Information and Coding Theory

MA41024/ MA60020/ MA60262

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 6
January 23, 2023

Kullback Leibler (KL) divergence

Also known as *relative entropy* is a measure of how different two distributions are.

Kullback Leibler (KL) divergence

Also known as *relative entropy* is a measure of how different two distributions are.

Definition Let P and and Q be be two distributions on a sample space $\mathcal{X}$. The KL-divergence between P and Q is defined as:

$$D(P\|Q) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

Kullback Leibler (KL) divergence

Also known as *relative entropy* is a measure of how different two distributions are.

Definition Let P and and Q be be two distributions on a sample space $\mathcal{X}$. The KL-divergence between P and Q is defined as:

$$D(P\|Q) = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right)$$

Example Suppose $\mathcal{X} = \{a, b, c\}$ with $p(x) = \frac{1}{3}$, $x \in \mathcal{X}$ and $q(a) = \frac{1}{2}$, $q(b) = \frac{1}{2}$, $q(c) = 0$. Then

Kullback Leibler (KL) divergence

Also known as *relative entropy* is a measure of how different two distributions are.

Definition Let P and and Q be be two distributions on a sample space $\mathcal{X}$. The KL-divergence between P and Q is defined as:

$$D(P\|Q) = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right)$$

Example Suppose $\mathcal{X} = \{a, b, c\}$ with $p(x) = \frac{1}{3}$, $x \in \mathcal{X}$ and $q(a) = \frac{1}{2}$, $q(b) = \frac{1}{2}$, $q(c) = 0$. Then

$$\begin{aligned} D(P\|Q) &= \frac{2}{3}\log\frac{2}{3} + \infty = \infty \\ D(Q\|P) &= \log\frac{3}{2} + 0 = \log\frac{3}{2} \end{aligned}$$

KL divergence

$\rightarrow$ $D(P\|Q)$ and $D(Q\|P)$ are necessarily equal

KL divergence

$\rightarrow$ $D(P\|Q)$ and $D(Q\|P)$ are necessarily equal

$\rightarrow$ $D(P\|Q)$ may be infinite

KL divergence

- $\rightarrow$ $D(P\|Q)$ and $D(Q\|P)$ are necessarily equal
- $\rightarrow$ $D(P\|Q)$ may be infinite
- $\rightarrow$ Let $Supp(P) = \{x : p(x) > 0\}$. Then we must have $Supp(P) \subseteq Supp(Q)$ if $D(P\|Q) < \infty$

KL divergence

- $\rightarrow$ $D(P\|Q)$ and $D(Q\|P)$ are necessarily equal
- $\rightarrow$ $D(P\|Q)$ may be infinite
- $\rightarrow$ Let $Supp(P) = \{x : p(x) > 0\}$. Then we must have $Supp(P) \subseteq Supp(Q)$ if $D(P\|Q) < \infty$

Even though the KL-divergence is not symmetric, it is often used as a measure of "dissimilarity" between two distributions

KL divergence

- $\rightarrow$ $D(P\|Q)$ and $D(Q\|P)$ are necessarily equal
- $\rightarrow$ $D(P\|Q)$ may be infinite
- $\rightarrow$ Let $Supp(P) = \{x : p(x) > 0\}$. Then we must have $Supp(P) \subseteq Supp(Q)$ if $D(P\|Q) < \infty$

Even though the KL-divergence is not symmetric, it is often used as a measure of "dissimilarity" between two distributions

Lemma Let P and Q be distributions on a finite space $\mathcal{X}$. Then $D(P\|Q) \geq 0$ with equality if and only if $P = Q$.

KL divergence

→ $D(P\|Q)$ and $D(Q\|P)$ are necessarily equal

→ $D(P\|Q)$ may be infinite

→ Let $Supp(P) = \{x : p(x) > 0\}$. Then we must have $Supp(P) \subseteq Supp(Q)$ if $D(P\|Q) < \infty$

Even though the KL-divergence is not symmetric, it is often used as a measure of "dissimilarity" between two distributions

Lemma Let P and Q be distributions on a finite space $\mathcal{X}$. Then $D(P\|Q) \geq 0$ with equality if and only if $P = Q$.

$$\begin{aligned} D(P\|Q) = \sum_x p(x) \log \frac{p(x)}{q(x)} &= \sum_{x \in Supp(P)} p(x) \log \frac{p(x)}{q(x)} \\ &\geq -\log \left(\sum_{x \in Supp(P)} p(x) \cdot \frac{q(x)}{p(x)} \right) \\ &= -\log \left(\sum_{x \in Supp(P)} q(x) \right) \geq -\log 1 = 0 \end{aligned}$$

Kl divergence

Homework When $D(P\|Q) = 0$ implies $p(x) = q(x)$ for all $x \in Supp(P)$, which gives $p(x) = q(x)\ \forall x \in \mathcal{X}$

Kl divergence

Homework When $D(P\|Q) = 0$ implies $p(x) = q(x)$ for all $x \in Supp(P)$, which gives $p(x) = q(x)\ \forall x \in \mathcal{X}$

Chain rule for KL-divergence Let $P(X, Y)$ and $Q(X, Y)$ be two distributions for a pair of variables X and Y. Then,

$$\begin{aligned} & D(P(X,Y)\|Q(X,Y)) \\ = \; & D(P(X)\|Q(X)) + \mathbb{E}_x[D(P(Y|X=x)\|Q(Y|X=x))] \\ = \; & D(P(X)\|Q(X)) + D(P(Y|X)\|Q(Y|X)) \end{aligned}$$

KL divergence

$$\begin{aligned} & D(P(X,Y)\|Q(X,Y)) \\ = & \sum_{x,y} p(x,y)\log\frac{p(x,y)}{q(x,y)} \end{aligned}$$

KL divergence

$$
\begin{aligned}
& D(P(X,Y)\|Q(X,Y)) \\
= & \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
= & \sum_{x,y} p(x)p(y|x) \log \left(\frac{p(x)}{q(x)} \cdot \frac{p(y|x)}{q(y|x)} \right)
\end{aligned}
$$

KL divergence

$$
\begin{aligned}
& D(P(X,Y)\|Q(X,Y)) \\
= & \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
= & \sum_{x,y} p(x)p(y|x) \log \left(\frac{p(x)}{q(x)} \cdot \frac{p(y|x)}{q(y|x)} \right) \\
= & \sum_{x} p(x) \log \frac{p(x)}{q(x)} \sum_{y} p(y|x) + \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{p(y|x)}{q(y|x)}
\end{aligned}
$$

KL divergence

$$
\begin{aligned}
& D(P(X,Y)\|Q(X,Y)) \\
= \;& \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
= \;& \sum_{x,y} p(x)p(y|x) \log \left(\frac{p(x)}{q(x)} \cdot \frac{p(y|x)}{q(y|x)} \right) \\
= \;& \sum_{x} p(x) \log \frac{p(x)}{q(x)} \sum_{y} p(y|x) + \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{p(y|x)}{q(y|x)} \\
= \;& D(P(X)\|Q(X)) + \sum_{x} p(x) \cdot D(P(Y|X=x)\|Q(Y|X=x))
\end{aligned}
$$

KL divergence

$$
\begin{aligned}
& D(P(X,Y)\|Q(X,Y)) \\
= \quad & \sum_{x,y} p(x,y)\log\frac{p(x,y)}{q(x,y)} \\
= \quad & \sum_{x,y} p(x)p(y|x)\log\left(\frac{p(x)}{q(x)}\cdot\frac{p(y|x)}{q(y|x)}\right) \\
= \quad & \sum_{x} p(x)\log\frac{p(x)}{q(x)}\sum_{y} p(y|x) + \sum_{x} p(x)\sum_{y} p(y|x)\log\frac{p(y|x)}{q(y|x)} \\
= \quad & D(P(X)\|Q(X)) + \sum_{x} p(x)\cdot D(P(Y|X=x)\|Q(Y|X=x)) \\
= \quad & D(P(X)\|Q(x)) + D(P(Y|X)\|Q(Y|X))
\end{aligned}
$$

KL divergence

$$
\begin{aligned}
& D(P(X,Y)\|Q(X,Y)) \\
= \; & \sum_{x,y} p(x,y)\log\frac{p(x,y)}{q(x,y)} \\
= \; & \sum_{x,y} p(x)p(y|x)\log\left(\frac{p(x)}{q(x)}\cdot\frac{p(y|x)}{q(y|x)}\right) \\
= \; & \sum_{x} p(x)\log\frac{p(x)}{q(x)}\sum_{y} p(y|x) + \sum_{x} p(x)\sum_{y} p(y|x)\log\frac{p(y|x)}{q(y|x)} \\
= \; & D(P(X)\|Q(X)) + \sum_{x} p(x)\cdot D(P(Y|X=x)\|Q(Y|X=x)) \\
= \; & D(P(X)\|Q(x)) + D(P(Y|X)\|Q(Y|X))
\end{aligned}
$$

Note If $P(X,Y) = P_1(X)P_2(Y)$ and $Q(X,Y) = Q_1(X)Q_2(Y)$ then

$$D(P\|Q) = D(P_1\|Q_1) + D(P_2\|Q_2)$$

Kl divergence

Interpretation of KL divergence in terms of source coding

$$D(P\|Q) = \sum_x p(x)\log\frac{p(x)}{q(x)} = \sum_x p(x)\log\frac{1}{q(x)} - \sum_x p(x)\log\frac{1}{p(x)}$$

Kl divergence

Interpretation of KL divergence in terms of source coding

$$D(P\|Q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_x p(x) \log \frac{1}{q(x)} - \sum_x p(x) \log \frac{1}{p(x)}$$

→ This can be interpreted as the number of extra bits we use (on average) if we designed a code according to the distribution P, but used it to communicate outcomes of a random variable X distributed according to Q

Kl divergence

Interpretation of KL divergence in terms of source coding

$$D(P\|Q) = \sum_x p(x)\log\frac{p(x)}{q(x)} = \sum_x p(x)\log\frac{1}{q(x)} - \sum_x p(x)\log\frac{1}{p(x)}$$

→ This can be interpreted as the number of extra bits we use (on average) if we designed a code according to the distribution P, but used it to communicate outcomes of a random variable X distributed according to Q

→ The first term in the RHS, which corresponds to the average number of bits used by the "wrong" encoding, is also referred to as *cross entropy*