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:

E step:

for n in 1, 2, ... N

for k in 1, 2, ... K:

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

Bishop 9.23

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

M step:

for k in 1, ... K:

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

9.26

M step

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

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

Big idea:

Coordinate ascent algorithm

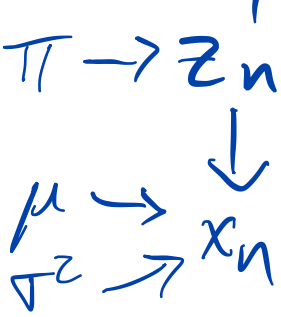
each substep updates subset of variables (coordinates)

Open Questions (goals ~~for today~~)

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

Step Back: Latent Variable Models

GMM w/ latent variables "z"



z_n is one-hot, indicates cluster assigned to example n

$z_n \in \text{onehot}(1 \ldots K)$

z_n is latent or "hidden" we cannot observe it directly

$$p(z_n) = \text{Cat}(\pi_1, \ldots \pi_K) = \prod_k \pi_k^{z_{nk}}$$

$$p(x_n | z_n) = \prod_{k=1}^{K} \text{NormPDF}(x_n | \mu_k, \sigma_k^2)^{z_{nk}}$$

remember, z_n is one hot, so only one term in each product will be used (rest are all 1)

anything to zero power is 1

$1 = \pi_k^0$

$1 = \text{NormPDF}^0$

Recall the posterior over z given x & parameters

$$p(z_n | x_n) = \text{CatPMF}(\gamma_{n1}, \gamma_{n2}, \ldots \gamma_{nK})$$

$$= \prod_{k=1}^{K} \gamma_{nk}^{z_{nk}}, \quad \text{where } \gamma_{nk} = \frac{\pi_k \, N(x_n | \mu_k, \sigma_k^2)}{\sum_l \pi_l \, N(x_n | \mu_l, \sigma_l^2)}$$

for each k in $1, \ldots K$

Now we have tools we need...

Idea 1: Complete likelihood easier to optimize

Incomplete likelihood aka marginal of x integrating away z

$$p(x) = \prod_n \text{GMMPDF}(x_n \mid \mu, \pi, \tau)$$

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

notice: π, μ, τ all interdependent

Complete likelihood aka joint of x and z eg $\mu_1, \mu_2, \dots \mu_K$ all intract

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

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

notice: π depends on z only

μ_k only depends on z_{nk}, x_n, τ_k

Key Idea: Complete likelihood easier to optimize

If only we "knew" z, we could estimate π, μ, τ, 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 $q(z_n) = Cat(r_{n1}, r_{n2}, \dots r_{nK})$

↳ $E_q[z_{nk}] = r_{nk}$

Suppose we wanted to compute $E_{q(z)}[\log p(x, z \mid \theta)]$, could we?

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

$$= \sum_n \sum_k E_q(z_{nk}) \log \pi_k + E_q\left[z_{nk} \log \text{NormalPDF}(x_n \mid \mu_k, \sigma_k^2)\right]$$

easy to optimize π_k, μ_k, σ_k if we know x and r

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

$$\log p(x_n) = \sum_{z_n} q(z_n) \log p(x_n)$$

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

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

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

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

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

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

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

That's it! Plus, we know $KL \geq 0$ 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, recall $q(z_n) = Cat(z_n | r_n)$

$$\mathcal{L}(x_n, r_n, \pi, \mu, \tau) = \mathbb{E}_{q(z_n|r_n)}\left[\log \frac{p(x_n, z_n)}{q(z_n|r_n)}\right]$$

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

expected complete log likelihood

entropy of $q(z_n|r_n)$

$$= \sum_k \left(r_{nk} \log \pi_k + r_{nk} \log \text{NormPDF}(x_n | \mu_k, \tau_k^2) - r_{nk} \log r_{nk} \right)$$

That's it. We can easily evaluate $\mathcal{L}$ given data x_n

probability vector r_n

weights π

means μ

variances τ^2

New View of EM:

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

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

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

$$r_n = \underset{r_n \in \Delta^K}{\operatorname{argmax}} \mathcal{L}(x_n, r_n, \pi, \mu, \Sigma^2)$$

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

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

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

How good is the bound?

Earlier, we saw

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

KL always ≥ 0

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

Earlier, we said

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

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

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

$\log p(x_n) = \mathcal{L}(x_n, \gamma_n, \pi, \mu, \tau)$ with equality

Turns out, this is optimal E step update.

Math behind E-step update

$$\max_{r_n \in \Delta^K} \mathcal{L}(x_n, r_n, \pi, \mu, \sigma^2)$$

$$= \log \pi_k + \log \text{Normal}(x_n | \mu_k, \sigma_k^2)$$

$$\max_{r_n \in \Delta^K} \sum_k r_{nk} \underbrace{w_k}_{\text{scalar}} - r_{nk} \log r_{nk}$$

Add Lagrange to make unconstrained

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

Take gradient, set to zero, solve

$$\frac{\partial J}{\partial r_{n1}} = w_1 - 1 - \log r_{n1} - \lambda = 0 \quad (1)$$

$$\frac{\partial J}{\partial r_{nK}} = w_K - 1 - \log r_{nK} - \lambda = 0 \quad (K)$$

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

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

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

plug into (K+1)

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

solve...

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

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

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 $\mathcal{L}$

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

E step: $r_n = \underset{r_n}{\operatorname{argmax}} \mathcal{L}(x_n, r_n, \theta)$

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