Lecture 10 - scribbles

Friday, October 5, 2018 13:29

today: · finish k-means
· EM ⭐ & GMM (!!)

properties of k-means:

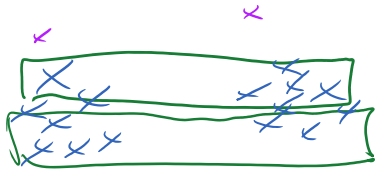
K-means demo: http://web.stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

1) converge in finite # of iterations to a local min
2) NP hard in general to find best z

k-means++: clever initialization scheme which guarantees that objective is with $\log K$ of global optimum (w.h.p.)

→ idea: spread out as much as possible the initial means

to avoid:



3) choice of k?

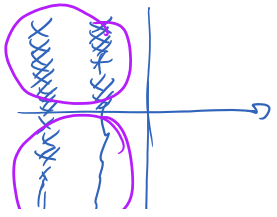
- one heuristic is: $J(\mu, z, K) = \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\|x_i - \mu_j\|^2 + \lambda K$

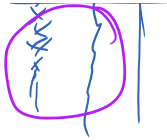
λ: hyperparameter

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

e.g. Dirichlet process mixture model

4) k-mean is very sensitive on distance measure: it assumes spherical cluster





↳ GMM fixes that

EM – maximum likelihood in latent variable model

setup:

x observed variable

z latent variable

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

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

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

problem?

→ gives multi-modal difficult optimization problem (non-convex)

options for ML in latent variable model

1) do gradient ascent on non-convex objective

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

nice interpretation in terms of filling "missing data"

ie. E step → fills z with "soft-values"

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

trick overview:

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

f(.) here is log(.)

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

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

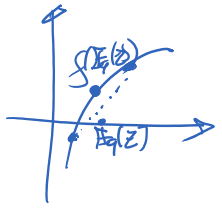
Jensen's inequality trick?

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

"entropy of q"

Jensen's inequality

$\mathbb{E}_q[f(z)] \leq f(\mathbb{E}_q(z))$ when f is concave



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

we get equality in Jensen if $\frac{p(x,z)}{q(z)} = \text{constant w.r. to } z \Rightarrow \frac{p(x,z)}{q(z)} = c$

$\mathbb{E}_q[c] = c$

$\log \mathbb{E}_q[c] = \log c$

$\mathbb{E}_q[\log c] = \log c$

(explanation: f(z) = log(z) is a *strictly* concave function; so this means Jensen's inequality becomes an equality only if the R.V. is constant -- this implies that p(x,z)/q(z) should not depend on z for this to happen)

ie. $q^*(z) \propto p(x,z)$

$$\Rightarrow q^*(z) = \frac{p(x,z)}{\sum_{z'} p(x,z)} = \frac{p(x,z)}{p(x)} = p(z|x)$$

this means that $\underset{q \in \text{distributions over } z}{\text{argmax}} \; \mathcal{L}(q,\theta_t) = p(z|x;\theta_t)$

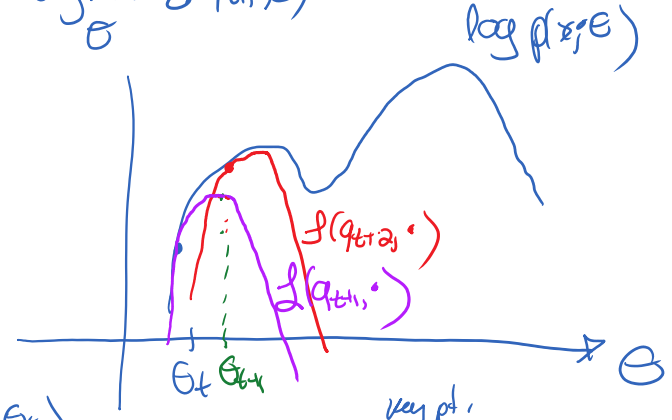
inference

$q \in$ distributions over z

EM algorithm: E step: $q_{t+1} \triangleq \arg\max_q \mathcal{J}(q, \theta_t) \Rightarrow q_{t+1}(z) = p(z|x;\theta_t)$ (inference)

M step: $\theta_{t+1} \triangleq \arg\max_\theta \mathcal{J}(q_{t+1}, \theta)$

block-coordinate ascent on $\mathcal{J}(q,\theta) \leq \log p(x;\theta)$



we have $\log p(x;\theta_t) = \mathcal{J}(q_{t+1}, \theta_t)$

$\Downarrow$

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

key pt.

$\log p(x;\theta_t) = \mathcal{J}(q_{t+1}, \theta_t) \leq \mathcal{J}(q_{t+1}, \theta_{t+1}) \leq \log p(x;\theta_{t+1})$

properties:

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

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

ie. $\nabla_\theta \log p(x;\theta)\big|_{\theta^*} = 0$

like k-means, initialization is crucial

→ usually random restarts

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

c) $\mathcal{J}(q,\theta) = E_q[\log p(x,z;\theta)]$ $p(z|x;\theta)$

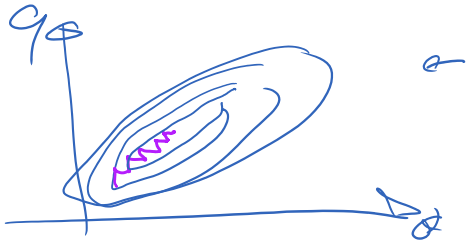
c) $\mathcal{L}(q,\theta) = \mathbb{E}_q\left[\log \frac{p(x,z;\theta)}{q(z)}\right]$

$p(z \mid x;\theta)$

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

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

(KL divergence)



← where conjugate gradient could be faster

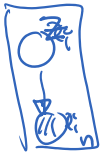
M step: $\theta^{t+1} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathbb{E}_{q_{t+1}}\left[\log p(x,z;\theta)\right] + cst.$

expected complete log-likelihood

↑ this is another ML problem, but for complete information

(usually replace z with $\mathbb{E}_q[z]$)

for GMM model:



$z_i \overset{iid}{\sim} Mult(\pi)$

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

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

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

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

shorthand for $z_{ij} = 1$

notation here $x = x_{1:n}$, $z = z_{1:n}$

complete log-likelihood

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

Gaussian log-likelihood; multinoulli

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

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

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

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

$$\propto p(x \mid z; \theta_t)\, p(z; \theta_t)$$

$$\prod_{i=1}^{n} \left(p(x_i \mid z_i; \theta_t)\, p(z_i; \theta_t) \right)$$

$$\Rightarrow q_{t+1}(z_i) \propto p(x_i \mid z_i; \theta_t)\, p(z_i; \theta_t)$$

$p(x_i \mid z_i; \theta_t) \rightarrow N(x_i \mid \mu_{z_i}, \Sigma_{z_i})$; $p(z_i; \theta_t) \rightarrow \pi_{z_i}$

weight $r_{ij}^{(t)} \triangleq p(z_{ij} = 1 \mid x_i; \theta_t)$

$$= q_{t+1}(z_{ij} = 1) = \frac{\pi_j^{(t)} N(x_i \mid \mu_j^{(t)}, \Sigma_j^{(t)}) \;\}\; p(x_i, z_{ij} = 1 \mid \theta^{(t)})}{\sum_{\ell} \pi_\ell^{(t)} N(x_i \mid \mu_\ell^{(t)}, \Sigma_\ell^{(t)}) \;\}\; p(x_i \mid \theta^{(t)})}$$

$$\sum_{\ell=1}^{k} \pi_\ell^{(t)} \mathcal{N}(x_i \mid \mu_\ell^{(t)}, \Sigma_\ell^{(t)}) \;\Big\}\; p(x_i \mid \theta^{(t)})$$

E step : compute τ_{ij}^t for $i = 1, \dots, n$, $j = 1, \dots, k$

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

Exercise : $\hat{\pi}_j^{(t+1)} = \sum_{i=1}^{n} \tau_{ij}^t$ ← "soft count"

$\hat{\mu}_j^{(t+1)} = \dfrac{\sum_{i=1}^{n} \tau_{ij}^t x_i}{\left(\sum_{i=1}^{n} \tau_{ij}^t\right)}$ ← soft-cluster assignment; [weighted empirical mean]

$\hat{\Sigma}_j^{(t+1)} = \dfrac{\sum_{i=1}^{n} \tau_{ij}^t (x_i - \hat{\mu}_j^{(t+1)})(x - \hat{\mu}_j^{(t+1)})}{\left(\sum_{i} \tau_{ij}^t\right)}$

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

$\Sigma_j^{(0)}$ big spherical variance $\Sigma_j^{(0)} = \sigma^2 I$, σ^2 big

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