EM: can we use latent variables to devise algo?

- We will consider this during lecture to be an MLE setting, allow it will be fine as MAP (just would have prior term ever finding around).

We will $\theta^{(0)}$ randomly. Run til "convergence"

i.e. iterate (i.e.)

**E step:** we have cluster assignments $C_k (c \in 1, \ldots, K)$

define a distribution over x and $\theta$:

$$Q^{(i)}(c_k) = P(c_k | x, \theta^{(i-1)})$$

$$= \left[ \frac{1}{\sum_{c_k} q^{(i)}(c_k)} \right]$$

maximize w.r.t. $\theta^{(i)}$.

This gives you a way to have a distribution to take expectation w.r.t.

**M step:**

for $Q_{c_k}^{(i)}$ fixed, maximize a "weighted" likelihood:

$$\theta^{(i)} = \arg\max_{\theta} \sum_{c_k} Q^{(i)}(c_k) \log P(x, c_k | \theta^{(i)})$$

looks like a likelihood

- may have to multiply adjust this

*eq.: repeat what 1st step is doing*

example for GMM:

- for every $k \in [K]$,

  $$Q^{(i)}_{c_k}(c_t = k) = P(c_t = k | x, \theta^{(i-1)})$$

- for $c_t = k$,

  $$= \frac{P(x_t, c_t = k | \theta^{(i-1)})}{\sum_{c_t} P(x_t, c_t = k | \theta^{(i-1)})}$$

  $$\propto \frac{P(x_t, c_t = k, \theta^{(i-1)})}{P(c_t = k | \theta^{(i-1)})}$$

- this is $\pi_k$, our mixture guess (marginal prob)

- you could expand this out

  this is easy in them

  - now, for $\theta^{(i)}$:

    $$\theta^{(i)} = \arg\max_{\theta} \sum_{c_t} Q_{c_t}(c_t = k) \log P(x_t, c_t = k | \theta)$$

    (note: if you take derivative, the $\frac{1}{\theta}$ will go inside the sum.)

  - question: are the $\theta$'s like what we saw last time??

    - yes, by making the $Q^{(i)}$'s to be "indicator funs".

    - for GMM: in slow dumping, average:

      $$\frac{1}{\sum_{c_t} \theta} \sum_{c_t} \theta Q^{(i)}(c_t) (\log N(x_t, \mu_k, \Sigma_k) + \log \pi_k)$$
and the solution will be:

\[ \pi_{k}^{(i)} = \frac{\sum_{t} Q_{t}^{(i)}(k)}{\sum_{t} Q_{t}(k)} \]

will post the proof for $\pi$, but let's try $\bar{X}$.

Actually, let's talk about $\bar{X}$, it adds some, but go back to derivation of $\bar{X}$.

Why does this work? Intuition:

Entry at never decreases log-likelihood, and usual does $\Pi$-steps.

- to show $\log L(\theta^{(m)}) \geq \log L(\theta^{(i)})$.

Steps: insert latent vars, use Jensen's inequality, massage.

Jensen's $\log$ is concave, with

\[ \sum \log P(x_t | \theta) = \sum \log \left[ \sum_{t} P(x_t, \zeta_t = \lambda | \theta) \right] \]

Consider:

and let's call it $H(k)$, so to set

\[ \frac{1}{\sum_{t} E_{\theta^{(i)}}(k) \log \hat{Q}_{t}(k) \log \tilde{Q}_{t}(k)} \]

for the $i^{th}$ iteration, we're picking the $\theta$ that maximizes this term.

\[ \theta = \text{argmax} \theta \sum_{t} Q_{t}^{(i)}(k) \log Q_{t}^{(i)}(k) \]

\[ \sum_{t} \sum_{k} Q_{t}^{(i)}(k) \left[ \log P(x_t, \zeta_t = \lambda | \theta^{(i)}) \right] - \sum_{t} \sum_{k} Q_{t}^{(i)}(k) \log Q_{t}(k) \]
So, then, why did we do the expectation step?

This comes in handy due:

\[
\sum_{i} \sum_{\theta} Q_{e}^{(m)}(\theta) \log \left( \frac{P(x_{i}, \theta) \theta^{(i)}}{Q_{e}^{(m)}(\theta)} \right) = \sum_{i} \sum_{\theta} Q_{e}^{(m)}(\theta) \log \frac{P(x_{i} | \theta^{(i)})}{Q_{e}^{(m)}(\theta)}
\]

\[
= \frac{1}{n} \sum_{i} \sum_{\theta} Q_{e}^{(m)}(\theta) \log P(x_{i} | \theta^{(i)})
\]

So, \[
= \sum \log P(x_{i} | \theta^{(i)})
\]

Can hit a stationary point.

Can initialize: hillclimb the hidden variables' values.

This isn't specific to cluster assignments.

K-means: a hard version (hard assignment) on Q's. (and assume spherical clusters)

Soft k-means: still assume spherical Gaussians.