

Day 20 in 2021s

SPR Day 16

A New View of EM
as a principled, general purpose
optimization algorithm

Reading: Bishop PRML
9.3
9.4

Topics: Recap: EM for GMMs
GMM as latent variable model

Idea 1: Complete likelihood easier than incomplete

Idea 2: Expectation of complete likelihood also easy

Idea 3: Can formulate objective function that:

- Principled: Is a lower bound of incomplete likelihood
- Tractable: Uses expectations of complete likelihood

Idea 4: EM is coordinate ascent optimization applied to this objective

Recap: EM for GMMs

Given: Data $\{x_n\}_{n=1}^N$, $x_n \in \mathbb{R}$ one dim. for now

Goal: Estimate GMM parameters to maximize (penalized) likelihood

$\pi = \{\pi_k\}_{k=1}^K$ mixture weights

$\mu = \{\mu_k\}_{k=1}^K$ means

$\sigma^2 = \{\sigma_k^2\}_{k=1}^K$ variances OR $\{\Sigma_k\}_{k=1}^K$ covars

Init: π, μ, σ

while not converged:

$\gamma_n = [\gamma_{n1} \dots \gamma_{nK}]$
is a vector w/ non-negative entries that sums to one

E step

for n in $1, 2, \dots, N$
for k in $1, 2, \dots, K$:

$$\gamma_{nk} = \frac{\pi_k \text{Norm PDF}(x_n | \mu_k, \sigma_k^2)}{\sum_l \pi_l \text{Norm PDF}(x_n | \mu_l, \sigma_l^2)}$$

Bishop 9.23

M step

for k in $1, \dots, K$:

$$\pi_k = \frac{\sum_n \gamma_{nk}}{N}$$

9.26

M step

$$\mu_k = \frac{\sum_n \delta_{nk} x_n}{\sum_n \delta_{nk}} \quad 9.24$$

$$\sigma_k^2 = \frac{\sum_n \delta_{nk} (x_n - \mu_k)^2}{\sum_n \delta_{nk}} \quad 9.25$$

Big idea:

Coordinate ascent algorithm
each substep updates
subset of variables (coordinates)

Open Questions (goals for today)

- What principles let us use EM?
- what objective are we optimizing?

Idea 1: Complete likelihood easier to optimize

Incomplete likelihood aka marginal of x
integrating away z

$$P(x) = \prod_n \text{GMMPDF}(x_n | \mu, \pi, \Sigma)$$

$$\log p(x) = \sum_n \log \sum_k \pi_k \text{NormPDF}(x_n | \mu_k, \Sigma_k^2)$$

notice:

π, μ, Σ
all interdependent

Complete likelihood aka joint of x and z

eg $\mu_1, \mu_2, \dots, \mu_k$
all interact

$$p(x, z) = \prod_{k=1}^K \pi_k^{z_{nk}} \text{NormPDF}(x_n | \mu_k, \Sigma_k^2)^{z_{nk}}$$

$$\log p(x, z) = \sum_k z_{nk} \log \pi_k + \sum_k z_{nk} \log \text{NormPDF}(x_n | \mu_k, \Sigma_k^2)$$

notice:
 π depends on z only

μ_k only depends on z_{nk}, x_n, Σ_k

Key Idea: Complete likelihood easier to optimize

If only we "knew" z , we could estimate π, μ, Σ easier

What if we knew a distribution over z ,
would that help?

Idea 2: Can optimize expectation of complete likelihood

Suppose we had a distribution over z that we thought was accurate

call it $q(z_n) = \text{Cat}(r_{n1}, r_{n2}, \dots, r_{nk})$

Suppose we wanted to compute $\hookrightarrow E_q[z_{nk}] = r_{nk}$

$E_{q(z)}[\log p(x, z | \theta)]$, could we?

$$E_{q(z)}[\log p(x, z)] = E_{q(z)}\left[\sum_{nk} z_{nk} \log \pi_k + z_{nk} \log \text{Normal PDF}(x_n | \mu_k, \sigma_k^2)\right]$$

$$= \sum_n \sum_k E_q(z_{nk}) \log \pi_k + E_q[z_{nk}] \log \text{Normal PDF}(x_n | \mu_k, \sigma_k^2)$$

easy to optimize π_k, μ_k, σ_k^2 if we know x and k

Punchline: Expectations of complete log likelihood

are easy to evaluate, easy to optimize
for π, μ, σ^2

Idea 3: We can develop an objective that is (a) a lower bound on $\log p(x)$ so can be interpreted in principled way as optimizing likelihood and (b) uses expectations of complete likelihood so it is tractable

Start with incomplete likelihood

$$\begin{aligned} \log p(x_n) &= \sum_{z_n} q(z_n) \log p(x_n) \\ &= \mathbb{E}_{q(z_n)} [\log p(x_n)] \end{aligned}$$

because $q(z)$ is valid PMF sum is 1

by defn of expectation

$$= \mathbb{E}_{q(z_n)} \left[\log \left(\frac{p(x_n, z_n)}{p(z_n | x_n)} \right) \right]$$

by Bayes rule

$$p(z_n | x_n) = \frac{p(x_n, z_n)}{p(x_n)}$$

$$= \mathbb{E}_{q(z_n)} \left[\log \left(\frac{p(x_n, z_n) q(z_n)}{p(z_n | x_n) q(z_n)} \right) \right]$$

multiply by $1 = \frac{q(z)}{q(z)}$ doesn't change value



$$= \mathbb{E}_{q(z_n)} \left[\log \frac{p(x_n, z_n)}{q(z_n)} + \log \frac{q(z_n)}{p(z_n|x_n)} \right]$$

because $\log ab = \log a + \log b$

$$= \mathbb{E}_{q(z_n)} \left[\log \frac{p(x_n, z_n)}{q(z_n)} \right] - \mathbb{E}_{q(z_n)} \left[\frac{p(z_n|x_n)}{q(z_n)} \right]$$

by linearity of expectations

$$= \mathbb{E}_{q(z)} \left[\log \frac{p(x_n, z_n)}{q(z_n)} \right] + \text{KL}(q(z) \| p(z|x))$$

That's it! Plus, we know $\text{KL} \geq 0$

always, so

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

we have a principled lower bound of the incomplete likelihood!

Let's define our lower bound as a function, recalling $q(z_n) = \text{Cat}(z_n/r_n)$

$$d(x_n, r_n, \pi, \mu, \Sigma) = \mathbb{E}_{q(z_n/r_n)} \left[\log \frac{p(x_n, z_n)}{q(z_n/r_n)} \right]$$

$$= \mathbb{E}_{q(z_n/r_n)} [\log p(x_n, z_n)] - \mathbb{E}_{q(z_n/r_n)} [\log q(z_n/r_n)]$$

expected complete log likelihood
entropy of $q(z_n/r_n)$

$$= \sum_k \left(r_{nk} \log \pi_k + r_{nk} \log \text{Norm PDF}(x_n / \mu_k, \Sigma_k^2) - r_{nk} \log r_{nk} \right)$$

That's it. We can easily evaluate d given data x_n

probability vector r_n
 weights π
 means μ
 variances Σ^2

New View of EM:

Optimizing an objective \mathcal{L}
that is lower bound of
incomplete likelihood

$$\log p(x|\pi, \mu, \sigma) \geq \sum_n \mathcal{L}(x_n, r_n, \pi, \mu, \sigma)$$

E step: Visit every example n ,
find $q(r_n)$ that
maximizes \mathcal{L}
given current params
 π, μ, σ

$$r_n = \operatorname{argmax}_{r_n \in \Delta^K} \mathcal{L}(x_n, r_n, \pi, \mu, \sigma)$$

M step: Find point estimate of
each param π, μ, σ
that maximizes whole dataset objective

$$\pi, \mu, \sigma = \operatorname{argmax}_{\pi, \mu, \sigma} \sum_n \mathcal{L}(x_n, r_n, \pi, \mu, \sigma)$$

We've said \mathcal{L} is a lower bound

How good is the bound?

Earlier, we saw

$$\log p(x_n) = \mathcal{L}(x_n, r_n, \pi, \mu, \Sigma) + \text{KL}(q(z) \parallel p(z|x))$$

KL always ≥ 0

Recall KL equals 0 ^{ONLY} when $q(z) = p(z|x)$

Earlier, we said

$$p(z_n|x_n) = \text{Cat}(\gamma_{n1}, \gamma_{n2}, \dots, \gamma_{nk})$$

$$\text{where } \gamma_{nk} = \frac{\pi_k \text{Norm}(x_n | \mu_k, \Sigma_k^2)}{\sum_l \pi_l \text{Norm}(x_n | \mu_l, \Sigma_l^2)}$$

So if we set $r_n = \gamma_n$, then KL term = 0.0 and bound is tight

Turns out, this is optimal E step update, $\log p(x_n) = \mathcal{L}(x_n, \gamma_n, \pi, \mu, \Sigma)$ with equality \triangle

Math behind E-step update

$$\max_{r_n \in \Delta^K} d(x_n, r_n, \pi, \mu, \Sigma^2)$$

$$\max_{r_n \in \Delta^K} \sum_k r_{nk} \omega_k - r_{nk} \log r_{nk}$$

\uparrow
Scaler

$$= \log \pi_k + \log \text{Norm}(x_n / \mu_k \Sigma_k^2)$$

Add Lagrange to make unconstrained

$$J = \sum_k r_{nk} \omega_k - r_{nk} \log r_{nk} + \lambda \left(1 - \sum_k r_{nk}\right)$$

Take gradient, set to zero, solve

$$\frac{\partial J}{\partial r_{n1}} = \omega_1 - 1 - \log r_{n1} - \lambda = 0 \quad (1)$$

$$\frac{\partial J}{\partial r_{nk}} = \omega_k - 1 - \log r_{nk} - \lambda = 0 \quad (k)$$

$$\frac{\partial J}{\partial \lambda} = 1 - \sum_k r_{nk} = 0 \quad (K+1)$$

add up eqns (1) ... (K)

$$\sum_k r_{nk} = (e^{-\lambda-1}) \sum_k e^{w_k} \quad \begin{array}{l} \text{plug into} \\ \text{(K+1)} \end{array}$$

$$1 - \sum_k r_{nk} = 1 - (e^{-\lambda-1}) \sum_k e^{w_k} = 0$$

solving...

$$1 = (e^{-\lambda-1}) \sum_k e^{w_k}$$

$$\lambda+1 = \log \sum_k e^{w_k}$$

sub into (1) ... (K):

$$r_{nk} = e^{-(\lambda+1)} e^{w_k} = \frac{e^{w_k}}{\sum_l e^{w_l}}$$

thus, optimal r_n is given by

$$r_{nk} = \frac{\pi_k \text{ NormPDF}(x_n / \mu_k, \sigma_k^2)}{\sum_l \pi_l \text{ NormPDF}(x_n / \mu_l, \sigma_l^2)}$$

Can we use EM to do
penalized likelihood maximization?
Yes!

Goal w/ incomplete likelihood

$$\min_{\theta} -\sum_n \log p(x_n | \theta) + \text{penalty}(\theta)$$

Goal w/ lower bound \downarrow

$$\min_{r_n, \theta} -\sum_n d(x_n, r_n, \theta) + \text{penalty}(\theta)$$

$$\text{E step: } r_n = \underset{r_n}{\text{argmax}} d(x_n, r_n, \theta)$$

$$\text{M step: } \theta = \underset{\theta}{\text{argmin}} -\sum_n d(x_n, r_n, \theta) + \text{penalty}(\theta)$$