Variational inference

Consider a prob. dist over observed + latent variables

$$p(x, z) = p(z)\, p(x \mid z).$$

The inference problem is to compute the posterior

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

$$= \frac{p(x, z)}{\int_z p(x, z)\, dz} \quad \leftsquigarrow \text{ this is pretty hard!}$$

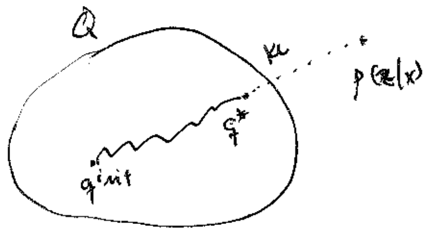
Variational inference :

Optimize. **First** : Assume there is a family of distributions Q over z.

Second : Find $q \in Q$ so that

$$q^*(z) = \underset{q \in Q}{\operatorname{argmin}}\ KL\big(q(z) \,\|\, p(z \mid x)\big)$$

Third : Approximate the posterior with $q^*(\cdot)$.



Notation $p(x)$ is also known as the evidence. $\quad p(x) = \int_z p(z, x)\, dx.$

The example we will use this entire note is that of a mixture of Gaussians i.e. a GMM.

To elaborate :

a Gaussian mixture model has K mixture components, each cluster a Gaussian w/ mean μ_k, $k=1,\dots,K$.

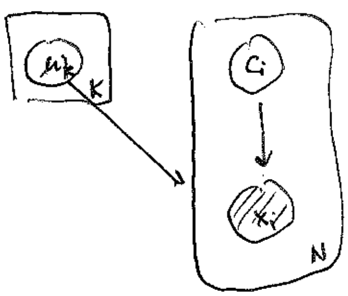
By assumption, each mean is sampled $\mu_k \sim p(\mu_k) = N(0, \sigma^2)$

$\uparrow$ hyperparameter.

As a generative model, a GMM is

$$\boxed{\begin{aligned} &\mu_k \sim N(0,\sigma^2) \\ &c_i \sim \text{Categorical}\left(\tfrac{1}{K},\dots,\tfrac{1}{K}\right) = \text{Uniform}(K) \\ &x_i \mid c_i, \mu \sim N(c_i^T \mu, 1) \end{aligned}}$$

As a probabilistic graphical network



where the boxes are plates.

$$p(\mu, c, x) = p(\mu) \prod_{i=1}^{n} p(c_i)\, p(x_i \mid c_i, \mu)$$

Here, $z = \{\mu, c\}$ are the latent variables.

What is the evidence?

$$p(x) = \int p(\mu) \prod_{i=1}^{n} \sum_{c_i} p(c_i) p(x_i \mid c_i, \mu)\, d\mu$$

$$= \sum_{c_i} p(c_i) \int p(\mu) \prod_{i=1}^{n} p(x_i \mid c_i, \mu)\, d\mu$$

Oh god.

So $p(x)$ is hard, $p(z|x)$ is hard. lets try VI.

We want

$$q^*(z) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \underbrace{KL(q(z) \,\|\, p(z|x))}_{\text{So what's this?}}.$$

$$KL(q(z) \,\|\, p(z|x)) = \mathbb{E}_{z \sim q(z)}\left[\log q(z) - \log p(z|x)\right]$$

$$= \underbrace{\mathbb{E}_{z\sim q(z)}[\log q(z)]}_{\text{yes}} - \underbrace{\mathbb{E}_{z\sim q(z)}[\log p(x,z)]}_{\text{yes}} + \underbrace{\log p(x)}_{\text{NO}}.$$

What's the point if we gotta compute $p(x)$ anyway !?

KEY IDEA : But note that KL divergence is always positive !

$$0 \leq \mathbb{E}_{z\sim q(z)}[\log q(z)] - \mathbb{E}_{z\sim q(z)}[\log p(x,z)] + \log p(x)$$

So rearranging we get

$$\log p(x) \geq \underbrace{\mathbb{E}_{z\sim q(z)}[\log p(x,z)] - \mathbb{E}_{z\sim q(z)}[\log q(z)]}_{\text{ELBO}(q)}$$

"evidence lower bound"

KEY: Maximizing this = Minimizing KL

This is totally tractible. We will seek to optimize this !

Aside The ELBO can be written as

$$\mathrm{ELBO}(q) = \mathbb{E}_{z\sim q(z)}[\log p(x,z)] - \mathbb{E}_{z\sim q(z)}[\log q(z)]$$

$$= \mathbb{E}_{z\sim q(z)}[\log p(x|z)] - \mathbb{E}_{z\sim q(z)}[\log p(z)] - \mathbb{E}_{z\sim q(z)}[\log q(z)]$$

$$\boxed{\mathrm{ELBO}(q) = \underbrace{\mathbb{E}_{z\sim q(z|x)}[\log p(x|z)]}_{\text{reconstruction error}} - KL\left(q(z)\,\|\,p(z)\right).}$$

Writing it this way expresses z (latent variables) as a "code" with encoder given by $q(z)$ and $p(x|z)$ an "decoder".

Maximizing the ELBO means we want to maximize the log likelihood of getting the original representation back, while if given a good prior, penalizing representations that merely copy the data.

This gives the theoretical backing behind <u>variational autoencoders</u>.

Now, $\log p(x) \geq \mathrm{ELBO}(q)$ for any $q(z)$.

<u>ex)</u> What is the ELBO in the GMM model?

To solve this question, first we need a variational family to pull q from. Note that the complexity of the family determines the complexity of the optimization.

Usually we impose stronger independence relations for distributions in Q, as these might be more computable.

A common family to use is the <u>mean-field variational family</u>.

$$q(z) = \prod_{j=1}^{m} q_j(z_j) \quad , \quad m = \# \text{ of latent variables}.$$

In the mean-field variational family, the latent variables are decoupled.

What are the q_j then? Depends on the latent variable.

If z_j is ~~[illegible]~~ $\begin{cases} \text{discrete} \rightsquigarrow \text{categorical, multinomial, ...} \\ \text{continuous} \rightsquigarrow \text{Gaussian, ...} \end{cases}$

<u>Note</u> If we add dependencies between the variables, we get families for <u>structured variational inference</u>.

In the GMM model, the mean-field VF comes as the form

$$q(\mu, c) = \prod_{k=1}^{K} \underbrace{q(\mu_k; m_k, s_k^2)}_{\text{Gaussians}} \prod_{i=1}^{n} \underbrace{q(c_i; \varphi_i)}_{\text{distribution of } x_i\text{'s mixture assignment}}$$

Gaussians $\mu_k \sim N(\mu_k; m_k, s_k^2)$

distribution of x_i's mixture assignment. φ_i = K-vector of assignment probabilities.

The total parameter space is given by

$$\{ m_k, s_k^2 : k=1,..,K, \ \varphi_i \in \mathbb{R}^K ; i=1,..,N \}$$

$$q_{\{m_k, s_k^2, \varphi_i\}}(\mu, c) \in \mathcal{Q}_{\text{mean-field}}.$$

The ELBO here is thus given by

$$\boxed{\begin{aligned} \text{ELBO}(q_{\{m_k, s_k^2, \varphi_i\}}) = & \sum_{k=1}^{K} \mathbb{E}_{z \sim q(z)}[\log p(\mu_k); m_k, s_k^2] \\ & + \sum_{i=1}^{n} \Big(\mathbb{E}[\log p(c_i); \varphi_i] + \mathbb{E}[\log p(x_i | c_i, \mu); \varphi_i, m, s^2] \Big) \\ & - \sum_{i=1}^{n} \mathbb{E}[\log q(c_i; \varphi_i)] - \sum_{k=1}^{K} \mathbb{E}[\log q(\mu_k; m_k, s_k^2)]. \end{aligned}}$$

Now to optimize. We use a strategy called <u>coordinate-ascent mean-field variational inference</u> (CAVI).

CAVI iteratively optimizes each factor of the mean-field variational density, while holding the others fixed.

<u>How to do it</u>: let z_j be a latent variable.

The <u>complete conditional</u> of z_j is $p(z_j \mid z_{\neq j}, x)$

Fix the other variational factors $q_\ell(z_\ell)$, $\ell \neq j$.

We want the optimal $q_j(z_j)$, called $q_j^*(z_j)$ to maximize ELBO.

<u>Claim</u>: $q_j^*(z_j) \propto \exp\left(\mathbb{E}_{z_{\neq j} \sim q(z)}\left[\log p(z_j \mid z_{\neq j}, x)\right]\right)$

<u>Proof</u>: $\mathrm{ELBO}(q) = \mathbb{E}[\log p(z,x)] - \mathbb{E}[\log q(z)]$

$$= \mathbb{E}_{z_j \sim q_j(z_j)}\left[\mathbb{E}_{q_{\neq j}(z)}\left[\log p(z_j, z_{\neq j}, x)\right]\right] - \mathbb{E}_{z_j \sim q_j(z_j)}\left[\log q_j(z_j)\right] + \text{const.}$$

$$= \text{const} - \underbrace{KL\left(q_j(z_j) \,\|\, \exp\left(\mathbb{E}_{z_{\neq j} \sim q(z)}\left[\log p(z_j, z_{\neq j}, x)\right]\right)\right)}_{\text{minimize this}}$$

So take $q_j^*(z_j) \propto \exp\left(\mathbb{E}_{z_{\neq j} \sim q(z)}\left[\log p(z_j \mid z_{\neq j}, x)\right]\right)$. $\square$.

ex) What are the optimal updates then for the GMM?

- $q^*(c_i; \varphi_i) \propto \exp\Big(\underbrace{\log p(c_i)}_{\log \frac{1}{K}} + \mathbb{E}_{\{m_k, s_k^2\}}\left[\log p(x_i \mid c_i, \mu); m, s^2\right]$

$$p(x_i \mid c_i, \mu) = \prod_{k=1}^{K} p(x_i \mid \mu_k)^{c_{ik}}$$ (as c_i is a one-hot vector)

$$\mathbb{E}_{\{m_k, s_k^2\}}[\log p(x_i|c_i,\mu)] = \sum_{k=1}^{K} c_{ik}\, \mathbb{E}[\log p(x_i|\mu_k)]$$

$$= \sum_{k=1}^{K} c_{ik}\, \mathbb{E}[-(x_i-\mu_k)^2/2\ ;\ m_k, s_k^2] + \text{const}$$

$$= \sum_{k=1}^{K} c_{ik}\left(\mathbb{E}[\mu_k\ ;\ m_k, s_k^2]x_i - \mathbb{E}[\mu_k^2\ ;\ m_k, s_k^2]/2\right) + \text{const}$$

ie the variational update for the i^{th} cluster assignment is

$$\boxed{\varphi_{ik} \propto \exp\left(\mathbb{E}[\mu_k\ ;\ m_k, s_k^2]x_i - \mathbb{E}[\mu_k^2\ ;\ m_k, s_k^2]/2\right)}$$

- $q^*(\mu_k; m_k, s_k^2) \propto \exp\left(\log p(\mu_k) + \sum_{i=1}^{n} \mathbb{E}[\log p(x_i|c_i,\mu)\ ;\ \varphi_i, m_{\neq k}, s_{\neq k}^2]\right)$

Now, as $\varphi_{ik} = \mathbb{E}_{\varphi_i}[c_{ik}]$ (c_{ik} is an indicator random variable),

$$\log q^*(\mu_k; m_k, s_k^2) = \log p(\mu_k) + \sum_{i=1}^{n} \mathbb{E}[\log p(x_i|c_i,\mu); \varphi_i, m_{\neq k}, s_{\neq k}^2] + \text{const}$$

$$= -\mu_k^2/2\sigma^2 + \sum_{i=1}^{n} \mathbb{E}[c_{ik}] \log p(x_i|\mu_k) + \text{const}$$

$$= -\mu_k^2/2\sigma^2 + \sum_{i=1}^{n} \varphi_{ik} \cdot \left(-(x_i-\mu_k)^2/2\right) + \text{const}$$

$$= \left(\sum_{i=1}^{n} \varphi_{ik} x_i\right)\mu_k - \left(\frac{1}{2\sigma^2} + \sum_{i=1}^{n} \varphi_{ik}/2\right)\mu_k^2 + \text{const.}$$

ie $q^*(\mu_k; m_k, s_k^2)$ is a Gaussian   sufficient statistics.

Normalizing, we get the updates are

$$\boxed{m_k = \frac{\sum_{i=1}^{n} \varphi_{ik} x_i}{1/\sigma^2 + \sum_{i=1}^{n} \varphi_{ik}}, \quad s_k^2 = \frac{1}{\frac{1}{\sigma^2} + \sum_{i=1}^{n} \varphi_{ik}}}$$

$$\log p(\mu, c, x) = \sum_{k=1}^{K} \log p(\mu_k) + \sum_{i=1}^{n} \Big(\underbrace{\log p(c_i)}_{-\log K} + \log p(x_i \mid c_i, \mu) \Big)$$

$\mu_k \sim N(0, \sigma^2)$

$x_i \mid c_i, \mu \sim N(c_i^T \mu, 1)$

$$\propto \sum_{k=1}^{K} \left(-\frac{\mu_k^2}{2\sigma^2} \right) + \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \left(-\frac{(x_i - \mu_k)^2}{2} \right)$$

$$\log p(\mu, c, x) \propto \sum_{k=1}^{K} \left(-\frac{\mu_k^2}{2\sigma^2} \right) + \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \left(-\frac{(x_i - \mu_k)^2}{2} \right).$$

$$\mathbb{E}_{z \sim q(z)} [\log p(\mu, c, x)] = \sum_{k=1}^{K} -\frac{1}{2\sigma^2} \mathbb{E}_q [\mu_k^2] + \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}(c_{ik}) \left(-\frac{(x_i - \mu_k)^2}{2} \right)$$

$$= \sum_{k=1}^{K} -\frac{(m_k^2 + s_k^2)}{2\sigma^2} - \sum_{i=1}^{n} \sum_{k=1}^{K} \varphi_{ik} \mathbb{E}_q \left[\frac{(x_i - \mu_k)^2}{2} \right]$$

$$\mathbb{E}_{z \sim q(z)} [\log q(m, s^2, \varphi)] \propto \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_q \log \varphi_{ik} + \sum_{k=1}^{K} \mathbb{E}_q \left(-\frac{1}{2} \log s_k^2 - \frac{(\mu_k - m_k)^2}{2 s_k^2} \right)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \log \varphi_{ik} - \sum_{k=1}^{K} \Big(\frac{1}{2} \log s_k^2 + \underbrace{\frac{1}{2}}_{\text{constant.}} \Big)$$

$$\propto \sum_{i=1}^{n} \sum_{k=1}^{K} \log \varphi_{ik} - \frac{1}{2} \sum_{k=1}^{K} \log s_k^2 .$$

- $\mathbb{E}_q [(x_i - \mu_k)^2] = x_i^2 - 2 x_i m_k + (m_k^2 + s_k^2)$
- $\log p(x_i \mid c_i, \mu) = \log N(x_i \mid c_i^T \mu, 1) = -\frac{1}{2} \left(x_i - \sum_{k=1}^{K} c_{ik} \mu_k \right)$