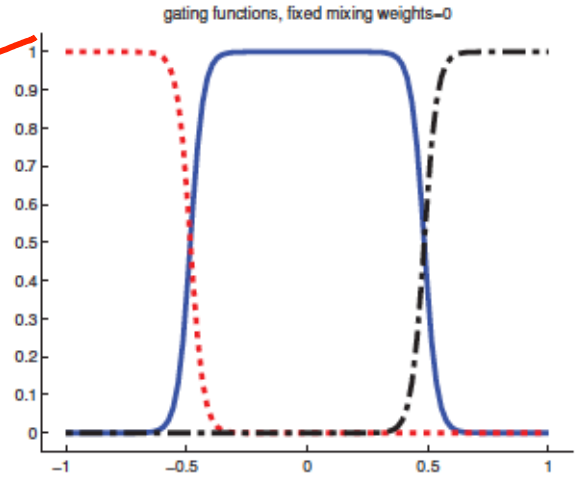
K mixing distr.

k-th component distr.

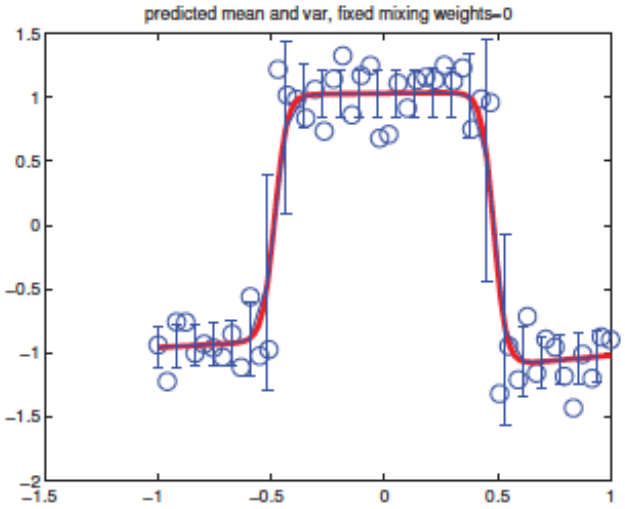
$$p(y|\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} p(z = k|\mathbf{x}, \boldsymbol{\theta}) p(y|\mathbf{x}, z = k, \boldsymbol{\theta})$$



(a)



(b)



(c)

Inference problem

$$p(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(z=k)p(x|z=k,\boldsymbol{\theta})$$

latent (unobserved)

- Given iid sample $X_1, X_2, \ldots, X_n$ from $p(x|\theta)$
- Need to estimate θ
- Maximum likelihood is NP-hard…

Soft clustering

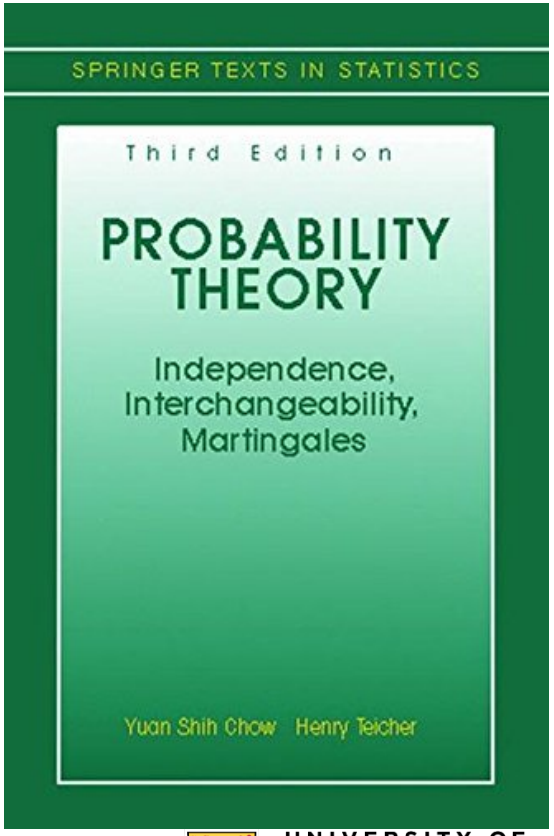
$$p(z = k|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(z = k|\boldsymbol{\theta})p(\mathbf{x}|z = k, \boldsymbol{\theta})}{\sum_{c=1}^{K} p(z = c|\boldsymbol{\theta})p(\mathbf{x}|z = c, \boldsymbol{\theta})}$$

(Stauffer & Grimson, CVPR'98)

Bigger issue: identifiability?

$$p(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(z=k)p(x|z=k,\boldsymbol{\theta})$$

- Is this factorization even unique?
- Yes, for GMMs!



Variational form of Max Likelihood

$$\min_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) := \sum_{i=1}^{n} -\log p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{i=1}^{n} -\log \sum_{z_i} p(\mathbf{x}_i, z_i|\boldsymbol{\theta})$$

$$= \min_{q_i(z_i)\geq 0, \sum_{z_i} q_i(z_i)=1} -\sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta}) + \sum_{z_i} q_i(z_i) \log q_i(z_i)$$

neg. entropy

$$\min_{q_i(z_i)} \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left[\sum_{z_i} q_i(z_i) \log q_i(z_i) - \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta}) \right]$$

KL divergence

$\neq KL(q\|p)$

$$\mathrm{KL}(\mathbf{p}\|\mathbf{q}) := \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \geq 0$$

- Both **p** and **q** are nonnegative and sum to 1
- Equality holds iff **p** == **q**
- Measures difference between distributions; asymmetric

Jensen's inequality
E(log(X)) <= log(E(X))

The EM algorithm

$$\min_{q_i(z_i)} \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left[\sum_{z_i} q_i(z_i) \log q_i(z_i) - \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i | \boldsymbol{\theta}) \right]$$

- Fix q, solve θ

$$\min_{\boldsymbol{\theta}} - \sum_{i=1}^{n} \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i | \boldsymbol{\theta})$$

often closed-form

- Fix θ, solve q

$$\min_{q_i(z_i) \geq 0, \sum_{z_i} q_i(z_i) = 1} - \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i | \boldsymbol{\theta}) + \sum_{z_i} q_i(z_i) \log q_i(z_i)$$

$p(z_i|x_i;\theta) \cdot p(x_i|\theta)$

$$q_i(z_i) = p(z_i | \mathbf{x}_i, \boldsymbol{\theta})$$

$KL(q_i \| p(z_i|x_i,\theta))$

$-\log(p(x_i|\theta))$

EM for GMM: step 1

$q_i(z_i = k) = r_{ik} \geq 0$

$\sum_{k=1}^{K} r_{ik} = 1$

$\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, S_k)$

$$\min_{r_{ik} \geq 0, \sum_k r_{ik} = 1} \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} r_{ik} \log r_{ik} - \sum_{k=1}^{K} r_{ik} \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta}) \right]$$

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} [-\log \pi_k + \frac{1}{2} \log |S_k| + \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top S_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)]$$

$$\pi_k = \frac{\sum_i r_{ik}}{n} \quad \sum_k \sum_i r_{ik}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} r_{ik} \mathbf{x}_i}{\sum_{i=1}^{n} r_{ik}}$$

$$S_k = \frac{\sum_{i=1}^{n} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top}{\sum_{i=1}^{n} r_{ik}} = \frac{\sum_{i=1}^{n} r_{ik} \mathbf{x}_i \mathbf{x}_i^\top}{\sum_{i=1}^{n} r_{ik}} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top$$

Just in case

or if you don't understand

$$\frac{\partial}{\partial \mu_k} = \sum_{i=1}^{n} r_{ik} \Sigma_k^{-1} (x_i - \mu_k) = 0$$

$$\Downarrow \text{ multiplying } \Sigma_k$$

$$\sum_{i=1}^{n} r_{ik} (x_i - \mu_k) = 0$$

$$\Downarrow$$

$$\mu_k = \frac{\sum_{i=1}^{n} r_{ik} x_i}{\sum_{i=1}^{n} r_{ik}}$$

$$\frac{\partial}{\partial \Sigma_k} = \sum_{i=1}^{n} r_{ik} \left[\frac{1}{2} \Sigma_k^{-1} - \frac{1}{2} \Sigma_k^{-1} (x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1} \right] = 0$$

$$\Downarrow \text{ multiplying } \Sigma_k \text{ twice}$$

$$\left(\sum_{i=1}^{n} r_{ik} \right) \cdot \Sigma_k = \sum_{i=1}^{n} r_{ik} (x_i - \mu_k)(x_i - \mu_k)^T$$

$$\Downarrow$$

$$\Sigma_k = \frac{\sum_{i=1}^{n} r_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} r_{ik}}$$

multiplying constant does not change min

adding constant does not change min

$$\min_{\substack{\pi_k \geq 0 \\ \sum_k \pi_k = 1}} \sum_{i=1}^{n} \sum_{k=1}^{k} r_{ik} \log \frac{1}{\pi_k} = \sum_{k=1}^{k} \left(\sum_{i=1}^{n} r_{ik} \right) \log \frac{1}{\pi_k} \equiv \min_{\substack{\pi_k \geq 0 \\ \sum_k \pi_k = 1}} \frac{1}{n} \sum_{k=1}^{k} r_k \log \frac{1}{\pi_k} + \sum_{k=1}^{k} \frac{r_k}{n} \log \frac{r_k}{n}$$

$$= \min_{\substack{\pi_k > 0 \\ \sum_k \pi_k = 1}} KL\left(\frac{r}{n} \| \pi \right) \Rightarrow \pi_k = \frac{r_k}{n}$$

where r_k is defined as $\sum_{i=1}^{n} r_{ik}$

EM for GMM: step 2

$$\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, S_k)$$

$$\min_{r_{ik}\geq 0, \sum_k r_{ik}=1} \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} r_{ik} \log r_{ik} - \sum_{k=1}^{K} r_{ik} \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta}) \right]$$

$$r_{ik} = p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) \longrightarrow \text{posterior}$$

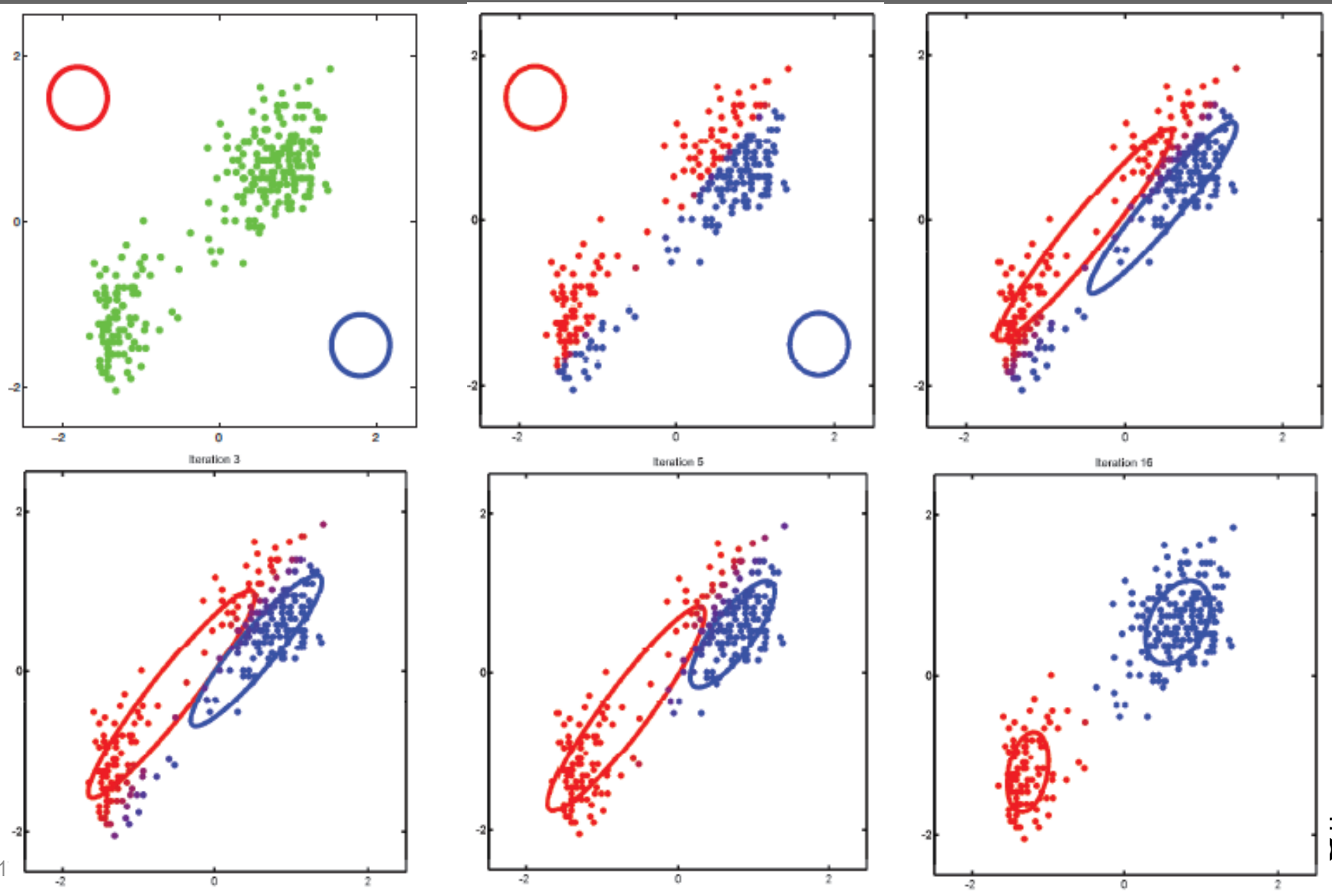
$$\propto p(z_i = k) \cdot p(\mathbf{x}_i|z_i = k, \boldsymbol{\theta})$$

prior ← $p(z_i = k)$; $p(\mathbf{x}_i|z_i = k, \boldsymbol{\theta})$ → likelihood

$$= \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, S_k)$$

$$r_{ik} = \frac{\pi_k |S_k|^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top S_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k))}{\sum_{c=1}^{K} \pi_c |S_c|^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top S_c^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_c))}$$

Example



Other uses of EM

- Simplify computation
  - t-distribution as a Gaussian scale-mixture
- Missing data