Tutorial 8

Theory review → what I learnt from the Tutorial

1. Variational Bayesian Methods

- parameters
- latent variables

unobserved var.

data

observed var

→ aim: provide approximation of posterior of unobserved variables

→ derive elbo → lower bound of marginal likelyhood.

(ELBO → evidence lower bound)

$p(Z|D)$ – not a "close form"

→ we don't know mean, variance

→ we cannot sample

In general:

→ directed graphical models with $\underbrace{\text{observed variables } X_{\mathbb{R}^N} \text{ and latent } Z_{\mathbb{R}^D}}$

→ we have access to joined distribution $p(x,z)$, we observe X as D (data)

e.g. X - images, Z - camera angle, lighting

task:

→ Use "inference" given X, estimate latent var. Z

→ look @ a couple of picture and estimate z.

Using Baye's rule

$$p(z|X=D) = \frac{\overbrace{p(X=D|z)}^{\text{likelihood of data}} \cdot \overbrace{p(z)}^{\text{prior}}}{\underbrace{p(X=D)}_{\text{Marginal}}}$$

Marginal → problem is we often do not know the normalization cnst. given by $p(x=D)$

$$p(x) = \int_{z_0} \dots \int_{z_{n-1}} p(x,z)\, dz_0 \dots dz_{n-1}$$

→ this is intractable → incomputable!

→ complex models, e.g. images where we want to find a close form for 10^4 dimensions (a lot) is computable infeasable (in general)

Solution:

Using surogate → Gaussian, we can capture most of important info

→ $q(z) \simeq p(z|X=D)$ as "good" as possible

→ "variation" optimize for a function

Metric

→ KL - divergance → distance between 2 distribution: → identical: 0; → diffrent: ∞

Optimization problem

$$q^*(z) = \underset{q(z)\in Q_{\text{simple distrib e.g. gauss}}}{\arg\min} \left(KL\left(q(z) \| p(z|\underbrace{D}_{X})\right)\right)$$

$$KL\left(q(z)\,\|\,p(z|D)\right) = \mathbb{E}_{z\sim q(z)}\left[\log \frac{q(z)}{p(z|D)}\right] = \int_{z_0}\cdots\int_{z_{D-1}} q(z)\log\frac{q(z)}{q(z|D)}\, dz_0\ldots dz_{D-1}$$

→ problem, we don't have the posterior, but we have joint $p(z,D)$

$$\cong \int_z q(z)\log\left(\frac{q(z)}{\underset{\frac{p(z,D)}{p(D)}}{\underset{\parallel}{p(z|D)}}}\right)dz = \int_z q(z)\log\left(\frac{q(z)\,p(D)}{p(z,D)}\right)dz = \underbrace{\int_z q(z)\log\left(\frac{q(z)}{p(z,D)}\right)dz}_{\text{expectation}}$$

$$+ \underbrace{\int_z q(z)\log p(D)\, dz}_{\text{expectation}} = \mathbb{E}_{z\sim q(z)} \log\frac{q(z)}{p(z,D)} + \mathbb{E}_{z\sim q(z)} \log p(D) \overset{\text{swap denominator}}{=} \underbrace{-\mathbb{E}_q \log\frac{p(z,D)}{q(z)}}_{\mathcal{L}(q)} + \mathbb{E}_q \log(p(D))$$

we have the access (GM)

const quantity

$$\underbrace{KL}_{\geq 0} = -\overbrace{\mathcal{L}(q)}^{\leq 0,} + \underbrace{\log \underbrace{p(D)}_{\text{marginal}}}_{\text{evidence}}$$

→ something ≤ 0 $\log(0:1) \leq 0$

→ fixed quantity

$$\underbrace{\mathcal{L}(q)} \leq \overbrace{\log p(D)}^{\text{evidence}}$$

lower bond of the evidence

→ ELBO

<u>ELBO</u> $\mathcal{L}(q) = -\mathbb{E}_q \log\frac{p(z,D)}{q(z)}$

$\mathcal{L}(q) = \log(p(D))$ i.f.f. $KL\left(q(z)\,\|\,p(z|D)\right) = 0$

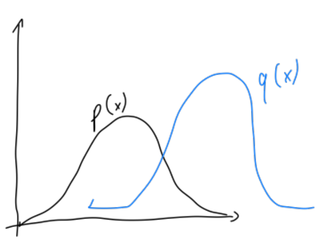
i.e. $q(z)$ is the same as $p(z|D)$

Maximizing ELBO ⇒ Minimizing KL

We can form an optimization problem → $$\underbrace{\log(p(x))}_{} = D_{KL}\left(q(z)\,\|\,p(z|D)\right) + \mathcal{L}(q)$$

this part is fixed! Even if we do not know the θ of p

This should be before but → intuition for KL divergance, distance between distributions



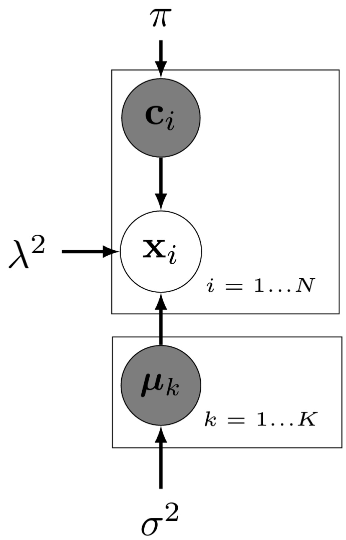
What is the distance, how far are they from each other

$$D_{KL}(p\|q) = \mathbb{E}_{x\sim p(x)}\left[\log\left(\frac{p(x)}{q(x)}\right)\right] = \begin{cases} \sum_x p(x)\log\left(\frac{p(x)}{q(x)}\right) \to \text{discrete} \\ \int p(x)\log\left(\frac{p(x)}{q(x)}\right) \to \text{cont.} \end{cases}$$

How <u>far</u> description of one RV is away from another?

Actual tutorial starts below!

ELBO (Assignment 1)



$$\mathcal{L}(x \mid m, s^2, \phi) = \underbrace{\sum_{k=1}^{K} E_q(\log p(\mu_k))}_{I} + \underbrace{\sum_{i=1}^{N} E_q[\log p(c_i)]}_{II} + \underbrace{\sum_{i=1}^{N} E_q[\log p(x_i \mid c_i, \mu)]}_{III} - \underbrace{\sum_{i=1}^{K} E_q[\log q(\mu_k)]}_{IV} - \underbrace{\sum_{j=1}^{N} E_q[\log q(c_i)]}_{V}$$

Compute closed form of ELBO

$\begin{cases} K \text{ – components} \\ N \text{ – sample size} \end{cases}$

I. $\mu_k \sim \mathcal{N}(\alpha, \sigma^2, I)$

$$\sum^{K} E_q \log(p(\mu_k)) = \sum E_q \log(\mathcal{N}(\alpha, \sigma^2, I)) = \sum E_q \log\left(\frac{1}{\sqrt{2\pi^p \sigma^2 I}} e^{-\frac{1}{2}(\mu_k - \alpha)^T (\sigma^2 I)^{-1} (\mu_k - \alpha)}\right)$$

$$= \sum E_q\left[\log\left(\frac{1}{\sqrt{2\sigma^2 \pi^p}}\right) - \frac{1}{2}(\mu_k - \alpha)^T \sigma^{-2} (\mu_k - \alpha)\right] = \sum^{K} \log \frac{1}{\sqrt{2\sigma^2 \pi^p}} - \frac{1}{2\sigma^2} E[(\mu_k - \alpha)^T (\mu_k - \alpha)]$$

$\Sigma p(\cdot) = 1 -$

Since we calculate the expectation E_q then [according to "Var. Bay. Model Selection for mix' distr."]

$$\begin{cases} E[\mu] = m \\ E[\mu^2] = S^2 + m^2 \quad * \\ E[\text{cnst.}] = \text{cnst.} \end{cases}$$

thus,

where "p" is the dim. of the data for multivar. gauss.

$$= K \log \frac{1}{\sqrt{2\sigma^2 \pi^p}} - \frac{1}{2\sigma^2} \sum^{K} \left(S_k^2 + m_k^T m_k - m_k^T \alpha - \alpha^T m_k + \alpha^T \alpha\right)$$

II

$$c_i \sim \text{Categorical}\left(\frac{1}{K}, \dots, \frac{1}{K}\right) = \log \prod_{k=1}^{K} \left(\frac{1}{K}\right)^{c_{i,k}} = \sum^{K} \log \left(\frac{1}{K}\right)^{c_{i,k}} = \sum^{K} -c_{i,k} \log K$$

$\log \frac{1}{a} = -\log a$

$$\sum^{N} E_q\left[\log \sum^{K} -c_{i,k} \log(K)\right] = \sum^{N} \underbrace{\sum^{K} -\phi_{i,k}}_{-1} \log(K) = \sum -\log(K) = \underbrace{-\log(K) - \log(K) \dots - \log K}_{N\text{-times}}$$

$E_q[c_{ik}] = \phi_{ik}$

$\Sigma p(\cdot) = 1 \to$

$$= -N \cdot \log K$$

=

III

$$x_i \mid c_i, \mu \sim \mathcal{N}(c_i^T \mu, \lambda^2, I) \overset{*}{=} \prod^{K} p(x \mid \mu_k)^{c_{k,i}}$$

$*$ using the fact that $p(x_i \mid c_i, \mu) = \prod^{K} p(x_i \mid \mu_k)^{c_{i,k}}$

$$\sum_{i=1}^{N} E_q[\log p(x_i \mid c_i, \mu)] = \sum_i E_q[\log p(x_i \mid c_i, \mu)] = \sum_i E_q\left[\log \prod^{K} p(x \mid \mu_k)^{c_{ki}}\right] = \sum_i^{N} E_q \sum^{K} c_{ik} \log p(x \mid \mu_k) = \sum_i^{N} E_q \sum^{K} \phi_{ik} \log p(x \mid \mu_k) = \sum \sum \phi_{ik} E_q \log(p(x \mid \mu_k))$$

$$\sum^{N} \sum^{K} \phi_{ik} E_q\left[\log\left(\frac{1}{\sqrt{2\pi^p \lambda^2}} e^{-\frac{1}{2}(x_i - \mu)^T (\lambda^2)^{-1} (x_i - \mu)}\right)\right] =$$

$$= \sum^{N} \sum^{K} \phi_{i,k} \log \frac{1}{\sqrt{2\pi^p \lambda^2}} - \frac{1}{2\lambda^2} E_q\left[(x_i - \mu)^T (x_i - \mu)\right] \Rightarrow \text{Similarly here as in I}$$

$$\sum^{N}\sum^{k}\Phi_{i,k}\left(\log\left(\frac{1}{\sqrt{2\pi^{p}\lambda^{2}}}\right)\right)-\frac{1}{2\lambda^{2}}\left(x_i^Tx_i-x_i^Tm_k-m_k^Tx_i+M_k^TM+S_k^2\right)\Big)$$

$\underline{IV}$

$$q(\mu_k)\sim\mathcal{N}(\mu_k|m_k,s_k^2I)=\frac{1}{\sqrt{2\pi^{p}s_k^2}}\cdot e^{-\frac{1}{2}(\mu_k-m_k)^Ts_k^{-2}(\mu_k-m_k)}$$

$$\sum^{k}E\left[\log q(\mu_k)\right]$$

$$=\sum^{k}E_q\left[\log\left(\frac{1}{\sqrt{2\pi s_k^2}}\right)-\frac{1}{2s_k^2}(\mu_k-m_k)^T(\mu_k-m_k)\right]=\sum^{k}\log\frac{1}{\sqrt{2\pi^{p}s_k^2}}-\frac{1}{2s_k^2}s_k^2=\sum^{k}\left(\log\frac{1}{\sqrt{2\pi^{p}s_k^2}}-1\right)$$

Expectation $\rightarrow$

$$m_k^Tm_k-m_k^Tm_k-m_k^Tm_k+m_k^Tm_k+s_k^2=s_k^2$$

$\underline{V}$ $\sum^{N}E_q\left[\log(q(c_i))\right]$

$$q(c_i)\sim\text{cat}(\phi_i)=\prod\phi_{i,k}^{c_{i,k}}$$

$$\sum^{N}E\log\prod\phi_{i,k}^{c_{i,k}}=\sum^{N}\sum^{k}E\left[c_{i,k}\log\phi_{i,k}\right]=\sum^{N}\sum^{K}\phi_{i,k}\cdot\log\phi_{i,k}$$

Assignment 2

show that $\phi_{i,k}\propto\exp\left\{\frac{x_i^TE[\mu_k]}{\lambda^2}-\frac{E[\mu_k^T\mu_k]}{2\lambda^2}\right\}$

In order to show variational update, we have to calculate derivative of ELBO w.r.t. $\phi_{i,k}$. As suggested in 1st assig. and using the fact that $\frac{\partial(f+g)}{dx}=\frac{\partial f}{\partial x}+\frac{\partial g}{dx}$ I will split the assigment into 5 terms

$$\text{I.}\quad\underbrace{\frac{\partial}{\partial\phi}K\log\frac{1}{\sqrt{2\sigma^2\pi^{p}}}}_{\frac{\partial}{\partial\phi}\text{const}=0}-\underbrace{\frac{1}{2\sigma^2}\sum^{k}\left(s_k^2+m_k^Tm_k-m_k^T\alpha-\alpha^Tm_k-\alpha^T\alpha\right)}_{\frac{\partial(\text{const})}{\partial\phi}=0}=0\quad\rightarrow\text{derivative w.r.t. }\phi_{i,k}=0$$

$$\text{II}\quad\frac{\partial}{\partial\phi}-N\cdot\log k=0$$

$$\text{III}\quad\frac{\partial}{\partial\phi_{i,k}}\sum^{N}\sum^{k}\underbrace{\Phi_{i,k}}_{=1}\left(\log\left(\frac{1}{\sqrt{2\pi^{p}\lambda^{2}}}\right)-\frac{1}{2\lambda^{2}}x_i^Tx_i+x_i^Tm_k-m_k^Tx_i+M_k^TM+S_k^2\right)=$$

$$= \sum^{N}\sum^{k} \log\left(\frac{1}{\sqrt{2\pi^{p}\lambda^{2}}}\right) - \frac{1}{2\lambda^{2}}\left(x_i^T x_i + x_i^T m_k - m_k^T x_i + m_k^T m + s_k^2\right)$$

$$\text{IV} \quad \frac{\partial}{\partial\phi_{i,k}} \sum^{k}\left(\log\frac{1}{\sqrt{2\pi^{p} s_k^2}} - 1\right) = 0$$

$$\text{V} \quad \frac{\partial}{\partial\phi_{ik}} \sum^{N}\sum^{K} \phi_{i,k}\cdot\log\phi_{i,k} = \sum^{N}\sum^{k} \log\phi_{ik} + \phi_{ik}\cdot\frac{1}{\phi_{ik}} = NK + \sum^{N}\sum^{k}\log\phi_{ik}$$

Finally

$$\frac{\partial}{\partial\phi}\mathcal{L} = \sum^{N}\sum^{k} \log\left(\frac{1}{\sqrt{2\pi^{p}\lambda^{2}}}\right) - \frac{1}{2\lambda^{2}}\left(x_i^T x_i + x_i^T m_k - m_k^T x_i + m_k^T m + s_k^2\right) + \log\phi_{ik} + 1$$

if we compare the term within summation to 0, then

$$\log\left(\frac{1}{\sqrt{2\pi^{p}\lambda^{2}}}\right) - \frac{1}{2\lambda^{2}}\left(x_i^T x_i + x_i^T m_k - m_k^T x_i + m_k^T m + s_k^2\right) + \log\phi_{ik} + 1 = 0 \qquad / \exp()$$

$$\phi_{i,k} = \underbrace{\frac{1}{\sqrt{2\pi^{p}\lambda}}}_{=\text{const.}} \underbrace{\cdot \exp\left(\frac{1}{2\lambda^{2}}\left(x_i^T x_i + x_i^T m_k - m_k^T x_i + m_k^T m + s_k^2\right)\right.}_{*} + \underbrace{e}_{\text{const.}}$$

since we do not care about exact value of $\phi_{i,k}$ but the proportionality, let's focus on term $*$. If we look back to the assignment 1, we can see that this term was

$$\underline{E_q (x-\mu)^T (x-\mu)} = -x^T E[\mu_i] + x_i^T x_i - E[\mu_i^T] x_i + E[\mu_i^T \mu_i]$$

if we assume that

$$x_i^T E[\mu_i] = E[\mu_i^T] x_i \text{ , then}$$

$$\exp\left(\frac{E[\mu_i^T\mu_i]}{2\lambda^2} - \frac{\cancel{2} x_i^T E[\mu_i]}{\cancel{2}\lambda^2} + \underbrace{\frac{x_i^T x_i}{2\lambda^2}}_{\text{const.}}\right) \text{, thus} \quad \phi_{i,k} \propto \exp\left(\frac{E[\mu_i^T\mu_i]}{2\lambda^2} - \frac{x_i^T E[\mu_i]}{\lambda^2}\right)$$

$\blacksquare$

Assignment 3

$$= -\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2\sigma^2} + \sum_{i=1}^{N}\phi_{i,k}\Big(-\frac{1}{2\lambda^2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\Big) + const$$

Assignment 3
Complete the square to find the parameters of the optimal Gaussian $\boldsymbol{\mu}$
$\mathcal{N}(\boldsymbol{m}_k, s_k^2\boldsymbol{I})$. Those parameters will be used for variational updates of the
terior of the mixture component means.

find velue of m_k & s_k^2

$$\mu_k \sim N(m_k, s_k^2 I) = \frac{1}{\sqrt{2\pi^p s_k^2}} \cdot e^{\underbrace{-\frac{1}{2}(\mu_k - m_k)^T s_k^{-2}(\mu_k - m_k)}}$$

, since we are comparing exp to exp we should focus on that term

$$-\frac{1}{2s_k^2}(\mu_k - m_k)^T(\mu_k - m_k) = -\frac{1}{2s_k^2}\Big(\mu_k^T\mu_k \underbrace{- \mu_k^T m_k - m_k^T\mu_k}_{\text{as in ass.2 this term are equal}} + \underbrace{m_k^T m_k}_{\text{const. } \to \text{skip}}\Big) = \underbrace{-\mu_k^T\mu_k\Big(\frac{1}{2s_k^2}\Big)}_{I} + \underbrace{\mu_k^T\Big(\frac{m_k}{s_k^2}\Big)}_{II}$$

$$q^*(\mu_k) \propto -\frac{\mu_k^T\mu_k}{2\sigma^2} + \sum_{i=1}^{N}\phi_{i,k}\Big(-\frac{1}{2\lambda^2}(x_i-\mu_k)^T(x_i-\mu_k)\Big) + \underbrace{const}_{\text{does not matter}} = -\frac{\mu_k^T\mu_k}{2\sigma^2} + \sum_i\Big(-\frac{\phi_{i,k}}{2\lambda^2}\Big)\Big(\underbrace{x_i^Tx_i}_{\text{cst.} \downarrow \text{ we can skip}} - x_i^T\mu_k - \mu_k^T x_i + \mu_k^T\mu_k\Big) =$$

$$= -\frac{\mu_k^T\mu_k}{2\sigma^2} + \sum_i^N \frac{-\phi_{ik}\mu_k^T\mu_k + \phi_{i,k}\cdot x_i\mu_k^T + \phi_{ik}\mu_k^T x_i}{2\lambda^2} = \underbrace{\mu_k^T\mu_k\Big(\frac{1}{2\sigma^2} + \frac{\sum_i^N\phi_{i,k}}{2\lambda^2}\Big)}_{I^*} + \underbrace{\mu_k^T\Big(\frac{\sum_i^N\phi_{ik}\cdot x_i}{\lambda^2}\Big)}_{II^*}$$

Now, by comparing I^* and II^* with I and II

$$\begin{cases} \dfrac{1}{2s_k^2} = \dfrac{1}{\sigma^2} + \sum_i^N \dfrac{\phi_{i,k}}{2\lambda^2} \Rightarrow s_k = 2\Big(\dfrac{1}{\sigma^2} + \sum^N \dfrac{\phi_{i,k}}{2\lambda^2}\Big)^{-1/2} \qquad s_k = \dfrac{1}{2\big(\frac{1}{\sigma^2} + \sum^N \frac{\phi_{ik}}{2\lambda^2}\big)} \\ \dfrac{m_k}{s_k^2} = \dfrac{\sum\phi_{ik}\cdot x_i}{\lambda^2} \Rightarrow m_k = \underbrace{s_k^2}_{\text{calc. above}}\dfrac{\sum\phi_{ik}\cdot x_i}{\lambda^2} \end{cases}$$