3/24/15
EM

EM q: can we use latent variables to devise algs?
- will consider this during lecture to be in an MLE setting, altho would be fine as MAP (just would have prior term floating around).

we init $\theta^{(0)}$ randomly. Run til "convergence"
i = iteration count.

E step: we have cluster assignments c_t $(c_t \in 1, \dots, k)$
define a distribution wrt each t:

$$Q_{(E)}^{(i)}(c_t) = P(c_t \mid x_t, \theta^{(i-1)})$$
$$= \begin{bmatrix} .3 \\ .7 \\ 0 \end{bmatrix} \begin{matrix} \text{lemons} \\ \text{oat} \\ \dots \end{matrix}$$

. This gives you a way to have a distribution to take expectation wrt.

maximization step:
for $Q_t^{(i)}(c_t)$ fixed, maximize a "weighted" likelihood:

$$\theta^{(i)} = \operatorname{argmax}_\theta \sum_t \sum_{c_t} \underbrace{Q_t^{(i)}(c_t)}_{\text{weight}} \log P(x_t, c_t \mid \theta)$$

looks like a likelihood

<may have to mentally ~~adjust the~~ adjust the i's to $i+1$'s>

<q: repeat what 1st step is doing>

example for GMM: for every $k \in [k]$,

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

<typo on slide>

$$\text{gaussian}(x_t; \mu_k^{(i-1)}, \Sigma_k^{(i-1)})$$

↳ you could expand this out.

this is π_k, our mixture guess (marginal prob)

this is easy (- in theory).

- now, for the M step: not similarity to last time:

$$\theta^{(i)} = \operatorname{argmax}_{\theta \in \Theta} \sum_t \sum_k Q_t^{(i)}(k) \log P(x_t, c_t = k \mid \theta)$$

(note: if you take derivatives, the $\frac{\partial}{\partial}$ will go inside the sum.)

Question: are the Q's like what we saw last time??
Yes, by making the $Q_t^{(i)}$'s to be indicator fns.

↳ for GMM, given above derivs: $\operatorname{argmax}_{\pi, \mu, \Sigma} : \sum_t \sum_{c_t=1}^{k} Q_t^{(i)}(k) \left(\log \text{Gauss}(x_t; \mu, \Sigma) + \log \pi_k \right)$.

and the sol'n will come out to: for each k:

$$\mu_k^{(i)} = \frac{\sum_t Q_t^{(i)}(k) x_i}{\sum_t Q_t^{(i)}(k)} \qquad \Sigma_k^{(i)} = \frac{\sum_t Q_t^{(i)}(k)(x_t - \mu_k^{(i)})(x_t - \mu_k^{(i)})^T}{\sum_t Q_t(k)}$$

→ weight on outer product that is your computation of the empirical covariance.

$$\pi_k^{(i)} = \frac{\sum_t Q_t^{(i)}(k)}{n}$$

---

will post the calc's for Σ, π, but let's try μ:
→ actually, let's talk about "why it makes sense", then go back to derivation of μ.

Why does this work? Intuition:

E step ~~at~~ never decreases log-likelihood, and neither does M-step.
→ to show: $\log \text{Lik}(\Theta^{(i+1)}) \geq \log \text{Lik}(\Theta^{(i)})$. → cornerstone of convex (did he say analysis) (line is above the curve). log is concave, which goes the other way.

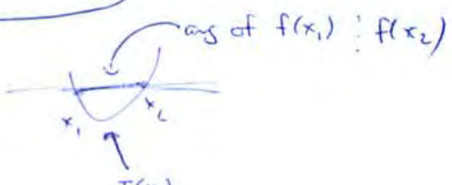
steps: insert latent vars, use ~~Jensen~~ Jensen's inequality, massage.

$$\sum_t \log P(x_t|\Theta^{(i+1)}) = \sum_t \log \sum_{k=1}^{K} P(x_t, c_t = k | \Theta^{(i+1)})$$

← intro of variable you will marginalize out

$$= \sum_t \log \sum_{l=1}^{K} \frac{Q_t^{(i+1)}(l)}{Q_t^{(i+1)}(l)} P(x_t, c_t = l | \Theta^{(i+1)})$$

Jensen's ineq: for f convex, $f(E(X)) \leq E f(x)$ — avg value



→ (it's practically a def'n of convexity)

consider: and let's call it ~~...~~. $H(k)$, so to get ~~$\sum_t \log \sum_{l=1}^{K} Q_t^{(i+1)}(l)$~~

notice fn of l

$$\sum_t \log \mathop{E}_{l \sim Q_t^{(i+1)}} [H(l)]$$

– by Jensen's on concave fns, such as the log.

$$\geq \sum_t \sum_k Q_t^{(i+1)}(k) \cdot \log H(k)$$

$$= \sum_t \sum_k Q_t^{(i+1)}(k) \left[\log P(x_t, c_t = l | \Theta^{(i+1)})\right] - \sum_t \sum_l Q_t^{(i+1)}(l) \log Q_t^{(i+1)}(l)$$

↳ ~~has a dependence on~~ Θ ↳ no dependence on Θ.

for the $i+1^{\text{th}}$ iteration, we're picking the ~~thing~~ Θ that maxes 1^{st} term.
⇒ 1^{st} term must be bigger than its value @ $\Theta = \Theta^{(i)}$.

So, then, why did we do the expectation step?

this comes in handy here:

$$\sum_n \sum_\ell Q_t^{(i+1)}(\ell) \log \left(\frac{P(x_t, c_t = \ell \mid \theta^{(i)})}{Q_t^{(i+1)}(\ell)} \right) = \sum_n \sum_\ell Q_t^{(i+1)}(\ell) \log \frac{P(c_t = \ell \mid x_t, \theta^{(i)}) \cancel{(} P(x_t \mid \theta^{(i)})}{Q_t^{(i+1)}(\ell)}$$

$$= \sum_n \sum_\ell Q_t^{(i+1)}(\ell) \underbrace{\log P(x_t \mid \theta^{(i)})}_{\text{no dependence on } \ell}, \text{ so } \boxed{= \sum_n \log P(x_t \mid \theta^{(i)})}$$

can hit a stationary point

an intuition: hallucinate the hidden variables' values.
this isn't specific to cluster assignments

k-means: a hard version (hard assignment) on Q's. (and assume spherical covariance).
soft k-means: still assume spherical Gaussians.