KL Divergence

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 16

Kullback-Leibler (KL) divergence

- Consider some unknown distribution $p(\mathbf{x})$. Suppose we model this using an approximating distribution $q(\mathbf{x})$.
- The additional information required to specify the value of $\mathbf{x}$ as a result of using q instead of p is called the **relative entropy**, or Kullback-Leibler **(KL) divergence**, given by

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x})\, \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x}.\end{aligned}$$

- Fact: $\mathrm{KL}(p\|q) \geqslant 0$

Kullback-Leibler (KL) divergence

- Suppose that data ~~is~~ are being generated from an unknown distribution $p(\mathbf{x})$ that we wish to model. We can try to approximate this distribution using some parametric distribution $q(\mathbf{x}|\theta)$.
- One way to determine θ is to minimize the KL divergence from $p(\mathbf{x})$ to $q(\mathbf{x}|\theta)$ with respect to θ.
- We cannot do this directly because we don't know $p(\mathbf{x})$. Suppose, however, that we have observed a finite set of training points $\mathbf{x}_n$,for $n = 1, \cdots, N$, drawn from $p(\mathbf{x})$. Then

$$\mathrm{KL}(p\|q) \simeq \frac{1}{N}\sum_{n=1}^{N}\{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\} \approx \mathbb{E}_p \ln\left(\frac{p(x)}{q(x)}\right)$$

- What are we doing if we minimize this KL divergence?

mutual information

- If $\mathbf{x}$ and $\mathbf{y}$ are independent, then $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$.
- For a general $p(\mathbf{x}, \mathbf{y})$, how close is it to being independent? We can use KL to measure

$$\begin{aligned} \mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \end{aligned}$$

- This is called the **mutual information** between $\mathbf{x}$ and $\mathbf{y}$

mutual information

by symmetry

- Fact: $\mathrm{I}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] \overset{\downarrow}{=} \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$

$$\text{RHS} = -\int p(x)\ln p(x)\,dx + \int p(x,y)\ln p(x|y)\,dx\,dy$$

$$= -\int p(x,y)\ln p(x)\,dx\,dy + \int p(x,y)\ln p(x|y)\,dx\,dy$$

$$= -\int p(x,y)\ln\left(\frac{p(x)}{p(x|y)}\right)dx\,dy$$

$$= -\int p(x,y)\ln\left(\frac{p(x)\,p(y)}{p(x|y)\,p(y)}\right)dx\,dy$$

$$= -\int p(x,y)\ln\left(\frac{p(x)\,p(y)}{p(x,y)}\right)dx\,dy = \text{LHS}$$

information does not hurt

- Fact: $\mathrm{H}[\mathbf{y}|\mathbf{x}] \leq \mathrm{H}[\mathbf{y}]$

$I[x,y]$, defined as a KL divergence, is nonnegative.

Therefore, $H[y] - H[y|x] = I[x,y] \geq 0$

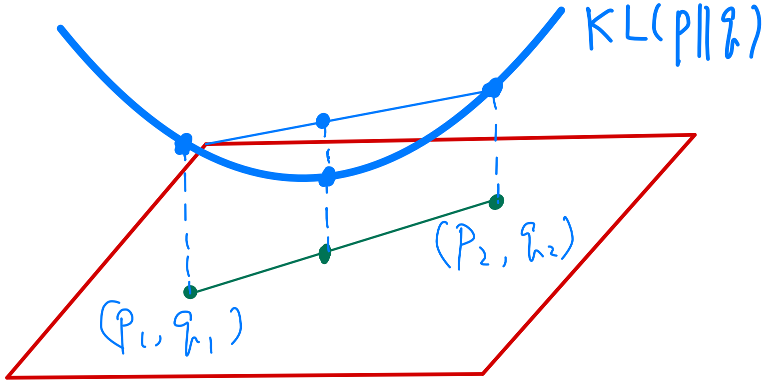
independence bound on entropy

- Fact: Let $\mathbf{x}_1, \cdots, \mathbf{x}_N$ be drawn according to $p(\mathbf{x}_1, \cdots, \mathbf{x}_N)$. Then

$$H[\mathbf{x}_1, \cdots, \mathbf{x}_N] \leq \sum_{n=1}^{N} H[\mathbf{x}_n]$$

$$\begin{aligned}
\text{LHS} &= H[x_N \mid x_1, \cdots, x_{N-1}] + H[x_1, \cdots, x_{N-1}] \\
&= H[x_N \mid x_1, \cdots, x_{N-1}] + H[x_{N-1} \mid x_1, \cdots, x_{N-2}] + H[x_1, \cdots, x_{N-2}] \\
&= \cdots \\
&= H[x_N \mid x_1, \cdots, x_{N-1}] + H[x_{N-1} \mid x_1, \cdots, x_{N-2}] + \cdots + \\
&\qquad H[x_3 \mid x_1, x_2] + H[x_2 \mid x_1] + H[x_1] \\
&\leq H[x_N] + H[x_{N-1}] + \cdots + H[x_1] = \text{RHS}
\end{aligned}$$

KL divergence is convex



- KL divergence $\mathrm{KL}(p \parallel q)$ is convex in (p, q) :

  For any $(p_1, q_1), (p_2, q_2), 0 \leq \lambda \leq 1,$

$$\mathrm{KL}(\lambda p_1 + (1-\lambda) p_2 \parallel \lambda q_1 + (1-\lambda) q_2) \leq \lambda\, \mathrm{KL}(p_1 \parallel q_1) + (1-\lambda) \mathrm{KL}(p_2 \parallel q_2)$$

KL divergence is convex

To prove the convexity of KL, we need the Log-Sum Inequality:

For nonnegative numbers $\{a_n\}_{n=1}^{N}$,$\{b_n\}_{n=1}^{N}$,

$$\sum_{n=1}^{N} a_n \ln\left(\frac{a_n}{b_n}\right) \geq \left(\sum_{n=1}^{N} a_n\right) \ln\left(\frac{\sum_{n=1}^{N} a_n}{\sum_{n=1}^{N} b_n}\right)$$

First note a fact: the function $f(u) = u \ln u$ is convex,

because $f'(u) = \ln u + 1$, $f''(u) = \frac{1}{u} > 0$ since $u > 0$

KL divergence is convex

$f(u) = u \ln u$ is convex

To prove the convexity of KL, we need the Log-Sum Inequality:

For nonnegative numbers $\{a_n\}_{n=1}^N, \{b_n\}_{n=1}^N$,

$$\sum_{n=1}^{N} a_n \ln\left(\frac{a_n}{b_n}\right) \geq \left(\sum_{n=1}^{N} a_n\right) \ln\left(\frac{\sum_{n=1}^{N} a_n}{\sum_{n=1}^{N} b_n}\right)$$

Let $p_n = \dfrac{b_n}{\sum_{m=1}^{N} b_m}$, $u_n = \dfrac{a_n}{b_n}$. Then by Jensen's Inequality.

$\sum_n p_n f(u_n) \geq f\left(\sum_n p_n u_n\right)$. Here, $p_n u_n = \dfrac{a_n}{\sum_m b_m}$.

That gives $\displaystyle\sum_n \frac{b_n}{\sum_{m=1}^{N} b_m} \frac{a_n}{b_n} \ln\left(\frac{a_n}{b_n}\right) \geq \sum_n \frac{a_n}{\sum_m b_m} \ln\left(\sum_n \frac{a_n}{\sum_m b_m}\right)$

KL divergence is convex

Now we have

$$\sum_n \frac{\cancel{b_n}}{\sum_{m=1}^{N} b_m} \frac{a_n}{\cancel{b_n}} \ln\left(\frac{a_n}{b_n}\right) \geq \sum_n \frac{a_n}{\sum_m b_m} \ln\left(\sum_n \frac{a_n}{\sum_m b_m}\right)$$

To prove the convexity of KL, we need the Log-Sum Inequality:

For nonnegative numbers $\{a_n\}_{n=1}^N, \{b_n\}_{n=1}^N$,

$$\sum_{n=1}^{N} a_n \ln\left(\frac{a_n}{b_n}\right) \geq \left(\sum_{n=1}^{N} a_n\right) \ln\left(\frac{\sum_{n=1}^{N} a_n}{\sum_{n=1}^{N} b_n}\right)$$

Therefore,

$$\frac{\sum_n a_n \ln\left(\frac{a_n}{b_n}\right)}{\sum_m b_m} \geq \frac{\sum_n a_n \ln\left(\frac{\sum_n a_n}{\sum_m b_m}\right)}{\sum_m b_m}$$

That is,

$$\sum_n a_n \ln\left(\frac{a_n}{b_n}\right) \geq \left(\sum_n a_n\right) \ln\left(\frac{\sum_n a_n}{\sum_n b_n}\right)$$

☺

KL divergence is convex

Proof: In order to show $\mathrm{KL}(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \leq \lambda \mathrm{KL}(p_1 \parallel q_1) + (1-\lambda)\mathrm{KL}(p_2 \parallel q_2)$,we only need to show that

$$(\lambda p_1(x) + (1-\lambda)p_2(x)) \ln \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)}$$

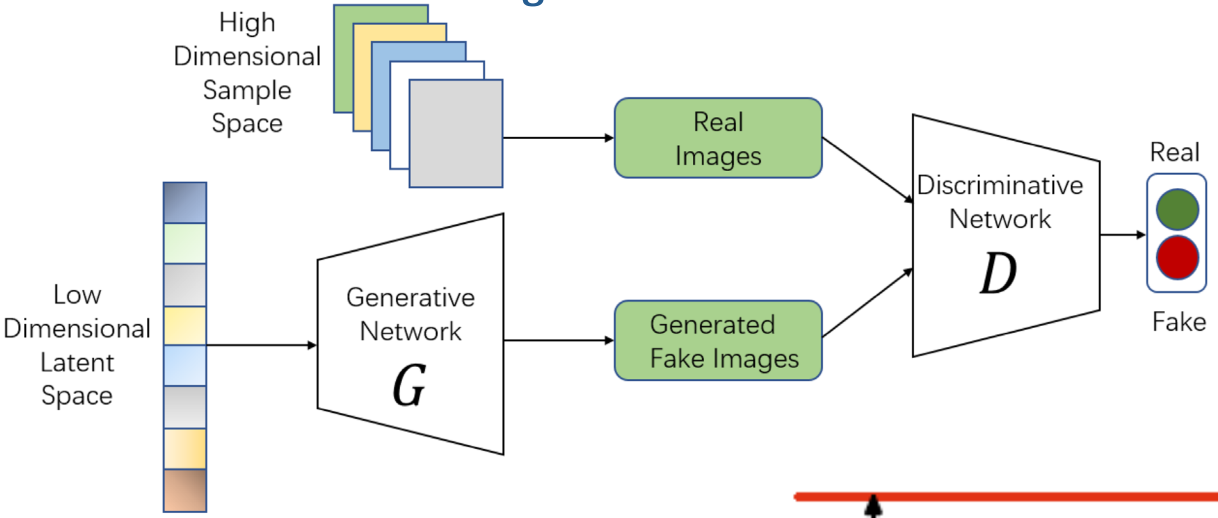
$$\leq \lambda p_1(x) \ln \frac{p_1(x)}{q_1(x)} + (1-\lambda)p_2(x) \ln \frac{p_2(x)}{q_2(x)}$$

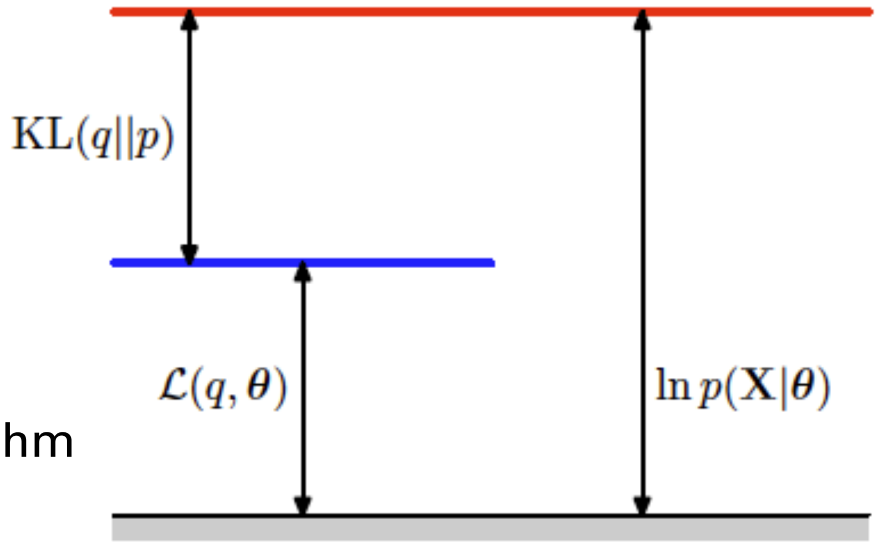
But this immediately follows the Log-Sum Inequality, with $N=2$,

$a_1 = \lambda p_1$, $a_2 = (1-\lambda) p_2$; $b_1 = \lambda q_1$, $b_2 = (1-\lambda) q_2$

applications of KL divergence

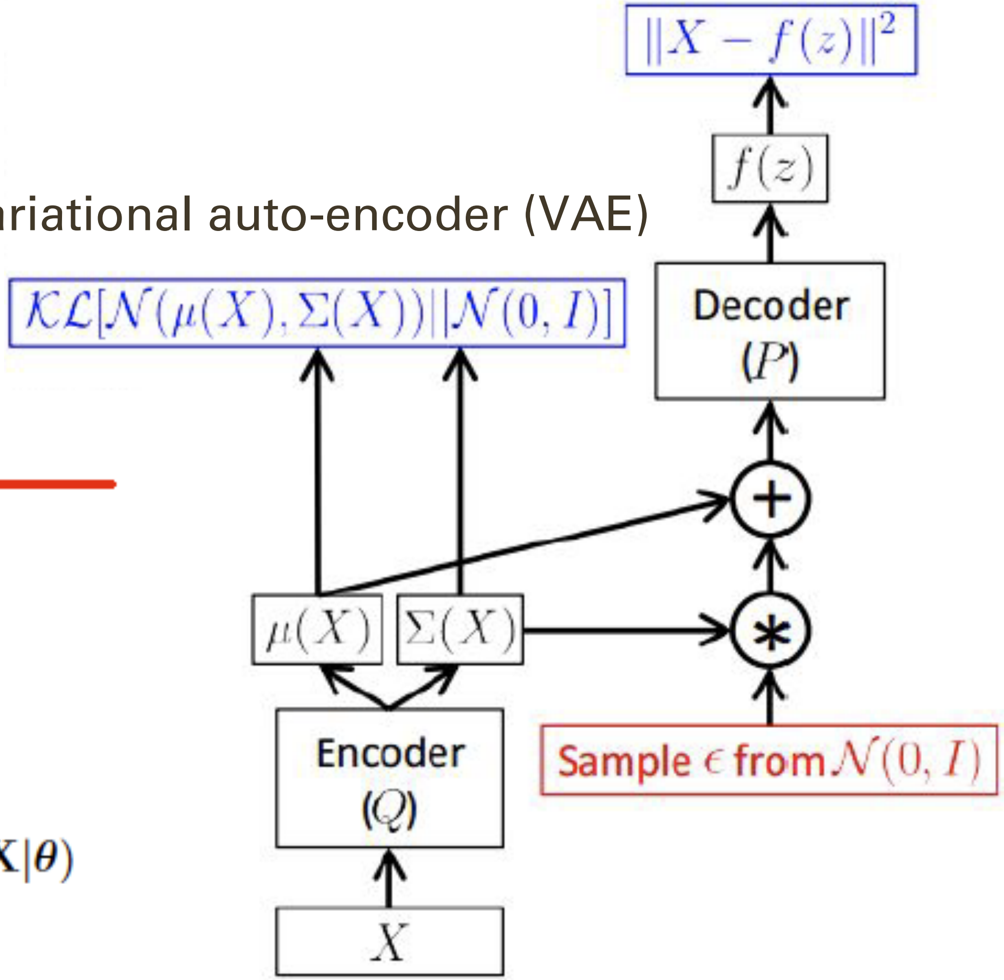
generative adversarial network (GAN)





EM algorithm

variational auto-encoder (VAE)



example: KL of Gaussian

- Let $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^D$, $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$, and $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I})$. Find $\mathrm{KL}(p \parallel q)$.

By definition, $KL(p\|q) = -\int p \ln q - \left(-\int p \ln p\right)$

Here,

$$-\int p(x) \ln p(x)\, dx$$

$$= -\int p(x) \ln N(x|\mu, \Sigma)\, dx$$

$$= -\int p(x) \ln\left(\frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right) dx$$

$$= -\int p(x)\left(-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) dx$$

$$= \underbrace{\frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma|}_{C} + \frac{1}{2}\int p(x)\,(x-\mu)^T\Sigma^{-1}(x-\mu)\, dx$$

$$= C + \frac{1}{2}\int \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \cdot (x-\mu)^T\Sigma^{-1}(x-\mu)\, dx$$

Change of variable $y = \Sigma^{-\frac{1}{2}}(x-\mu)$

$|J| = |\Sigma|^{-\frac{1}{2}}$

(if $\Sigma = \sum_j \lambda_j v_j v_j^T$, then $\Sigma^{\frac{1}{2}} = \sum_j \lambda_j^{\frac{1}{2}} v_j v_j^T$)

$$= C + \frac{1}{2}\int \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}\|y\|^2\right)\|y\|^2\, dy$$

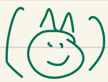
$$= C + \frac{1}{2} \int \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}\|y\|^2\right) \|y\|^2 \, dy$$

$$= C + \frac{1}{2} \int \mathcal{N}(y \mid 0, I) \, \|y\|^2 \, dy$$

$$= C + \frac{1}{2} \sum_{i=1}^{D} \int \mathcal{N}(y \mid 0, I) \, |y_i|^2 \, dy_1 \cdots dy_D$$

$$= C + \frac{1}{2} \sum_{i=1}^{D} 1$$

$$= C + \frac{D}{2} .$$



On the other hand,

$$- \int p(x) \ln q(x) \, dx$$

$$= - \int p(x) \ln \mathcal{N}(x \mid 0, I) \, dx$$

$$= - \int p(x) \ln \left(\frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}\|x\|^2\right) \right) dx$$

$$= \frac{D}{2} \ln(2\pi) + \frac{1}{2} \int p(x) \|x\|^2 \, dx$$

$$= \frac{D}{2} \ln(2\pi) + \frac{1}{2} \int \mathcal{N}(x \mid \mu, \Sigma) \|x\|^2 \, dx$$

$$= \frac{D}{2}\ln(2\pi) + \frac{1}{2}\int \|x+\mu\|^2 \, \mathcal{N}(x|0,\Sigma)\, dx$$

$$= \frac{D}{2}\ln(2\pi) + \frac{1}{2}\int \left(\|x\|^2 + \|\mu\|^2 + \underbrace{2x^T\mu}\right)\mathcal{N}(x|0,\Sigma)\, dx$$

odd function,
integral equal to zero

$$= \underbrace{\frac{D}{2}\ln(2\pi) + \frac{1}{2}\|\mu\|^2}_{\tilde{C}} + \frac{1}{2}\int \|x\|^2 \, \mathcal{N}(x|0,\Sigma)\, dx$$

$$= \tilde{C} + \frac{1}{2}\int \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right) \|x\|^2 \, dx$$

change of variable: $y = \Sigma^{-\frac{1}{2}}x$

$|J| = |\Sigma|^{-\frac{1}{2}}$

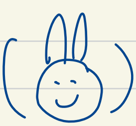
$$= \tilde{C} + \frac{1}{2}\int y^T\Sigma y \, \mathcal{N}(y|0,I)\, dy$$

Suppose $\lambda_1, \cdots, \lambda_D$ are the eigenvalues of Σ

$$= \tilde{C} + \frac{1}{2}\int \left(\sum_{i=1}^{D} \lambda_i y_i^2\right)\mathcal{N}(y|0,I)\, dy$$

$$= \tilde{C} + \frac{1}{2}\sum_{i=1}^{D}\lambda_i$$

$$= \tilde{C} + \frac{1}{2}\mathrm{tr}(\Sigma)$$



Therefore,

$KL(p\|q) =$ () $-$ ()

$$= \left(\tilde{C} + \frac{1}{2}\mathrm{tr}(\Sigma)\right) - \left(C + \frac{D}{2}\right)$$

$$= \left(\cancel{\frac{D}{2}\ln(2\pi)} + \frac{1}{2}\|\mu\|^2 + \frac{1}{2}\mathrm{tr}(\Sigma)\right) - \left(\cancel{\frac{D}{2}\ln(2\pi)} + \frac{1}{2}\ln|\Sigma| + \frac{D}{2}\right)$$

$$= \frac{1}{2}\left(\|\mu\|^2 + \mathrm{tr}(\Sigma) - \ln|\Sigma| - D\right).$$

f-divergence: generalization of KL

- In general, if f is a differentiable convex function satisfying $f(1) = 0$, then we can define a "divergence", called f-divergence, by

$$D_f(p \parallel q) = \int p(\mathbf{x}) f\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right) d\mathbf{x}$$

- For instance, take $f(u) = \frac{1}{2}(u-1)^2$, then

$$D_f(p \parallel q) = \frac{1}{2}\int \frac{(p(\mathbf{x}) - q(\mathbf{x}))^2}{p(\mathbf{x})} d\mathbf{x}.$$

Questions?

Reference

- *Information theory:*
  - *[Bi] Ch.1.6*