Introduction to Deep Learning M2177.0043
Seoul National University

Homework #(**HW7**)
Seo Junwon

INSTRUCTIONS

- Anything that is received after the deadline will be considered to be late and we do not receive late homeworks. We do however ignore your lowest homework grade.
- Answers to every theory questions need to be submitted electronically on ETL. Only PDF generated from LaTex is accepted.
- Make sure you prepare the answers to each question separately. This helps us dispatch the problems to different graders.
- Collaboration on solving the homework is allowed. Discussions are encouraged but you should think about the problems on your own.
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution.

1 Getting familiar with KL Divergence

Theorem 1. *if p follows N(* μ, Σ*), Pr(p = x) =* $\frac{e^{-0.5(x-\mu)^T\Sigma^{-1}(x-\mu)}}{2\pi^{n/2}det(\Sigma)^{0.5}}$

$D(p\|q)$

= $E_p[logp - logq]$

= 0.5 * $E_p[-log\det\Sigma_1 - (x-\mu_1)^T\Sigma_{-1}(x-\mu_1) + log\det\Sigma_2 + (x-\mu_1)^T\Sigma_{-1}(x-\mu_1)]$

= 0.5 * ($log\frac{\det\Sigma_2}{\det\Sigma_2} + E_p[-tr(\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^T) + tr(\Sigma_2^{-1}(x-\mu_2)(x-\mu_2)^T)]$)

= 0.5 * ($log\frac{\det\Sigma_2}{\det\Sigma_2} - n + E_p[tr(\Sigma_2^{-1}(x-\mu_2)(x-\mu_2)^T)]$)

= 0.5 * ($log\frac{\det\Sigma_2}{\det\Sigma_2} - n + tr(\Sigma_2^{-1}(\Sigma_1 + (\mu_1\mu_1^T) - 2\mu_2\mu_1^T + \mu_2(\mu_2)^T))$)

= 0.5 * ($log\frac{\det\Sigma_2}{\det\Sigma_2} - n + tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^T{\Sigma_2}^{-1}(\mu_2-\mu_1)$)

Introduction to Deep Learning M2177.0043
Seoul National University

Homework #(**HW7**)
Seo Junwon

---

2 Forward and Reverse KL Divergence

2.1 Forward KL

$\theta^* = argmin_\theta D_{KL}(p(x) \| q_\theta(x)$

= $argmin_\theta E_p[log(p(x)) - log(q(x))]$

= $argmin_\theta E_p[-log(q(x))] - H(p(x))$

= $argmin_\theta E_p[-log(q(x))]$, since p is fixed distribution.

2.2 Reverse KL

$\theta^* = argmin_\theta D_{KL}(q_\theta(x) \| p(x))$

= $argmin_\theta E_q[-log(p(x)) + log(q(x))]$

= $argmin_\theta E_q[-log(q(x))] - H(q(x))$

2.3 Meaning

$E_p[-log(q(x))]$ and $E_q[-log(p(x))]$ is log - likelyhood between p and q.

Whenever P has hight probability, Q must also have high probability. We consider this mean-seeking behavior in Forward KL, and Mode-seeking behavior in reverse KL.

$H(q(x))$ is Entropy of q.

When p is a unimodal Gaussian, q_θ will follow same Gaussian distribution with same parameter.

When P is a bimodal Gaussian, The forward KL q will approximate distribution centers itself between the two modes, so that it can have high coverage of both. The forward KL divergence does not penalize Q for having high probability mass where P does not.

However, in reverse KL, q will approximate distribution within a mode of P since it's required that sample from q have high probability under P. The entropy term prevents the approximate distribution collapsing to very narrow mode.

3 Mutual Information and independence

3.1 1

$I(X;Y) = D_{KL}(p(X,Y)\|p(X)p(Y))$

$= E_{X,Y}(log\frac{p(X,Y)}{p(X)p(Y)})$

As X and Y are independent,

$p(X,Y) = p(X) * p(Y)$

Therefore, $E_{X,Y}(log\frac{p(X,Y)}{p(X)p(Y))} = E_{X,Y}(log\frac{p(X)p(Y)}{p(X)p(Y)})$

= $E_{X,Y}(log(1)) = 0$

3.2 2

Using Jensen's Inequality, we can show KL Divergence $D_{KL}(p\|q) \geq 0$,

and it meets equality (= 0) if and only if $p(x) = q(x), \forall x$

which means p and q have exactly same density function.

Jensen's Inequality has equality condition if and only if inner function (log in I(X;Y)) is affine function or variable($P(X,Y)/P(X)P(Y)$ in $I(X;Y)$) is constant. Therefore, for KL Divergence to be 0, $P(X,Y)/P(X)P(Y)$ should be constant, exactly "1" as $pdf <= 1$,

Thus $I(X;Y) = D_{KL}(p(X,Y)\|p(X)p(Y)) = 0->$

$p(X,Y) = p(X)p(Y)$.

Homework #(**HW7**)
Seo Junwon

Therefore X and Y are Independent.

4 Entropy of a multivariate normal distribution

$H(x) = -\int p(x) * log(\frac{e^{-0.5(x-\mu)^T\Sigma^{-1}(x-\mu)}}{2\pi^{n/2}det(\Sigma)^{0.5}})dx$

= - $\int p(x) * log(e^{-0.5(x-\mu)^T\Sigma^{-1}(x-\mu)})dx$ + $\int p(x) * log(2\pi^{n/2}det(\Sigma)^{0.5})dx$

Using trace trick,

= 0.5 * $\int p(x) * (x-\mu)^T\Sigma^{-1}(x-\mu)dx$ + $log(2\pi^{n/2}det(\Sigma)^{0.5})dx$

= 0.5 * $\int p(x) * tr((x-\mu)^T\Sigma^{-1}(x-\mu))dx$ + $\frac{N}{2}log(2\pi) + \frac{1}{2}log(det(\Sigma))$

= 0.5 * $tr(\int \Sigma^{-1}p(x)(x-\mu)(x-\mu)^T)dx$+ $\frac{N}{2}log(2\pi) + \frac{1}{2}log(det(\Sigma))$

= 0.5 * $tr(\Sigma^{-1}\Sigma)$+ $\frac{N}{2}log(2\pi) + \frac{1}{2}log(det(\Sigma))$

= $\frac{N}{2} + \frac{N}{2}log(2\pi) + \frac{1}{2}log(det(\Sigma))$