Kullback-Leibler (KL) Divergence

Ravi Kothari, Ph.D.
ravi.kothari@ashoka.edu.in

"I always begin a room with a rug; it is literally the foundation of the space. I then go on to the furniture" – Lee Radziwill

What is it (A Probabilistic Perspective)?

What is it (A Probabilistic Perspective)?

- Indicates how well a distribution $q(x)$ approximates a distribution $p(x)$,

$$D_{\mathrm{KL}}(p(x) \parallel q((x)) = H(p(x), q(x)) - H(p(x)) \tag{1}$$

What is it (A Probabilistic Perspective)?

- Indicates how well a distribution $q(x)$ approximates a distribution $p(x)$,

$$D_{\mathrm{KL}}(p(x) \parallel q((x)) = H(p(x), q(x)) - H(p(x)) \tag{1}$$

- To estimate how well $q(x)$ approximates $p(x)$, the likelihood ratio seems like a good place to start,

$$\mathrm{LR} = \frac{p(x)}{q(x)} \tag{2}$$

What is it (A Probabilistic Perspective)?

- Indicates how well a distribution $q(x)$ approximates a distribution $p(x)$,

$$D_{\mathrm{KL}}(p(x) \parallel q((x)) = H(p(x), q(x)) - H(p(x)) \tag{1}$$

- To estimate how well $q(x)$ approximates $p(x)$, the likelihood ratio seems like a good place to start,

$$\mathrm{LR} = \frac{p(x)}{q(x)} \tag{2}$$

- If the data points are independent then one can define an average likelihood ratio as,

$$\mathrm{LR} = \frac{1}{N} \prod_{i=1}^{N} \frac{p(x_i)}{q(x_i)} \tag{3}$$

- Taking the log,

$$\log \mathrm{LR} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(x_i)}{q(x_i)} \tag{4}$$

- Taking the log,

$$\log \mathrm{LR} = \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(x_i)}{q(x_i)} \tag{4}$$

- A value of $\log \mathrm{LR}$ greater than 0 implies that $p(x)$ better models the data; a value less than 0 implies that $q(x)$ better models the data, and 0 implies that $q(x)$ and $p(x)$ model the data equally well

- Taking the log,

$$\log \mathrm{LR} = \frac{1}{N}\sum_{i=1}^{N} \log \frac{p(x_i)}{q(x_i)} \tag{4}$$

- A value of $\log \mathrm{LR}$ greater than 0 implies that $p(x)$ better models the data; a value less than 0 implies that $q(x)$ better models the data, and 0 implies that $q(x)$ and $p(x)$ model the data equally well
- In the limiting case,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \log \frac{p(x_i)}{q(x_i)} = E_{x\sim p(x)}\left[\log \frac{p(x)}{q(x)}\right] \tag{5}$$

which is defined as the KL-Divergence $D_{\mathrm{KL}}(p(x) \parallel q((x))$

What is it (An Information Theoretic Perspective)?

What is it (An Information Theoretic Perspective)?

- Indicates how well a distribution $q(x)$ approximates a distribution $p(x)$,

$$D_{\mathrm{KL}}(p(x) \parallel q((x)) = H(p(x), q(x)) - H(p(x)) \tag{6}$$

What is it (An Information Theoretic Perspective)?

- Indicates how well a distribution $q(x)$ approximates a distribution $p(x)$,

$$D_{\text{KL}}(p(x) \parallel q((x)) = H(p(x), q(x)) - H(p(x)) \tag{6}$$

- In this form, we can say that the KL-Divergence computes the cross-entropy minus the entropy

What is it (An Information Theoretic Perspective)?

- Indicates how well a distribution $q(x)$ approximates a distribution $p(x)$,

$$D_{\mathrm{KL}}(p(x) \parallel q((x)) = H(p(x), q(x)) - H(p(x)) \tag{6}$$

- In this form, we can say that the KL-Divergence computes the cross-entropy minus the entropy
- Recall,

$$\begin{aligned} H(p(x), q(x)) &= E_{x \sim p(x)}[-\log q(x)] \\ H(p(x)) &= E_{x \sim p(x)}[-\log p(x)] \end{aligned} \tag{7}$$

$$
\begin{aligned}
D_{\mathrm{KL}}(p(x) \parallel q(x)) &= E_{x\sim p(x)}[-\log q(x)] - E_{x\sim p(x)}[-\log p(x)] \\
&\geq E_{x\sim p(x)}\left[-\log q(x) - (-\log p(x))\right] \\
&\geq E_{x\sim p(x)}\left[-\log q(x) + \log p(x)\right] \\
&\geq E_{x\sim p(x)}\left[\log \frac{p(x)}{q(x)}\right] \\
&\geq \sum_{x} p(x) \log \frac{p(x)}{q(x)}
\end{aligned} \tag{8}
$$

Characteristics

Characteristics

- It is non-negative (if $p(x) = q(x)$, then $\log 1 = 0$. If $p(x) \neq q(x)$,

$$D_{\text{KL}}(p(x) \parallel q(x)) = E_{x\sim p(x)}\left[-\log \frac{q(x)}{p(x)}\right] \tag{9}$$

Since, $-\log$ is a convex function, we can use Jensen's inequality,

$$\begin{aligned} D_{\text{KL}}(p(x) \parallel q(x)) &\geq -\log E_{x\sim p(x)}\left[\frac{q(x)}{p(x)}\right] \\ &= -\log\left(\int p(x)\frac{q(x)}{p(x)}dx\right) \\ &= -\log\left(\int q(x)dx\right) \\ &= -\log(1) = 0 \end{aligned} \tag{10}$$

Characteristics

- It is non-negative (if $p(x) = q(x)$, then $\log 1 = 0$. If $p(x) \neq q(x)$,

$$D_{\mathrm{KL}}(p(x) \parallel q(x)) = E_{x \sim p(x)} \left[-\log \frac{q(x)}{p(x)} \right] \tag{9}$$

Since, $-\log$ is a convex function, we can use Jensen's inequality,

$$\begin{aligned} D_{\mathrm{KL}}(p(x) \parallel q(x)) &\geq -\log E_{x \sim p(x)} \left[\frac{q(x)}{p(x)} \right] \\ &= -\log \left(\int p(x) \frac{q(x)}{p(x)} dx \right) \\ &= -\log \left(\int q(x) dx \right) \\ &= -\log(1) = 0 \end{aligned} \tag{10}$$

- It is not symmetric (cross-entropy is itself asymmetric)