4. Variational method. EM algorithm as a variational method

Variational Bayes.

Def. KL-divergence: $D_{KL}(Q \| P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$;

$D_{KL}(Q\|P) \geq 0$, $D_{KL}(Q\|P) = 0 \Leftrightarrow Q = P$. $D_{KL}(Q\|P) \neq D_{KL}(P\|Q)$.

~~Suppose we are to derive the posterior~~ $P(\Theta \mid \{x_n\}_{n=1}^N)$

~~Given~~ prior $p(\Theta)$.

Suppose we have an assumed model $\mathcal{H}$, a set of data points $\{x^{(n)}\}_{n=1}^N$, a set of latent variables $\{z^{(n)}\}_{n=1}^N$ and a set of parameters Θ.

(e.g., $z^{(n)}$ is a class label in k-means)

Our task is to get the best (MAP) parameters Θ_{ML}, given X and $\mathcal{H}$.

$\left(\prod_{i=1}^n \pi_{z_i} N(x_i \mid \mu_{z_i}, \sigma^2_{z_i}) \text{ e.g.} \right)$

~~log~~

1) If we know $p(X, Z \mid \Theta, \mathcal{H})$, we can do the following:

$$\log p(X|\Theta,\mathcal{H}) = \int q(Z) \log p(X|\Theta,\mathcal{H})\, dZ =$$

$$= \int q(Z) \log \frac{p(X,Z|\Theta,\mathcal{H})}{q(Z)} dZ + \int q(Z) \log \frac{q(Z)}{p(Z|X,\Theta,\mathcal{H})} dZ =$$

$$\cancel{= D_{KL}(q \| p(X,Z|\Theta,\mathcal{H})) +}$$

$$= \underbrace{\int q(Z) \log \frac{p(X,Z|\Theta,\mathcal{H})}{q(Z)} dZ}_{\substack{\text{Variational lower bound} \\ \equiv \\ \mathcal{L}_V(q,\Theta)}} + \underbrace{D_{KL}(q \| p(Z|X,\Theta,\mathcal{H}))}_{\geq 0}$$

2) As a result, instead of solving

$$\mathcal{L}(\Theta) \equiv \log p(X|\Theta,\mathcal{H}) \longrightarrow \max_\Theta$$

we solve $\mathcal{L}_V(q, \Theta) \longrightarrow \max\limits_{q, \Theta}$

by moving step-by-step on q and Θ:

(M) $\Theta^{(t)} = \arg\max\limits_{\Theta} \mathcal{L}_V(q^{(t-1)}, \Theta^{(t-1)}) = \arg\max\limits_{\Theta} \mathbb{E} \log p(X, Z | \Theta, H)$

(E) $q^{(t)} = \arg\max\limits_{q} \mathcal{L}_V(q, \Theta^{(t)}) = p(Z | X, \Theta^{(t)}, H)$

$\quad \left(\arg\min\limits_{q} D_{KL}(q, p(Z|X, \Theta^{(t)}, H))\right)$

and we get EM algorithm ~~(in general)~~ (as a general idea)

3) If all $z_i \in \{1, \dots, k\}$, then:

$$p(x_i | \Theta) = \sum_{k=1}^{K} p(x_i | k, \Theta) \, p(z_i = k | \Theta),$$

and E-step becomes:

$$q^{(t)}(z_i = k) = \frac{p(x_i | k, \Theta^{(t)}) \, p(z_i = k | \Theta^{(t)})}{\sum_{l=1}^{K} p(x_i | l, \Theta^{(t)}) \, p(z_i = l | \Theta^{(t)})},$$

while M-step is: $\Theta^{(t)} = \sum_{i=1}^{n} \sum_{k=1}^{K} q^{(t-1)}(z_i = k) \log p(x_i, k | \Theta^{(t-1)})$.

4) if z_i is continuous, E-step can be done in closed form only in case of conjugate distributions.

(because we compute $q^{(t)}(z_i) = \dfrac{p(x_i | z_i, \Theta^{(t)}) \, p(z_i | \Theta^{(t)})}{\int p(x_i | z, \Theta) p(z | \Theta) \, dz}$)

5) To deal with non-conjugate priors we use Variational Bayes:

instead of $q^{(t)} = \arg\min\limits_{q} D_{KL}(q \| p(Z | X, \Theta^{(t)}, H))$

we approximate it with $q^{(t)} = \arg\min\limits_{q \in Q} D_{KL}(q \| p(Z | X, \Theta^{(t)}, H))$

$\qquad\qquad = \arg\max\limits_{q \in Q} \mathcal{L}_V(q, \Theta^{(t)})$.

Inference becomes optimization

$$\hat{\mathcal{L}}_V(q, \Theta^{(t)}) = \int q(Z | \varphi) \log \frac{p(X, Z | \Theta)}{q(Z | \varphi)} \, dZ \to \max_{\varphi} \quad \text{- Variational E-step.}$$

$$= \hat{\mathcal{L}}_V(\varphi, \Theta^{(t)})$$

~~8~~

~~6) In classical EM:~~

6) Variational Bayes with classical EM:

$$\varphi_n = \arg\max_{\varphi} \hat{\mathcal{L}}_V(\varphi, \theta_{n-1}), \text{ E-step}$$

$$\theta_n = \arg\max_{\theta} \hat{\mathcal{L}}_V(\varphi_n, \theta), \text{ M-step}$$

If we cannot solve this problems analytically we'd better use:

$$\varphi_n = \varphi_{n-1} + \eta \nabla_\varphi \hat{\mathcal{L}}_V(\varphi, \theta_{n-1}), \quad \text{Variational E-step}$$

$$\theta_n = \theta_{n-1} + \varepsilon \nabla_\theta \hat{\mathcal{L}}_V(\varphi_n, \theta), \quad \text{Variational M-step}$$

We can even use the stochastic gradients instead of the full. ~~true~~ $\hat{\mathcal{L}}_V$ - ELBO (Evidence Lower-BOund)

ELBO is often intractable, but:

$$\log p(x, z|\theta) = \sum_{i=1}^{n} \log p(x_i, z_i|\theta) = \sum_{i=1}^{n} (\log p(x_i|z_i, \theta) + \log p(z_i|\theta)),$$

if $\{x_i, \hat{z_i}\}_{i=1}$ is i.i.d. $\Rightarrow$ can use mini-batching for unbiased gradient estimations.

<u>Example</u> (Ada-Gram)

Word2Vec with Hierarchical softmax:

$$p(y|x, \theta) = \prod_{c \in Path(y)} \sigma(d_{c,y} In(x)^T Out(c)).$$

$$p(y|x, \theta) \to \max_{\theta}, \quad \theta = [In, Out],$$

$Path(y)$, $d_{c,y}$ are obtained from the Huffman tree, build for our dictionary.

How to account different meanings of the same word?
Let's define a latent variable z_i that indicates meaning of x_i:

$$p(y_i|x_i, z_i, \theta) = \prod_{c \in Path(y_i)} \sigma\left(d_{c,y_i} \, In(x_i, z_i)^T Out(c)\right),$$

$$p(z_i = k|x_i) = \frac{1}{K(x_i)}, \quad K(x_i) \text{ - total number of meanings for } x_i.$$

Now we can use a standard EM-algorithm with discrete latent variables (as in 3))...
But that we'll require a lot of time.
To make it scalable - use Variational EM:

M-step:

$$\nabla_\theta \mathbb{E}_Z \log p(Y, Z|X, \theta) = \nabla_\theta \mathbb{E}_Z \sum_{i=1}^{n} \left(\log p(y_i|z_i, x_i, \theta) + \log p(z_i|x_i)\right) =$$

$$= \sum_{i=1}^{n} \mathbb{E}_{z_i}\left(\nabla_\theta \log p(y_i|z_i, x_i, \theta) + \underbrace{\nabla_\theta \log p(z_i|x_i)}_{\text{doesn't depend on } \theta}\right) =$$

$$= \sum_{i=1}^{n} \mathbb{E}_{z_i}\left(\nabla_\theta \log p(y_i|z_i, x_i, \theta)\right)$$

Its unbiased estimate is $n \cdot \sum_{j=1}^{K(x_i)} \underbrace{p(z_i = k|y_i, x_i, \theta)}_{\text{we know from E-step}} \nabla_\theta \log p(y_i|k, x_i, \theta)$

~~During M-step, we update~~
During E-step we update probabilities only for single x_i.
To deal with different # of meanings for each x_i use DP:

$$p(z = k|x_i, \vec{\beta}) = \beta_{ik} \prod_{r=1}^{k-1} (1 - \beta_{ir}), \quad p(\beta_{ik}|\alpha) = Beta(\beta_{ik}|1, \alpha)$$

$$\Downarrow$$

$$p(Y, Z, \beta|X, \alpha, \theta) = \prod_{w=1}^{|V|} \prod_{k=1}^{\infty} p(\beta_{wk}|\alpha) \prod_{i=1}^{n} \left[p(z_i|x_i, \vec{\beta}) \prod_{j=1}^{C} p(y_{ij}|z_i, x_i, \theta)\right]$$