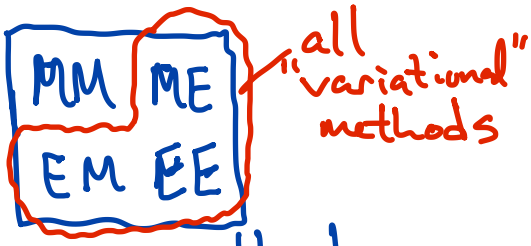
Statistical Pattern Recognition
Day 24 in 2021

From EM to Variational Inference

Readings: Bishop PRML Sec 10.1
" Sec 10.2

Outline



(1) Variational methods

(2) Overview of optimization methods for probabilistic models

(3) GMM case study: EM vs. EE side-by-side

(4) Choosing a family of approximate posteriors Q

Variational Methods

Consider any probabilistic model with random variables

X : observed data ← discrete or continuous

Z : hidden quantity ← discrete or continuous

e.g. means in GMM, weights in linear regression

The model defines joint $p(X, Z)$.

Our goal is to estimate posterior $p(Z|X)$ so we can { - evaluate PDF, - draw samples }

Usually, this is difficult due to intractable sum/integral

$$p(z|x) = \frac{p(x,z)}{p(x)}$$

known by definition of model

difficult $p(x) = \int p(x,z)\,dz$

Solution: Find a good approximation among an easier to work with set of distributions Q over the domain of r.v. Z.

Define set Q of easy-to-use distributions $q_\omega(z|\omega)$ with possible parameters $\omega \in Q$

We'll assume we can easily evaluate $q(z|\omega)$ as a PDF.

We want to find parameter ω^* that makes $q(z|\omega^*) \approx p(z|x)$

$$\omega^* = \min_{\omega \in Q} KL(q(z|\omega) \,||\, p(z|x))$$

How can we solve this?

Turns out, the following are equivalent optimization problems:

$$\arg\min_{w \in Q} \; KL\left(q(z|w) \,\|\, p(z|x)\right)$$

$$\arg\max_{w \in Q} \; \underbrace{\mathbb{E}_{q(z|w)}\left[\log p(x,z) - \log q(z|w)\right]}_{\text{evidence lower bound ("ELBO")}}$$

$$\mathcal{L}(x, w) \leq \log p(x)$$

with equality iff $q(z|w) = p(z|x)$

Thus, we can solve the tractable problem

$$w^* = \arg\max_{w \in Q} \mathcal{L}(x, w)$$

using our usual optimization toolkit (coordinate ascent, gradient ascent)

And use $q_{(z|w^*)}$ as an "approximate posterior" however we might use $p(z|x)$

Proof the two problems are equivalent:

$$\log p(x) = \int q_\phi(z) \log p(x) dz$$

$$= \mathbb{E}_{q_\phi(z)}[\log p(x)]$$ defn of expectation

$$= \mathbb{E}_{q_\phi(z)}\left[\log \frac{p(x,z)}{p(z|x)}\right]$$ Bayes rule

$$= \mathbb{E}_{q_\phi(z)}\left[\log \frac{p(x,z)}{p(z|x)} \frac{q_\phi(z)}{q_\phi(z)}\right]$$ multiply inside by $1 = \frac{q}{q}$

$$= \mathbb{E}_{q_\phi(z)}\left[\log \frac{p(x,z)}{q_\phi(z)}\right] + \mathbb{E}_{q_\phi(z)}\left[\log \frac{q_\phi(z)}{p(z|x)}\right]$$

$$= \mathcal{L}(x, w) + KL\left(q_\phi(z|w) \,\|\, p(z|x)\right)$$

always ≥ 0 for any z, discrete or continuous

$$\geq \mathcal{L}(x, w)$$

When sum of $a(w)$ and $b(w)$ is fixed, maximizing $a(w)$ is equivalent to minimizing $b(w)$ for any functions a and b.

Overview of Optimization Strategies for Probabilistic Models

Assume a model $p(x, z, \theta)$ with:

- x : observed data
- z : hidden variable
- θ : hidden variable

how to treat θ?

how to treat z?	$\theta \in \Theta$ point estimate	$q(\theta \mid \lambda)$ for $\lambda \in \Lambda$ approx. posterior
point estimate $z \in \Omega$	**M / M** $\underset{\substack{z \in \Omega \\ \theta \in \Theta}}{\operatorname{argmax}} \log p(x, z, \theta)$	**M / E** $\underset{\substack{z \in \Omega \\ \lambda \in \Lambda}}{\operatorname{argmax}} \mathbb{E}_{q(\theta \mid \lambda)}\left[\log \frac{p(x, z, \theta)}{q(\theta \mid \lambda)}\right]$
approx posterior $q(z \mid r)$ $r \in R$	**E / M** $\underset{\substack{r \in R \\ \theta \in \Theta}}{\operatorname{argmax}} \mathbb{E}_{q(z \mid r)}\left[\log \frac{p(x, z, \theta)}{q(z \mid r)}\right]$	**E / E** $\underset{\substack{r \in R \\ \lambda \in \Lambda}}{\operatorname{argmax}} \mathbb{E}\left[\log \frac{p(x, z, \theta)}{q(z \mid r) q(\theta \mid \lambda)}\right]$

EM is one of many possible variational approaches

all ☐ use variational objectives based on ELBO

EM vs EE side by side

(simpler version of PRML 10.2)

GMM with fixed variance σ^2

$$p(x, z, \pi, \mu) = \underbrace{\mathrm{Dir}(\pi \mid a_0 \cdots a_0) \prod_{k=1}^{K} \mathcal{N}(\mu_k \mid m_0, s_0^2)}_{\text{E/M}} \underbrace{\prod_{n=1}^{N} \left[\mathrm{Cat}(z_n \mid \pi)\, \mathcal{N}(x_n \mid \mu_{z_n}, \sigma^2)\right]}_{\text{E/E}}$$

	E/M	E/E
Update to Z or $q(z \mid r)$ $= \mathrm{Cat}(r_1 \cdots r_K)$	$\tilde{r}_{nk} = \pi_k \mathcal{N}(x_n \mid \mu_k, \sigma^2)$ $r_{nk} = \dfrac{\tilde{r}_{nk}}{\sum_{l=1}^{K} \tilde{r}_{nk}}$	$\log \tilde{r}_{nk} = \mathbb{E}_q[\log \pi_k] + \mathbb{E}_q[\log p(x_n \mid \mu_k, \sigma^2)]$ $r_{nk} = \dfrac{\tilde{r}_{nk}}{\sum_{l=1}^{K} \tilde{r}_{nk}}$
Update to μ or $q(\mu_k)$ $= \mathcal{N}(\mu_k \mid m_k, s_k)$	$\mu_k = \dfrac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$ ML estimate adjust for MAP	$q(\mu_k) = \mathcal{N}(m_k, s_k^2)$ $m_k^* = f(m_0, s_0, \sum_n r_{nk} x_n)$ see 10.61 $s_k^* = f(m_0, s_0, \underbrace{\sum_n r_{nk}}_{\text{suff. stat}})$ see 10.62 prior
Update to π or $q(\pi) = \mathrm{Dir}(a_1 \cdots a_K)$	$\pi_k = \dfrac{\sum_n r_{nk}}{N} = \dfrac{N_k}{N}$ ML estimate adjust for MAP $\pi_k = \dfrac{N_k + a_0 - 1}{N + K a_0 - K}$	$q(\pi) = \mathrm{Dir}(a_1, \cdots a_K)$ $a_k^* = a_0 + \underbrace{\sum_n r_{nk}}_{\text{sufficient statistic}}$ prior term See 10.58

Side-by-Side view in terms of ELBO objective

Where do updates come from?

All variational methods set up an objective function to maximize: the evidence lower bound (ELBO).

If we do coordinate ascent and try to optimally solve at each "coordinate"'s update, we get the updates on previous page.

	E/M	E/E
ELBO objective	$\mathcal{L}(r, \pi, \mu) = \mathbb{E}_{q_\beta(z\|r)}\left[\log \frac{p(x,z\|\pi,\mu)}{q_\beta(z\|r)} p(\pi,\mu)\right]$	$\mathcal{L}(r, a, m, s) = \mathbb{E}_{\substack{q_\beta(z\|r) \\ q_\beta(\pi\|a) \\ q_\beta(\mu\|m,s)}}\left[\log \frac{p(x,z\|\pi,\mu)p(\pi,\mu)}{q_\beta(z) q_\beta(\pi) q_\beta(\mu)}\right]$
$z / q_\beta(z)$ update	$\max_r \mathcal{L}(r, \pi^{t-1}, \mu^{t-1})$	$\max_r \mathcal{L}(r, a^{t-1}, m^{t-1}, s^{t-1})$
$\pi / q_\beta(\pi)$ update	$\max_\pi \mathcal{L}(r^t, \pi, \mu^{t-1})$	$\max_a \mathcal{L}(r^t, a, m^{t-1}, s^{t-1})$
$\mu / q_\beta(\mu)$ update	$\max_\mu \mathcal{L}(r^t, \pi^t, \mu)$	$\max_{m,s} \mathcal{L}(r^t, a^t, m, s)$

Note that we can use any algorithm we like to optimize our chosen objective function.

Given an ELBO optimization problem, the common approach (e.g. EM for GMMs) is to use coordinate ascent with closed-form updates. When possible, this is likely to be best choice (see CP4).

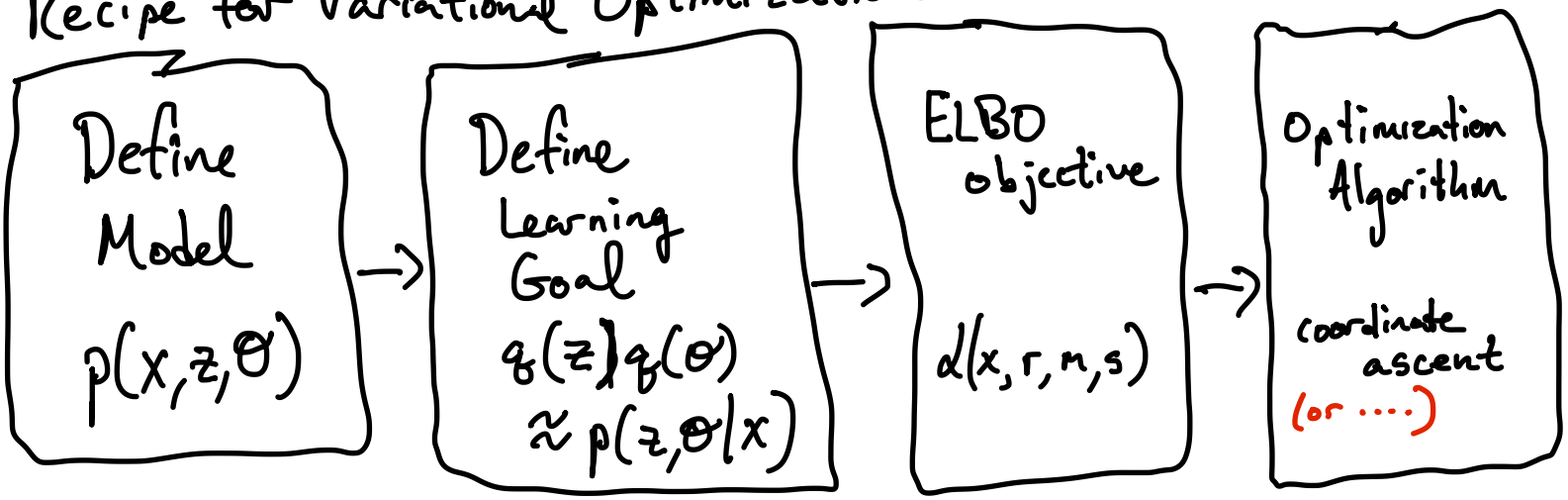
However, we could:

(1) do coordinate updates that just step uphill (esp. if closed-form optima unknown)

(2) do gradient ascent for all variables simultaneously

(3) use any other algorithm to optimize

Recipe for Variational Optimization



How to select the "approximate posterior" set of possible distributions Q?

Usually our goal is to balance accurate approximation with tractability.

- accurate approximation: ideally, $q(z) = p(z|x)$ "perfect" approximation
- tractability:
  - need $q_\phi(z)$ easy to evaluate
  - need ELBO easy to evaluate
  - need ELBO easy to optimize

Common strategies

(1) Impose independence assumptions on $q_\phi(z_1, \dots, z_V)$

Full Independence: $q_\phi(z_{1:V}) = q_\phi(z_1) q_\phi(z_2) \cdots q_\phi(z_V)$

for historical reasons, often called "mean field"

Some Structured Independence: $q_\phi(z_{1:V}) = q_\phi(z_1) \prod_{v=2}^{V} q_\phi(z_v | z_1)$

(2) Take advantage of conjugate model structure

e.g. in GMM, prior on π is $Dir(a_0, \dots a_0)$

make posterior on π also $Dir(a_1, \dots a_K)$

(3) To get tractable continuous r.v. density, use a Gaussian

either diagonal or low rank or full rank covariance matrix