Lecture 10 - K-means / EM

Friday, October 7, 2022 2:17 PM

today: · k-means
· EM

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

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

cluster assignment $z_1, \dots, z_n$ ∈ corners of Δ_K ("one hot encoding")

$\mu_1, \dots, \mu_K \in \mathbb{R}^d$ cluster centroids

cluster index represented by z_i

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

2) iterate until convergence:

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

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

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

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

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

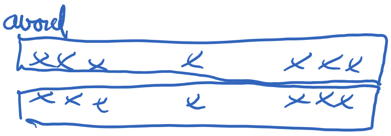
properties of k-means:

1) converge in finite # of iterations to a local min

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

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

→ idea: spread as much as possible the initial means to avoid



3) choice of K?

· one heuristic is $J(z, \mu, K) = \sum_i \sum_j z_{ij} \|x_i - \mu_j\|^2 + \lambda \cdot K$

hyperparameter

→ we'll see later in class "non-parametric" models

Side reference I mentioned:
see https://jmrl.cc/2012/papers/201.pdf

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]

→ we'll see later in class "non-parametric" models where "k" is basically infinite and can get $p(k \mid \text{data})$

hyperparameter

e.g. Dirichlet process mixture model

K-means solution

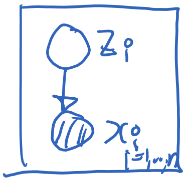
4) K-mean is very sensitive in distance measure: it assumes spherical clusters

→ GMM fixes that problem

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 $\log p(x_{1:n};\theta) = \log\left(\prod_{i=1}^{n} p(x_i;\theta)\right)$

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

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

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

options MLE in latent variable model

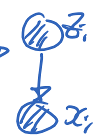
1) do gradient ascent on a non-concave obj.

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

nice interpretation in terms of filling "missing data"

i.e. E step → fill z with "soft-values"

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



trick overview:

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

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

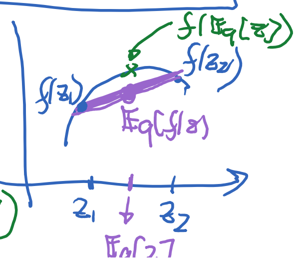
$$\geq \mathbb{E}_q\left[\log \frac{p(x,z)}{q(z)}\right]$$

Jensen's inequality trick?

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

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

Jensen's inequality
$\mathbb{E}_q[f(g(z))] \leq f(\mathbb{E}_q[g(z)])$
when f is concave



$\triangleq \mathcal{L}(q,\theta) \triangleq \underbrace{\mathbb{E}_q[\log p(x,z;\theta)]}_{\text{"expected complete log-likelihood"}} + \underbrace{H(q)}_{\text{"entropy of } q\text{"}}$

$z_1 \quad \mathbb{E}_q[z] \quad z_2$

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

EM algorithm: E step: $q_{t+1} \triangleq \underset{q \in \text{dist. over } z}{\operatorname{argmax}} \mathcal{L}(q,\theta_t) \Rightarrow q_{t+1}(z) = p(z|x;\theta_t)$

M step: $\theta_{t+1} \triangleq \underset{\theta}{\operatorname{argmax}} \mathcal{L}(q_{t+1},\theta)$

$= \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{q_{t+1}}[\log p(x,z;\theta)]$

this is another MLE problem, but for complete information

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

16h08

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 unless the dist. for $g(z)$ is degenerate (i.e. $g(z)$ takes only one value) i.e. when $g(z) =$ constant

(when f is strictly concave)
(log)

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

i.e. $\boxed{q(z) = p(z|x;\theta_t)}$

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

$\Rightarrow q_{t+1}$ maximizes $\mathcal{L}(q,\theta_t)$ w.r. to q $\qquad$ i.e. $\underset{q}{\operatorname{argmax}} \mathcal{L}(q,\theta_t) = p(z|x;\theta_t) = q_{t+1}$

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

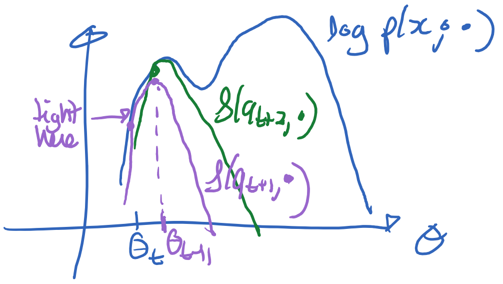
properties of EM algorithm

a) $\log p(x;\theta_{t+1}) \geq \log p(x;\theta_t)$ ✓

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

$\geq \mathcal{L}(q_{t+1},\theta_t)$ [because θ_{t+1} was maximizer]

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



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

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

like means, initialization is crucial

→ usually do random restarts

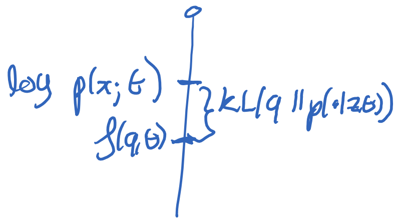
- for GMM, could use K-mean++ to initialize the μ's

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

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

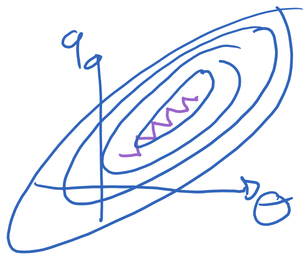
$$= \mathbb{E}_q\left[\log \frac{q(z)}{p(z|x;\theta)}\right] \triangleq KL\big(q(\cdot)\,\|\,p(\cdot|x;\theta)\big)$$

KL divergence



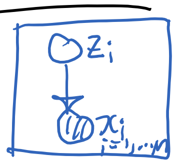
we will revisit this for variational inference $q \in \mathcal{Q}$

"simple distributions"



block-coordinate method sometimes slow

for GMM model:



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

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

shorthand to say $z_{i,j} = 1$

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

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:

$$\log p(x,z;\theta) = \sum_{i=1}^{n}\Big[\log \underbrace{p(x_i|z_i;\theta)}_{\text{Gaussian}} + \log \underbrace{p(z_i;\theta)}_{\text{multinoulli}}\Big]$$

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

$$\mathbb{E}_q[\log p(x,z;\theta)] = \sum_i\sum_j \mathbb{E}_q[z_{i,j}]\left(\log N(x_i|\mu_j,\Sigma_j) + \log \pi_j\right)$$

$\mathbb{E}_q[\mathbb{1}\{z_{i,j}=1\}] = q(z_{i,j}=1)$

$\mathbb{E}_q[z_{i,j}] = q(z_{i,j}=1)$ [marginal prob.]

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

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

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

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

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

E-step is computing $q_{t+1}(z) \triangleq p(z|x;\theta_t)$

$$= \prod_i p(z_i|x_i;\theta_t)$$

$$\Rightarrow q_{t+1}(z_i) = p(z_i|x_i;\theta_t)$$

$$\propto \underbrace{p(x_i|z_i;\theta_t)}_{\text{Gaussian}}\ \underbrace{p(z_i;\theta_t)}_{\pi^t_{z_i}}$$

$$\boxed{\tau_{ij}^t = q_{t+1}(z_{ij}=1) = \frac{\pi_j^{(t)} N(x_i|\mu_j^{(t)},\Sigma_j^{(t)})}{\sum_{l=1}^{k} \pi_l^{(t)} N(x_i|\mu_l^{(t)},\Sigma_l^{(t)})}} \quad \begin{matrix}\} p(x_i,z_i|\theta^{(t)}) \\ \} p(x_i|\theta^{(t)})\end{matrix}$$

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

M step: $\max_{\{\mu_j,\Sigma_j,\pi_j\}} \sum_{i=1}^n \sum_{j=1}^k \tau_{ij}^{(t)} [\log p(x_i|\mu_j,\Sigma_j) + \log \pi_j]$

exercise:

M step for EM for GMM

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

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

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

- initialize: e.g. $\mu_j^{(0)}$ from k-means++

  $\Sigma_j^{(0)}$ big spherical covariance $\Sigma_j^{(0)} = \sigma^2 I$ (big)

  $\pi_j^{(0)}$: proportions from k-means++

- EM step in GMM with fixed $\Sigma_j = \sigma^2 I$ with $\sigma^2 \to 0$

  $\rightsquigarrow$ get k-means alg. "hard EM"