Mixture Models

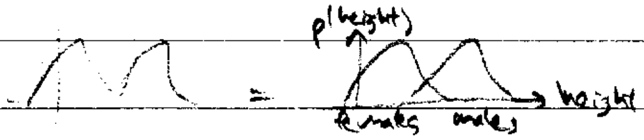
In e-fam: $\log\left(\prod_{n=1}^{N} p(x_n|\theta)\right)$

$\sum_{n=1}^{N} \log p(x_n|\theta)$ if $p(x|\theta) = \exp(H(x) + \theta^T x - A(\theta))$

log & exp annihilate

Easy MLE, nice unique max. step. $\frac{\partial}{\partial\theta} := 0$

But what if want more general / powerful distribs?



2 subpopulations in data
Can represent as 2 Gauss.

or nonlinear complicated distrib  3 Gaussis.

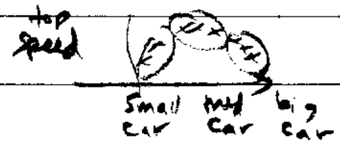
consider mixtures of Gaussians (multinom / e-families

$p(x) = \sum_i \alpha_i \frac{1}{(2\pi)^{D/2}\sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(x_i-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}$

$\quad\quad$ " $N(x|\mu_i, \Sigma_i)$

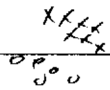
mixture → hidden variable / latent / unobserved class label

- male / female
- or conceptual hidden variable



$p(x|\theta) = \sum_i \pi_i N(x_i|\mu_i, \Sigma_i)$ or $\sum_i \pi_i \exp(H(x) + x^T\theta_i - A(\theta_i))$ $\quad \pi_i \geq 0$, $\sum_i \pi_i = 1$

- subpopulations in data
- hidden variable selects between them (coin flip)
- clustering: interpret the sub-pops. (i.e. word keys data contains 2 doc types)
- density est. more complicated distrib
- classification: we observed label, here it is hidden

 in mixture model, label is unobserved

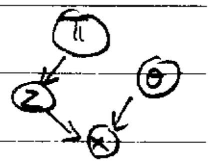
hidden vars: $Z = \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix}$ $Z^i = 1$ or not or $z = 1 \ldots K$

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

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

or

$$p(x|\theta) = \sum_i p(z^i=1|\pi_i)\, p(x|z^i=1,\theta)$$



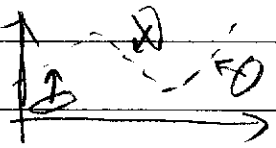
$\pi_i = p(z=i) = p(z^i=1|\pi_i)$ = prior probs ≡ mixing proportions, $\sum_i \pi_i = 1$

$p(x|z^i=1,\theta)$ = mixture components

$\tau_i = p(z^i=1|x,\theta)$ = posterior probs ≡ responsibilities, $\sum_i \tau_i = 1$

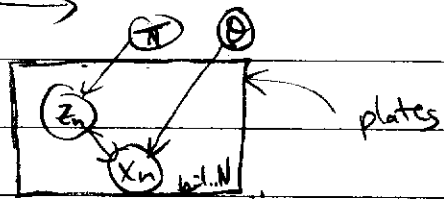
$= \dfrac{p(x|z^i=1,\theta)\,p(z^i=1|\theta)}{p(x|\theta)}$

Learning with Mixtures? Given Data how to fit model?



Maximum Likelihood

IID Data



$p(\theta|X) \propto p(X|\theta)$ if ML, uniform prior on model

$p(X|\theta) = \prod_{n=1}^{N} p(x_i|\theta)$

$l(\theta) = \log P(X|\theta) = \sum_{n=1}^{N} \log p(x_i|\theta) = \sum_{n=1}^{N} \log \sum_{i=1}^{K} \pi_i N(x_n|\mu_i,\Sigma_i)$ or expo-fam

$\frac{\partial}{\partial\theta}$ maximization step is ugly

- parameters are coupled via log-sum
- separate or decouple? Reuse E-Family Machinery?
- gradient descent / newton raphson θ_1 θ_2 slow steps

- Problem with log-sum:

  Σ e-family $\neq$ e-family

  Π e-family = e-family

  So mixture models break clean ML for mix models

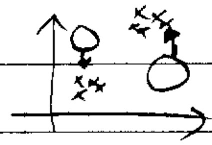
- We know how to estimate 1 single distrib in mix with ML

  Can we break down several mixture model estimates into multiple e-family estimates?

- Heuristic Divide & Conquer ... K-means ... formal ... EM

<u>K-Means</u> - Old heuristic algorithm

- "gobble up" data with divide & conquer



Timeline:

	old	K-means
	1977	Dempster, ..., Baum & Egon
	1980's	Csiszar & Tusnady
	1993	Neal & Hinton
	1995	em, Amari, information geometry

Chicken & Egg: If we know classes, easy to get model (e-family)

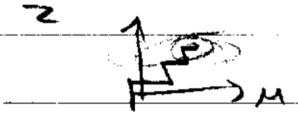
If we know model, easy to get classes

- for each x_n, define z_n indicator vector which classifies it, $\sum_i z_n^i = 1$

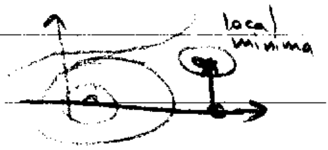
K-Means: $\mu_i = \frac{\sum_n z_n^i x_n}{\sum_n z_n^i}$ $\qquad z_n^i = \begin{cases} 1 & \text{if } i = \arg\min_j \|x_n - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$

closest point gets "gobbled up"

Kmeans Cost fn $\min_{\mu} \min_{z} J(\mu_1, \ldots, \mu_K, z_1, \ldots, z_N) = \sum_{n=1}^{N} \sum_{i=1}^{K} z_n^i \|x_n - \mu_i\|^2$



coordinate descent
axis parallel optimization
alternating minimization



EM - Maximum Likelihood for Mix Models

Outline:
- EM for mix of Gaussians
- EM as bound maximization
- EM as alternating maximization
- EM as shading nodes
- EM as Min KL & Info-Geometry...

EM mix of Gaussians - "fuzzify" Softening K-means, instead of winner takes all

K-means: $z = \arg\max_i p(x|z^i=1, \Theta)$

slower than K-means but less greedy

Soft assignment, $\propto \pi_i \frac{1}{(2\pi)^{D/2}} \exp(-\frac{1}{2}\|x_n - \mu_i\|^2)$

look at $\tau_n = p(z|x_n)$ as shared responsibility for a point

share instead of gobble up.

$\mu_i = \frac{\sum_n \tau_n^i x_n}{\sum_n \tau_n^i}$ ← use expectation of z, instead of hard select

$\tau_n^i = E[z_n^i | x_n] = p(z^i = 1 | x_n)$

Estep. $\tau_n^{i\,(t)} = \frac{\pi_i^{(t)} N(x_n | \mu_i^{(t)}, \Sigma_i^{(t)})}{\sum_j \pi_j^{(t)} N(x_n | \mu_j^{(t)}, \Sigma_j^{(t)})}$

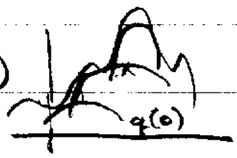
Mstep: $\mu_i^{(t+1)} = \frac{\sum_n \tau_n^{i\,(t)} x_n}{\sum_n \tau_n^{i\,(t)}}$ $\quad \Sigma_i^{(t+1)} = \frac{\sum_n \tau_n^{i\,(t)} (x_n - \mu_i^{(t+1)})(x_n - \mu_i^{(t+1)})^T}{\sum_n \tau_n^{i\,(t)}}$

$\pi_i^{(t+1)} = \frac{1}{N} \sum_n \tau_n^{i\,(t)}$

Not yet formal. Not coord ascent.
Why converge? Why divide & conquer?
divide & conquer need not always work...

EM as Bound Maximization

$f(\theta)$



$f(\theta) \geq q_t(\theta) \quad \forall \theta \ \& \ \forall_t$

$f(\theta^t) = q_t(\theta^t) \qquad q_t(\theta^{t+1}) > q_t(\theta^t)$

$f(\theta^{t+1}) \geq q(\theta^{t+1}) > q(\theta^t) = f(\theta^t)$

$\therefore \ f(\theta^{t+1}) > f(\theta^t)$ monotonic increase

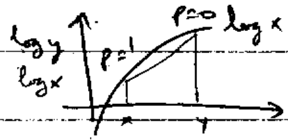
How to Bound? Jensen

for concave f:

$f(E\{x\}) \geq E\{f(x)\}$

i.e. if $\sum_i \rho_i = 1$ & $\rho_i \geq 0$

$\log(\sum_i \rho_i x_i) \geq \sum_i \rho_i \log(x_i)$



$\log(\rho x + (1-\rho)y) \geq \rho \log x + (1-\rho)\log y$

$\max_\theta l(\theta) = \sum_{n=1}^N \log \sum_z p(x_n, z|\theta)$

$\equiv \max_\theta \sum_{n=1}^N \log \sum_z p(x_n, z|\theta) - \underbrace{l(\theta^t)}_{\text{constant}} \equiv \max_\theta \Delta l(\theta)$

Same max locus yet $\Delta l(\theta^t) = 0$

$\sum_{n=1}^N \log \sum_z p(x_n, z|\theta) - \sum_{n=1}^N \log \sum_z p(x_n, z|\theta^t)$

$\sum_{n=1}^N \log \frac{\sum_z p(x_n, z|\theta)}{\sum_z p(x_n, z|\theta^t)} = \sum_n \log \sum_z \frac{p(x_n, z|\theta)}{\sum_{\hat z} p(x_n, \hat z|\theta^t)} \times 1$

$= \sum_n \log \sum_z \frac{p(x_n, z|\theta)}{\sum_{\hat z} p(x_n, \hat z|\theta^t)} \frac{p(z|x_n, \theta^t)}{p(z|x_n, \theta^t)}$

$= \sum_n \log \sum_z p(z|x_n, \theta^t) \frac{p(x_n, z|\theta)}{\left(\sum_{\hat z} p(x_n, \hat z|\theta^t)\right) p(z|x_n, \theta^t)}$ $\qquad$ $= p(x_n, z|\theta^t)$

$\geq \underbrace{\sum_n \sum_z \underbrace{p(z|x_n, \theta^t)}_{\tau_n} \log \frac{p(x_n, z|\theta)}{p(x_n, z|\theta^t)}}_{Q(\theta|\theta^t) \text{ auxilliary function}}$ $\leftarrow$ const

"Pulled" log inside sum

$\theta^{t+1} = \arg\max_\theta Q(\theta|\theta^t)$

$= \arg\max_\theta \sum_n \sum_z \tau_n^z \log p(x_n, z|\theta)$ $\qquad$ straight forward to derive w.r.t. θ & set to 0

$\frac{\partial}{\partial u_k}$ & $\sum_n \tau_n^k (\mu_k - x_n) := 0$

EM as Expected Likelihood

$Q(\theta|\theta^t)$ = expected "complete" likelihood + constant

$z \to x$ incomplete data since z unobserved (only x)

$z \to x$ complete data if z is observed (like classification)

$l(\theta)$ = incomplete log-likelihood = $\sum_n \log \sum_z p(x_n, z|\theta)$

$l^c(\theta)$ = complete log-likelihood = $\sum_n \log p(x_n, z_n|\theta)$ ← no log-sum

but we don't know z, so use expected values of z assuming current model θ^t, i.e. $p(z_n|x_n, \theta^t)$

$$E_{\substack{\text{distrib.}\\ \text{over } z}}\left[l^c(\theta)\right] = \sum_{z_1=1}^{k} \cdots \sum_{z_N=1}^{k} \prod_{i=1}^{N} p(z_n|x_n, \theta^t) \; l^c(\theta)$$

rewrite: $l^c(\theta) = \sum_n \log p(x_n, z_n|\theta)$

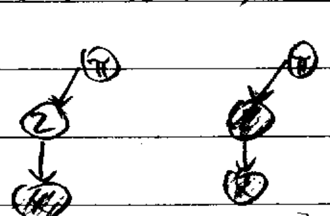
$$E[l^c] = \sum_{z_1} \cdots \sum_{z_n} \prod_{n=1}^{N} p(z_n|x_n, \theta^t) \sum_n \log p(x_n, z_n|\theta)$$

$$E[l^c] = \sum_n \sum_{z_n} p(z_n|x_n, \theta^t) \log p(x_n, z_n|\theta) \underbrace{\sum_{z_1} \cdots \sum_{z_i \neq z_n} \cdots \sum_{z_N} \prod_{i \neq n}^{N} p(z_i|x_i, \theta^t)}_{(1)\times(1)\times\cdots\times(1)}$$

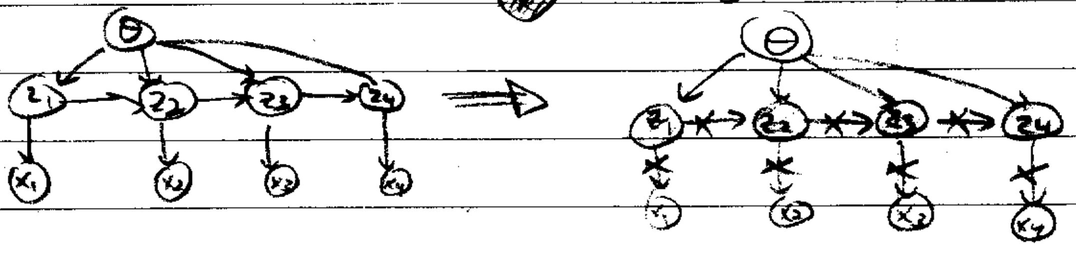
$$E[l^c] = \sum_{n=1}^{N} \sum_{j=1}^{k} p(j|x_n, \theta^t) \; \log p(x_n, j|\theta)$$

↑ like $Q(\theta|\theta^t)$ function + constant

EM as "Filling in" Nodes



Hmm



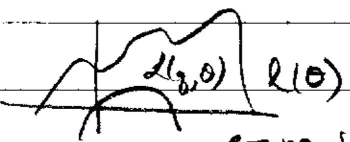
EM as Alternating Maximization ← more general (Neal & Hinton)

$$\ell(\theta) = \sum_{n=1}^{N} \log p(x_n|\theta) = \sum_{n=1}^{N} \log \sum_z p(x_n, z|\theta)$$

$$= \sum_{n=1}^{N} \log \sum_z q_n(z|x_n) \frac{p(x_n,z|\theta)}{q_n(z|x_n)} \qquad \underline{\underline{\text{any}}}\ q \geq 0\ \&\ \sum_z q_n(z|x_n) = 1$$

$$\geq \sum_{n=1}^{N} \sum_z q_n(z|x_n) \log \frac{p(x_n,z|\theta)}{q_n(z|x_n)}$$

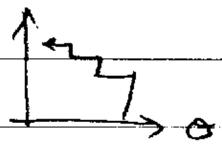
$$\geq \mathcal{L}(q,\theta)$$



← no tangential contact

E: $q^{t+1} = \arg\max_q \mathcal{L}(q, \theta^t)$

M: $\theta^{t+1} = \arg\max_\theta \mathcal{L}(q^{t+1}, \theta)$



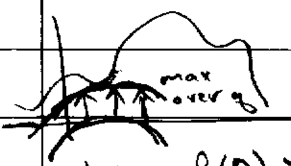
$$\mathcal{L}(q,\theta) = \underbrace{\sum_{n=1}^{N} \sum_z q_n(z|x_n) \log p(x_n, z|\theta)}_{\text{Expected}_q [\ell^c]} - \underbrace{\sum_{n=1}^{N} \sum_z q_n(z|x_n) \log q_n(z|x_n)}_{\text{constant with } \theta}$$

← different "q" though.

Mstep $\max_\theta$ or $\frac{\partial}{\partial\theta} := 0$ is easy, for e-family m-step is standard

else "GEM" is partial m-step

EStep maximize over q, closes gap with current tangential contact



$$\max_q \mathcal{L}(q,\theta) \Rightarrow q_n^* := p(z|x_n, \theta^t)$$

check: $\ell(\theta) \geq \mathcal{L}(\theta, q)$

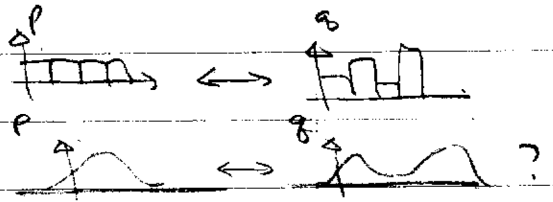
$$\ell(\theta^t) \geq \mathcal{L}(\theta^t, q^*) = \sum_{n=1}^{N} \sum_z p(z|x_n, \theta^t) \log \frac{p(x_n, z|\theta)}{p(z|x_n \theta^t)} \Big|_{\theta = \theta^t}$$

$$\ell(\theta^t) \geq \sum_{n=1}^{N} \underbrace{\sum_z p(z|x_n, \theta^t)}_{=1} \log p(x_n|\theta^t)$$

$$\ell(\theta^t) \geq \sum_{n=1}^{N} \log p(x_n|\theta^t) = \ell(\theta^t)$$

can't increase q any more since R.H.S. is equal to L.H.S. upper bound

else "Incremental EM" is partial E-step



Alternative Deriv. of ML

distance between two distribs

Dist $(p,q) = ?$ Euclidean? $\|p-q\|^2$ no, since $\int p = 1$ & $p \geq 0$

Generalized Divergences: Bregman Divs.

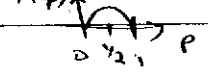
Euclidean, Mahalanobis, KL, $K(p) - K(q) - K'(q)^T(p-q)$

for any convex K function on the space of p

Shannon, Kullback-Leibler 1950's information theory

for K = neg entropy, $K(p) = -\int p \log p$, get $KL(p,q)$

entropy ↑, disorder, more uniform



KL-divergence: $KL(p,q) = \int p \log p/q = \sum_i p_i \log p_i/q_i$

natural metric on probability manifolds

$D(p\|q) \neq D(q\|p)$ assymmetric

abstract non-Euclidean prob. space

Empirical Distrib of a data set: $\hat{p}(x) = \frac{1}{N}\sum_{n=1}^{N} \delta(x, x_n)$

$\sum_x \hat{p}(x) = 1$, $\hat{p}(x) \geq 0$

not a very good model, ∞ if we observe a point, 0 otherwise

no generalization power.

our model

$$D(\hat{p}(x) \| p(x|\theta)) = \sum_x \hat{p}(x) \log \frac{\hat{p}(x)}{p(x|\theta)} = \sum_x \hat{p}(x) \log \hat{p}(x) - \sum_x \hat{p}(x) \log p(x|\theta)$$

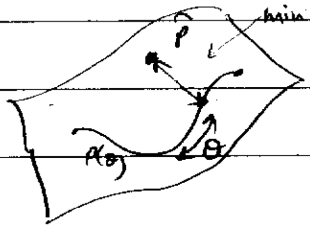
$$= \sum_x \hat{p}(x) \log \hat{p}(x) - \sum_x \frac{1}{N} \sum_{n=1}^{N} \delta(x, x_n) \log p(x|\theta)$$

$$= \sum_x \hat{p}(x) \log \hat{p}(x) - \frac{1}{N} \sum_{n=1}^{N} \log p(x_n|\theta)$$

constant (neg. entropy) $\quad$ $\frac{1}{N} \times$ log-likelihood

$\min D(\hat{p} \| p(x|\theta)) \equiv$ max. likelihood

min. KL, geometric picture.



Information Geometry of EM (Amari)

consider single data point (easy to extend to more)

$\ell(\theta) = \log p(x|\theta) \geq \mathcal{L}(q,\theta)$ from before but $i = 1 \ldots l$

$$\min_\theta D(\hat{p}(x) \,\|\, p(x|\theta)) = \underbrace{\sum_x \hat{p}(x) \log \hat{p}(x)}_{\text{constant}} - \sum_x \hat{p}(x) \underbrace{\log p(x|\theta)}_{\ell(\theta)}$$

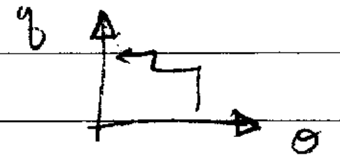
$$= \sum_x \hat{p}(x) \log \hat{p}(x) - \sum_x \hat{p}(x) \ell(\theta)$$

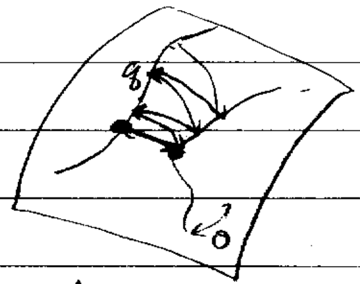
$$\leq \sum_x \hat{p}(x) \log \hat{p}(x) - \sum_x \hat{p}(x) \mathcal{L}(q,\theta)$$

$$\leq \sum_x \tilde{p}(x) \log \hat{p}(x) - \sum_x \hat{p}(x) \sum_z q(z|x) \log \frac{p(x,z|\theta)}{q(z|x)}$$

$$\leq \sum_x \hat{p}(x) \sum_z q(z|x) \left(\log \frac{\hat{p}(x)\, q(z|x)}{p(x,z|\theta)} \right)$$

$$\underbrace{D(\hat{p}(x) \| p(x|\theta))}_{\text{incomplete divergence}} \leq \underbrace{D(\hat{p}(x) q(z|x) \,\|\, p(x,z|\theta))}_{\text{"complete" divergence}}$$





iterated projections (i.e. min KL's)

e-step $\quad q^{t+1}(z|x) = \arg\min_q D(\hat{p} q \,\|\, p(x,z|\theta^t))$

m-step $\quad \theta^{t+1} = \arg\min_\theta D(\hat{p} q^t \,\|\, p(x,z|\theta))$

little em yields slightly different results than EM