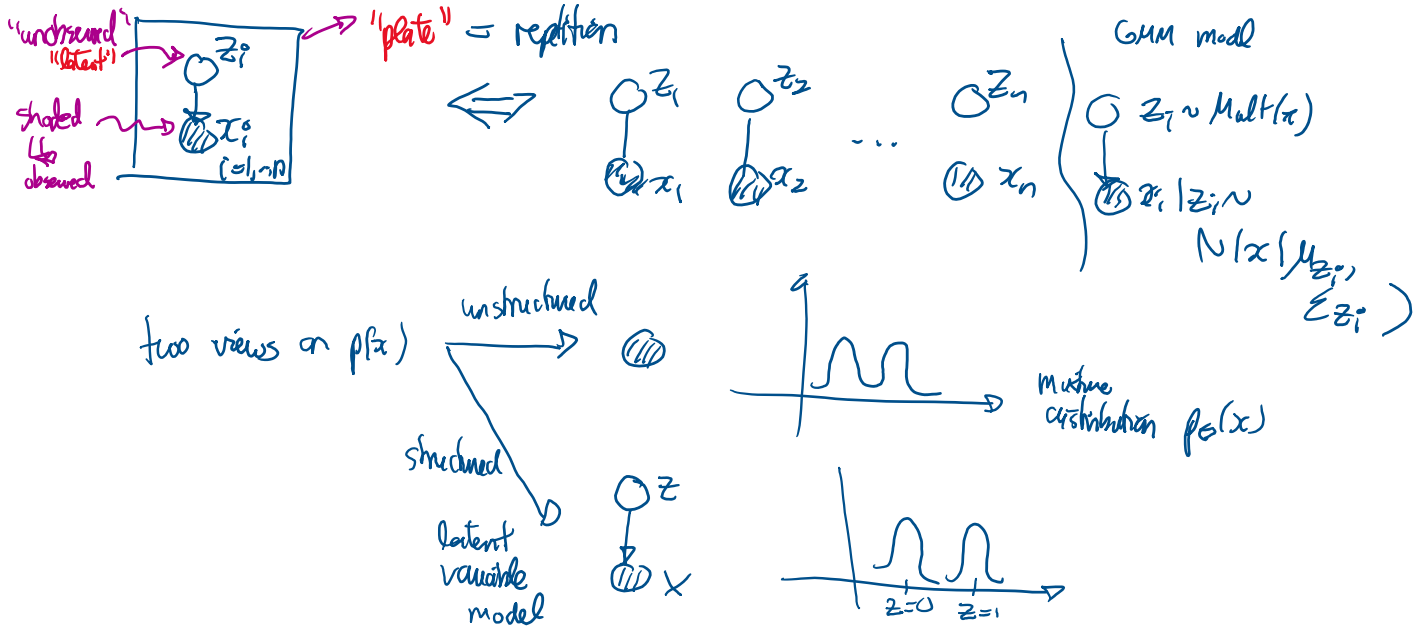
Lecture 10 - K-means / EM for GMM

Monday, October 7, 2024 12:30 PM

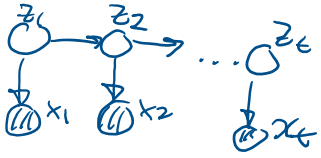
today: • k-means
• EM for GMM

graphical model for this "latent variable model"

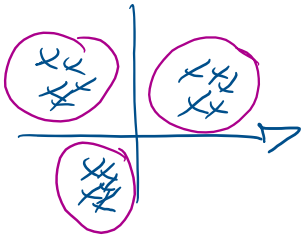


$$p(x) = \sum_z p_\theta(x|z)\, p_\theta(z)$$

(later in class, we will add time structure: HMM)



k-means → e.g. to clustering i.e. group data



we want to get a cluster assignment for every data pt. x_i

represent $z_{i,j} = 1$ to mean that x_i belongs to cluster j

$j = 1, \dots, k$ (specified in advance for k-means)

k = # of clusters

applications: • vector quantization (compression)

• in computer vision: use k-mean to get "bag of visual words" representation of image patches

• many many others?

K-mean alg: → can derive as block-coordinate minimization alg. of obj fct.:

"distortion measure" → $J(z, \mu) \triangleq \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|_2^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{k} z_{i,j} \|x_i - \mu_j\|^2 \right)$

- z: cluster assignment
- μ: cluster centroids
- z_i: cluster index represented by z_i
- $z_1, \dots, z_k \in$ corners of Δ_k ("one hot encoding")
- $\mu_1, \dots, \mu_k \in \mathbb{R}^d$

alg.: 1) initialize $\mu^{(1)}$

2) iterate until convergence

"E step": $z^{(t+1)} = \underset{z \text{ valid assign.}}{\operatorname{argmin}} J(z, \mu^{(t)})$

$\Rightarrow z_{i,j^*}^{(t+1)} = 1$ for $j^* = \underset{j}{\operatorname{argmin}} \|x_i - \mu_j^{(t)}\|$

"M step": $\mu^{(t+1)} = \underset{\mu \in \mathbb{R}^{d \times k}}{\operatorname{argmin}} J(z^{(t+1)}, \mu)$

$$\Rightarrow \boxed{\mu_j^{(t+1)} = \frac{\sum_i z_{i,j} x_i}{\left(\sum_i z_{i,j}\right)}}$$ empirical mean of cluster j

Visualization:
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

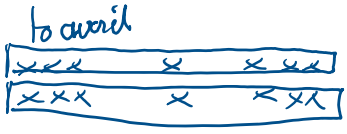
properties:

1) converge in finite # of iterations to a local min — note: run for multiple random initializations

2) NP hard in general to compute global min in z

K-means++: clever initialization scheme which guarantees that alg. is within $O(\log(k))$ of global opt. (w.h.p.) + run K-means

→ idea: spread as much as possible the initial means

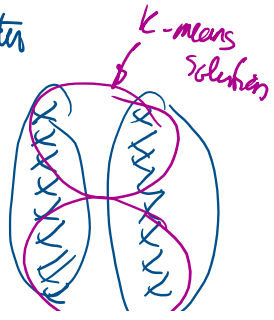


3) choice of k?

- one heuristic is $J(z, \mu, k) = \sum_i \sum_j z_{i,j} \|x_i - \mu_j\|^2 + \lambda k$ (λ: hyperparameter)

Side reference I mentioned: see https://icml.cc/2012/papers/291.pdf for interpreting regularized K-means as approximate inference in a Dirichlet process mixture model... [by Kullis & Jordan]

4) K-means is very sensitive in distance measure: it assumes spherical clusters



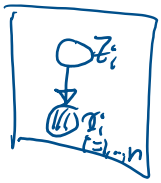
↳ GMM fixes this problem (partially)

↳ GMM fixes this problem (partially)

Mahalanobis distance $d_\Sigma(x,x') \triangleq \sqrt{(x-x')^T \Sigma^{-1} (x-x')}$

EM – maximum likelihood in latent variable model

setup



z latent variable

x observed variable

log-likelihood $\log p(x_1,\dots,x_n;\theta) = \log\left(\prod_{i=1}^n p(x_i;\theta)\right)$

$= \sum_{i=1}^n \log p(x_i;\theta)$

$= \sum_{i=1}^n \log\left(\sum_{z_i} p(x_i,z_i;\theta)\right)$

problem? → yields a multi-modal opt. problem (non-convex)

options MLE in latent variable model

1) do gradient ascent on a non-concave obj.

2) EM alg. → block coordinate ascent on a auxiliary fct. that lower bounds $\log p(x_{1:n};\theta)$

nice interpretation in terms of "filling "missing data")

ie. E step ⇝ fill z with "soft-values"

M step ⇝ max w.r. to θ for fully observed model

trick overview:

$\log \sum_z p(x,z) = \log \sum_z q(z) \frac{p(x,z)}{q(z)}$

$= \log \mathbb{E}_q\left[\frac{p(x,z)}{q(z)}\right]$

Jensen's ineq. trick

$\geq \mathbb{E}_q\left[\log \frac{p(x,z)}{q(z)}\right]$

$= \sum_z q(z) \log p_\theta(x,z) - \sum_z q(z) \log q(z)$

$\triangleq \mathcal{J}(q,\theta) \triangleq \mathbb{E}_q[\log p(x,z;\theta)] + H(q)$

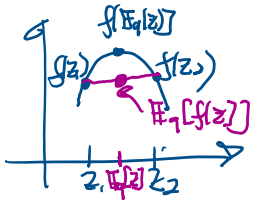
"expected complete log-likelihood"

"entropy" of q

Jensen's inequality

$\mathbb{E}_q[f(q(z))] \leq f(\mathbb{E}_q[q(z)])$

when f is concave



13h36

we have $\log p(x;\theta) \geq \mathcal{L}(q,\theta) \quad \forall q \; \& \; \theta$

EM algorithm: E step: $q_{t+1} \triangleq \underset{q \text{ distributions over } z}{\arg\max}\; \mathcal{L}(q,\theta_t) \Rightarrow$ turns out $q_{t+1}(z) = p(z|x;\theta_t)$

M step: $\theta_{t+1} \triangleq \underset{\theta}{\arg\max}\; \mathcal{L}(q_{t+1},\theta)$
$= \underset{\theta}{\arg\max}\; \mathbb{E}_{q_{t+1}}[\log p(x,z;\theta)]$

↳ this is another MLE problem, but for complete information

(often, replace z with $\mathbb{E}_q[z]$ in this expression)

E step derivation: by Jensen's inequality

* we had $\log p(x;\theta) \geq \mathcal{L}(q,\theta)$ $\qquad \log(\mathbb{E}_q[g(z)]) \geq \mathbb{E}_q[\log(g(z))]$

in Jensen's ineq., you get a strict inequality (when f is strictly concave) unless R.V. is deterministic

↳ here $g(z)$ being deterministic $\Rightarrow$ constant

above $g(z) = \frac{p(x,z)}{q(z)} \overset{\text{if}}{=} \text{constant } \forall z$

$\Rightarrow q(z) \propto p(x,z)$

i.e. $q^*(z) = p(z|x;\theta_t)$ $\quad (= q_{t+1})$

$\mathcal{L}(q_{t+1},\theta_t) = \log p(x;\theta_t) \geq \mathcal{L}(q,\theta_t) \; \forall q$

$\Rightarrow q_{t+1}$ maximizes $\mathcal{L}(q,\theta_t)$ w.r.t. q

i.e. $\underset{q}{\arg\max}\; \mathcal{L}(q,\theta_t) = p(z|x;\theta_t) \triangleq q_{t+1}$

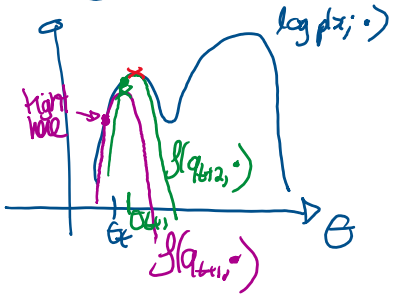
and $\mathcal{L}(q_{t+1},\theta_t) = \log p(x;\theta_t)$

properties of EM algorithm

a) $\log p(x;\theta_{t+1}) \geq \log p(x;\theta_t)$ $\qquad$ proof: $\log p(x;\theta_{t+1}) \geq \mathcal{L}(q_{t+1},\theta_{t+1})$

(by M step) $\geq \mathcal{L}(q_{t+1},\theta_t)$

$= \log p(x;\theta_t)$ //



b) θ_t in EM converges to a stationary pt. of $\log p(x;\theta)$

i.e. $\nabla_\theta \log p(x;\theta)\big|_{\hat{\theta}} = 0$

i.e. $\nabla_\theta \log p(x;\theta)\big|_{\hat\theta} = 0$

like k-means, initialization is crucial
↝ usually do random restarts

- for GMM, could use k-means++ to initialize μ's

c) $\mathcal{J}(q,\theta) = \mathbb{E}_q\left[\log \frac{p(x,z)}{q(z)}\right]$

$\log p(x;\theta) - \mathcal{J}(q,\theta) = -\mathbb{E}_q\left[\log \frac{p(x,z|\theta)}{q(z)\,p(x;\theta)}\right]$ ($p(z|x;\theta)$)

$\log p(x;\theta)$ } $KL(q \| p(\cdot|x;\theta))$
$\mathcal{J}(q,\theta)$ }

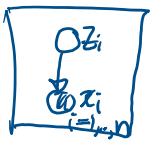
$= \mathbb{E}_q\left[\log \frac{q(z)}{p(z|x;\theta)}\right] \triangleq KL(q(\cdot) \| p(\cdot|x;\theta))$ KL divergence

we will revisit this for variational inference

$q \in \mathcal{Q}$ ("simple distributions")

block-coordinate method sometimes slow

for GMM model:



$z_i \sim \text{Mult}(\pi)$

$x_i | z_i = j \sim N(\mu_j, \Sigma_j)$

shorthand to say $z_{ij} = 1$

$\theta = (\pi, (\mu_j)_{j=1}^k, (\Sigma_j)_{j=1}^k)$

notation: $x = x_{1:n}$

$z = z_{1:n}$

exercise $p(z|x) = \prod_{i=1}^n p(z_i|x) = \prod_{i=1}^n p(z_i|x_i)$

complete log-likelihood

$\log p(x,z;\theta) = \sum_{i=1}^n \log p(x_i|z_i;\theta) + \log p(z_i;\theta)$

Gaussian, multinoulli

$= \sum_{i=1}^n \left[\sum_{j=1}^k z_{ij} \log N(x_i|\mu_j,\Sigma_j) + \sum_{j=1}^k z_{ij} \log \pi_j\right]$

$\mathbb{E}_q[\log p(x,z;\theta)] = \sum_{i=1}^n \left[\sum_{j=1}^k \mathbb{E}_q[z_{ij}] (\log N(x_i|\mu_j,\Sigma_j) + \log \pi_j)\right]$

$\mathbb{E}_q[\mathbb{1}\{z_{ij}=1\}] = q(z_{ij}=1)$ (marginal)

during EM: $q_{t+1}(z) = p(z|x;\theta_t)$

weight $\gamma_{ij}^t \triangleq p(z_{ij}=1|x;\theta_t) = q_{t+1}(z_{ij}=1)$

$q_{t+1}(z_i = j) = p(z_i = j \mid x_i; \theta^{(t)})$

↳ marginal of z_i ?

↳ weight $\gamma_{ij}^{(t)} \triangleq p(z_{ij}=1 \mid x_i; \theta^{(t)}) = q_{t+1}(z_{ij}=1)$

E-step: compute $q_{t+1}(z) \triangleq p(z \mid x; \theta^{(t)})$

$$= \prod_{i=1}^{n} p(z_i \mid x_i; \theta^{(t)})$$

$$\Rightarrow q_{t+1}(z_i) = p(z_i \mid x_i; \theta^{(t)})$$

$$\propto p(x_i \mid z_i; \theta^{(t)})\, p(z_i; \theta^{(t)})$$

(Gaussian; $\pi(z_i)$)

$$\gamma_{ij}^{(t)} = q_{t+1}(z_{ij}=1) = \frac{\pi_j^{(t)} \mathcal{N}(x_i \mid \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{\ell=1}^{k} \pi_\ell^{(t)} \mathcal{N}(x_i \mid \mu_\ell^{(t)}, \Sigma_\ell^{(t)})} \quad \left| \begin{array}{l} \propto p(x_i, z_i \mid \theta^{(t)}) \\ \propto p(x_i \mid \theta^{(t)}) \end{array} \right.$$

E step for GMM: compute $\gamma_{ij}^{(t)}$ for $i=1,\dots,n$ using $\theta^{(t)}$

M step: $\max_{\{\mu_j, \Sigma_j, \pi_j\}} \sum_i \sum_j \gamma_{ij}^{(t)} [\log p(x_i \mid \mu_j, \Sigma_j) + \log \pi_j]$

Exercise: M step for EM in GMM

$$\hat{\pi}_j^{(t+1)} = \frac{\sum_i \gamma_{ij}^{(t)}}{n}$$ ← soft-counts

$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{ij}^{(t)} x_i}{\sum_{i=1}^{n} \gamma_{ij}^{(t)}}$$

$$\hat{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{ij}^{(t)} (x_i - \hat{\mu}_j^{(t+1)})(x_i - \hat{\mu}_j^{(t+1)})^T}{\sum_{i=1}^{n} \gamma_{ij}^{(t)}}$$

- initialize e.g. $\mu_j^{(0)}$ from k-means++
  $\Sigma^{(0)}$ big spherical covariance $\Sigma_j^{(0)} = \sigma^2 I$ (big σ^2)
  $\pi_j^{(0)}$: from proportions from k-means
- EM step in GMM with $\Sigma_j = \sigma^2 I$ with $\sigma^2 \to 0$
  ⇝ get k-means alg. "hard EM"