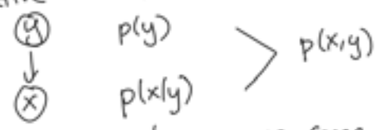
Last time: clustering. intuitive.
- how do you measure a "good" clustering?
- how to handle: unclear points?
  missing data / missing dims? $d(x, x')$

Today: probabilistic clustering ⇒ mixture models

View: 

[note: looks a lot like classification w/ generative models:



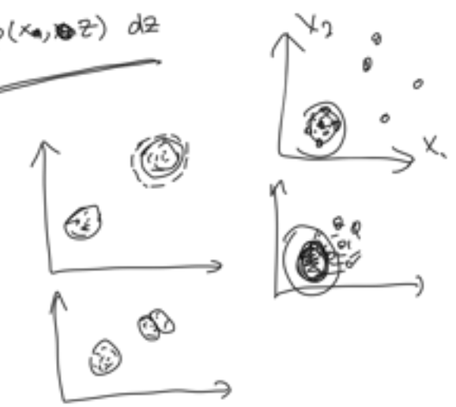
now: only given x's, assume some hidden z

Gen. Process:

$z_n \sim \pi$

$x_n \sim p(x_n | z_n, W)$

$p(x_n, z_n) \Rightarrow p(x) = \int_z p(x, z)\, dz$



Notes:

1) This is density estimation

2) We can predict which mixture is likely given x, W

$p(z | x, W) \propto p(x | z, W) p(z)$

3) e.g. masses of car

"source separation"

Let's make this specific: Gaussian mixture model:

$p(z_n = k) = \pi_k$

$p(x | z = k) = \mathcal{N}(\mu_k, \Sigma_k)$

- Complete data log likelihood: $p(\{x, z\} | W, \cdot)$

$$\log \mathcal{L}(W) = \sum_n \log p(x_n, z_n | W)$$
$$\underset{\text{complete}}{=} \sum_n \log p(x_n | z_n, W) + \log p(z | W)$$
$$= \sum_n \log \mathcal{N}(\mu_{z_n}, \Sigma_{z_n}) + \log \pi_{z_n}$$
$$= \sum_n \sum_k z_{nk} \log \mathcal{N}(x_n | \mu_k, \Sigma_k) + z_{nk} \log \pi_k$$

We can solve for the MLE analytically: set mixture μ, Σ to

$\pi_k = \frac{N_k}{N}$ $\quad \mu_k = \text{mean}(x_n \text{ s.t. } z_{nk}=k)$ $\quad \Sigma_k = \text{var}(x_n \text{ s.t. } z_{nk}=k)$ } empirical values.

Problem: we don't have z! [aside: search over the z's is combinatorial]

- Let's write down $p(x|W)$

$$\mathcal{L}_{marg}(W) = \sum_n \log p(x_n|W) = \sum_n \log \left[\sum_k \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)\right]$$ } model $p(x)$

no analytic solution
gradients are messy...

- Can we bound the marginal? [new idea]
  ↳ lower bound.

Turns out, the expected complete data log likelihood is a lower bound for the marginal likelihood $\max \int p(x,z)\,dz$ → we subbed in π_k's.

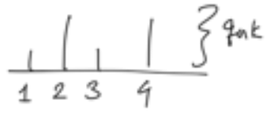
$$\mathcal{L}_{marg} = \log \int p(x,z)\,dz$$

instead

$$E[\mathcal{L}_{complete}] = \int \log p(x,z)\,dz$$

$$E_z[\mathcal{L}_{complete}(W)] = E_z[\log p(\{x_n, z_n\}|W)] = E_z\left[\sum_{n,k} z_{nk} \log \pi_k + z_{nk} \log \mathcal{N}(x_n|\mu_k, \Sigma_k)\right]$$

Let q_{nk} be a distribution over z_n
q_n has K elements, sum to 1.



$$= \sum_{n,k} q_{nk} \log \pi_k + q_{nk} \log \mathcal{N}(x_n|\mu_k, \Sigma_k)$$

- Apply block coordinate ascent [q] · [π, μ, Σ]
- This is nice for optimizing w.r.t. π, μ, Σ → corresponds to a weighted cluster assignment.

$$\pi_k = \frac{\sum_n q_{nk}}{N}$$

$$\mu_k: \frac{\partial \mathcal{L}_n}{\partial \mu_k} = q_{nk}\left(-\tfrac{1}{2}\right)\left(2\Sigma_k^{-1}\mu_k - 2\Sigma_k^{-1}x_n\right)$$

↳ gives us $\hat{\mu}_k = \frac{\sum_n q_{nk} x_n}{\sum_n q_{nk}}$

$$\Sigma_k = \frac{1}{\sum_n q_{nk}} \sum q_{nk}(x_n - \hat{\mu}_k)(x_n - \hat{\mu}_k)^T$$

↳ nice! [often the case w/in exponential families]

Name: M-step [Maximizing expected complete D.L. w.r.t. global params].

- 2nd part of the block update: q_{nk}

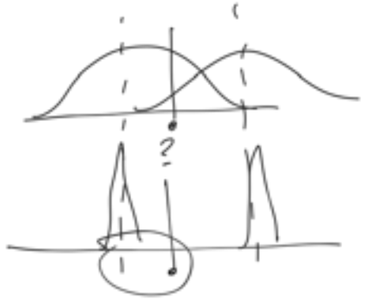
best choice of $q_{nk} = p(z_n \mid x_n, W)$

$$\propto \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$$

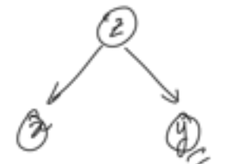
E-step Expectations over z [locals].

Back to properties of EM:

- general: E-step: local hiddens
  M-step: opt. globals
- ⇒ monotonically improves (lower bound [coding check]
- converges to local optima (random restarts)
- somewhat sensitive to init: $\pi_k = 1/K$ & all μ_k, Σ_k equal (saddle)
- variations MoG: Σ_k be full, diagonal, isotropic... shared Σ's. (versatile)
- MoG w/ small variance → K-means.



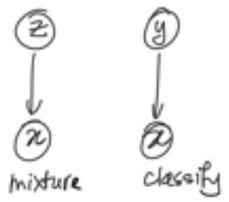
Concept Check:



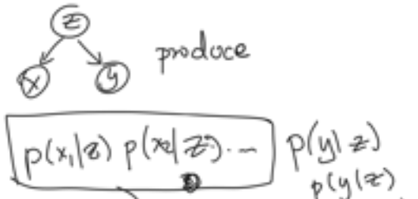
"supervised mixture" model

interested in predicting y

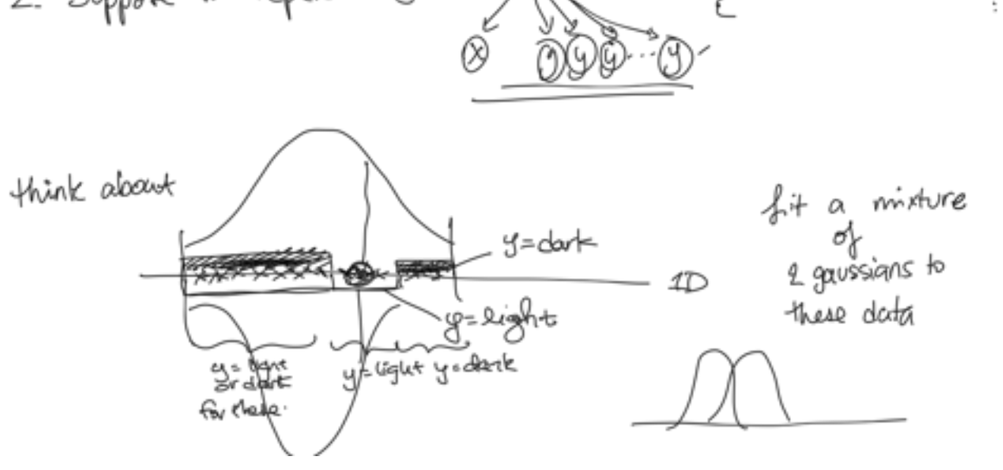
also want a generative model for x



1. Suppose $D_x \gg D_y$. Will mixture model different $\{\mu_k, \Sigma_k\}$ than  ?



2. Suppose we replicate y:



Mixture Models