Homework-2

- Theory part.

Q.1)

Acc. to bayes theorem we know, the probability of the hidden variables over the observations are

$$P(z/x) = \frac{P(x/z)\, P(z)}{P(x)}$$

$$= \frac{P(x/z)\, P(z)}{\int_z P(x,z)}$$

we can prove the variational lower bound 'L' by applying log of the observational probability on the denominator of the $P(z/x)$

$$\Rightarrow \log \int_z P(x,z)$$

As we approximate the $P(z/x)$ using $q(z)$ we add it to the equation of the ELBO.

$$\Rightarrow \log \int_z \frac{P(x,z)\, q(x)}{q(x)}$$

$$\Rightarrow \log \left(E_q \left[\frac{P(x,z)}{q(z)} \right] \right)$$

the above equation can also be written as.
$$E_q \left[\log \frac{P(x,z)}{q(z)} \right] \leq \log P(x)$$

$$\Rightarrow E_q \log P(x,z) - E_q \log q(z)$$

which is the variational lower bound.

→ Now we try to prove,

$$KL(q(z) \| P(z|x)) = -E_q [\log P(x,z)]$$
$$+ E_q \log q(z)$$
$$+ \log P(x)$$

We know K·L divergence is used to define the similarity between two different distributions in our case it's P & q

when we talk about the joint probability $P(z,x)$ ~~to~~ & factorize it.

$$P(z/x) = P(z,x)\ P(z) \quad \text{— } (A)$$

• Next if we apply log and expectations

$$E_q \log (P(x/z)\, P(z))$$

$$= E_q[\log P(x/z)] + E_q[\log P(z)]$$

- then we apply integral & write it as

$$\Rightarrow \int q(z) (\log P(z) - \log q(z))\, dz$$

$$\Rightarrow \int q(z) \log \frac{P(z)}{q(z)}\, dz.$$

by the help of the above equation we can write the KL divergence as

$$KL(q(z) \| P(z/x)) \Rightarrow \int q(z) \log \frac{q(z)}{P(z/x)}\, dz$$

As here our distributions are $q(z)$ & $P(z/x)$

$$= \int q(z)\left(\log q(z) - \log P(z|x)\right) dz$$

$$= \int q(z) \log q(z) - q(z) \log P(z|x)\, dz$$

$$= \int q(z) \log q(z)\, dz - \int q(z) \log P(z|x)\, dz$$

$$= E_q\left(\log q(z)\right) \cancel{dz} - E_q\left(\log P(z|x)\right)$$

$$= E_q\left(\log q(z)\right) - E_q\left(\log \frac{P(x,z)}{P(x)}\right)$$

$\Rightarrow$ we write $P(z|x) = P(z,x) \cdot P(z)$

~~so changed it as P(~~

$$= E_q\left(\log q(z)\right) - E_q\left[\log P(x,z) - \log P(x)\right]$$

$$= E_q\left(\log q(z) - \log P(x,z)\right) + E_q\left[\log P(x)\right]$$

$$= E_q \log q(z) - \cancel{\log} E_q \log P(x,z) + E_q \log P(x)$$

$$\Rightarrow -E_q \log P(x,z) + E_q \log q(z) + E_q \log P(x)$$

Q.2)

The given $\delta(x)$ is the binary classifier with the classification:

$$\delta(x) = \begin{cases} +1 & \text{if } P(y=+1|X=x) \geq P(Y=-1|X=x) \\ -1 & \text{if otherwise} \end{cases}$$

$t(x)$ is another classifier with an error rate of $R(t)$ which can be defined as

$$R(t) = P(Y \neq t(x))$$

which says when the classification $t(x)$ gives is not the same as the expected label or output. we can write that as

$$R(t) = E_X \left[P(y \neq t(x) / X = x) \right]$$

Now we need to show that the error rate of the binary classifier $R(\delta)$ is greater than or equal to $R(f)$ which is any binary classifier.

All the features $X \in \mathcal{X}$ & labels $Y \in \mathcal{Y}$ &
x, y are distributed ~~according to~~

for an binary classifier
$f : X \to Y$ its 0-1 loss $l(y f(x))$ is

$$l(y, f(x)) \Rightarrow P(y \neq f(x))$$ or

we can say in other words,

$$l(y, f(x)) = \begin{cases} 1 & \text{if } y \neq f(x) \\ 0 & \text{otherwise} \end{cases}$$

the error rate $R(f)$ is defined as

$$R(f) = E[(Y, f(x))] = P(Y \neq f(x))$$

if $R(\delta) \geq R(f)$ then,

$$\eta(x) \geq 1/2 \Leftrightarrow \frac{P(Y = +1 \mid X = x)}{P(Y = -1 \mid X = x)} \geq 1$$

for any δ

$$P(Y = \delta(x) \mid X = x) = 1 - \{P(Y = 1, \delta(x) = 1 \mid X = 1)$$

$$+ P(Y = -1, g(x) = -1 \mid X = x)$$

$$= 1 - \{E[1_\delta(x = 1)]\eta(x) + E[\delta(x = -1)\eta(x)]\}$$

so,

$$P(Y \neq g^{\delta}(x) \mid X = x) - P(Y \neq f(x) \mid X = x)$$

$$= \eta(x)\{E[f(x) = 1)] - E(g(x) = 1]\} +$$

$$(1-\eta(x))\{[E(f(x) = -1)] + E(g(x) = -1)]\}$$

this after solving

$$\Rightarrow (2\eta(x) - 1)\{[E(f(x) - 1)]$$

$$- E[\delta(x) = 1]\} \leq 0$$

taking the expectation with respect to X gives

$$~~R(\delta) - R(f) \leq 0~~$$

$$~~\Rightarrow R(\delta) \geq R(f)~~$$

$$R(\delta) - R(f) \leq 0$$

$$R(\delta) \leq R(f)$$

Q.3) Given a set of data with n-Paired samples in the form of $\{(x_i, y_i)\}_{i=1}^{n}$ where x_i is the d-dimentional vector of i^{th} sample and y_i is the label of the same i^{th} sample.

- The log likelihood function of the logistic regression.

$$l(\beta) = \sum_{i=1}^{n} y_i \beta^T x_i - \log(1 + e^{\beta^T x_i})$$

We need to compute $\frac{\partial l(\beta)}{\partial(\beta)}$ &

$$\frac{\partial l(\beta)}{\partial\beta\, \partial\beta^T}.$$

First computing $\frac{\partial l(\beta)}{\partial(\beta)}$

$$\Rightarrow \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} \frac{x_{ij} e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}$$

$$= \sum_{i=1}^{N} \left(y_i - \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \right) x_i$$

$$= \sum_{i=1}^{n} \left(y_i x_{ij} - \frac{x_{ij} e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \right)$$

$$= \sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} \frac{x_{ij} e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}$$

$$= \sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} p(x_i;\beta) x_{ij}$$

$$= \sum_{i=1}^{N} x_{ij} \left(y_i - p(x_i;\beta) \right) = 0$$

for $\frac{\partial l(\beta)}{\partial \beta} = 0, \; i = 0, 1, \cdots p.$

For the 2^{nd} order derivative

$$\frac{\partial l(\beta)}{\partial \beta \, \partial \beta^T}$$

$$= - \sum_{i=1}^{N} \frac{\left(1 + e^{\beta^T x_i}\right) e^{\beta^T x_i} x_i x_i^T - \left(e^{\beta^T x_i}\right)^2 x_i x_i^T}{\left(1 + e^{\beta^T x_i}\right)^2}$$

$$= -\sum_{i=1}^{N} x_{i\beta}\, x_{i\beta}^{T}\, P(x_i;\beta) - x_{i\beta}\, x_i^{T} P(x_i;\beta)^2$$

$$= -\sum_{i=1}^{N} x_{i\beta}\, x_i^{T}\, P(x_i;\beta)(1 - P(x_i;\beta))$$

$$\Rightarrow \frac{\partial l(\beta)}{\partial\beta\,\partial\beta^{T}} = -\sum_{i=1}^{N} x_i x_i^{T} P(x_i;\beta)(1-P(x_i;\beta))$$

we solve thes equations for the value of β, as with the 1^{st} derivative the β value is not easy or possible ~~for computing~~ using gradient decent. so we compute β'

$$\beta' = \beta - \frac{f(\beta)}{f'(\beta)}$$

$\Rightarrow f(\beta)$ is the 1^{st} derivative $\frac{\partial l(\beta)}{\partial(\beta)}$

and the $f'(\beta)$ is the 2^{nd} derivative

$$\frac{\partial l(\beta)}{\partial(\beta)\partial(\beta^{T})}$$