Lecture-15: Divergence

1 KL divergence

Definition 1.1. Let $(\mathcal{X},\mathcal{F})$ be a measurable space, we define the set of probability measures on $\mathcal{X}$ as

$$\mathcal{M}(\mathcal{X}) \triangleq \left\{P \in [0,1]^{\mathcal{F}} : P \text{ satisfies probability axioms } \right\}.$$

For $P,Q \in \mathcal{M}(\mathcal{X})$, we say P is *absolutely continuous* w.r.t. Q and denoted by $P \ll Q$ if $Q(E) = 0$ implies $P(E) = 0$ for all measurable $E \in \mathcal{F}$. If $P \ll Q$, then *Radon-Nikodym theorem* show that there exists a function $g : \mathcal{X} \to \mathbb{R}_+$ alled a *relative density* or a *Radon-Nikodym derivative* of P w.r.t. Q and denoted by $\frac{dP}{dQ} \triangleq g$, such that for any measurable set $E \in \mathcal{F}$,

$$P(E) = \int_E g dQ.$$

Remark 1. Note that $\frac{dP}{dQ}$ may not be unique. In the simple cases, $\frac{dP}{dQ}$ is the likelihood ratio.

(a) For discrete distributions, we can just take $\frac{dP}{dQ}(x)$ to be the ratio of probability mass functions.

(b) For continuous distributions, we can take $\frac{dP}{dQ}(x)$ to be the ratio of probability density functions.

Definition 1.2 (Kullback-Leibler (KL) divergence). Adopting the convention $0\ln 0 = 0$, we can define the *KL divergence* or *relative entropy* between any $P,Q \in \mathcal{M}(\mathcal{X})$ with Q being the reference measure, as

$$D(P\|Q) \triangleq \begin{cases} \mathbb{E}_P \ln \frac{dP}{dQ} = \mathbb{E}_Q \left[\frac{dP}{dQ} \ln \frac{dP}{dQ}\right], & P \ll Q, \\ +\infty, & P \not\ll Q. \end{cases}$$

2 f-divergence

Definition 2.1 (f-divergence). Let $f : (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$ and define $f(0) \triangleq \lim_{x\downarrow 0} f(x), f'(\infty) \triangleq \lim_{x\downarrow 0} x f\left(\frac{1}{x}\right)$. Let $P,Q \in \mathcal{M}(\mathcal{X})$ for a measurable space $(\mathcal{X},\mathcal{F})$. If $P \ll Q$ then the f-divergence is defined as

$$D_f(P\|Q) \triangleq \mathbb{E}_Q f\left(\frac{dP}{dQ}\right).$$

Suppose for some common dominating measure μ such that $P \ll \mu$ and $Q \ll \mu$, we have relative densities $q \triangleq \frac{dQ}{d\mu}$ and $p \triangleq \frac{dP}{d\mu}$, then we have

$$D_f(P\|Q) = \int_{q>0} q f\left(\frac{p}{q}\right) d\mu + f'(\infty) P\{q = 0\}$$

where the last term is taken to be zero when $P\{q = 0\} = 0$, regardless of the value of $f'(\infty)$ which could be infinite.

Example 2.2 (KL divergence). The map $x \mapsto f(x) \triangleq x \ln x$ results in KL divergence.

Example 2.3 (Total variation). The map $x \mapsto f(x) \triangleq \frac{1}{2}|x-1|$ results in the total variation divergence (distance). For $P,Q \in \mathcal{M}(\mathcal{X})$, we define total variation divergence as

$$\mathrm{TV}(P,Q) \triangleq \frac{1}{2}\mathbb{E}_Q \left|\frac{dP}{dQ} - 1\right| = \frac{1}{2}\int_{\mathcal{X}} |dP - dQ|.$$

Exercise 2.4. Show that $\mathrm{TV}(P,Q) = 1 - \int_{\mathcal{X}} d(P \wedge Q)$ for all $P,Q \in \mathcal{M}(\mathcal{X})$.

Example 2.5 (χ^2-divergence). The map $x \mapsto f(x) \triangleq (x-1)^2$ results in the χ^2 divergence. For $P,Q \in \mathcal{M}(\mathcal{X})$, we define χ^2 divergence as

$$\chi^2(P\|Q) \triangleq \mathbb{E}_Q\Big(\frac{dP}{dQ} - 1\Big)^2 = \int_{\mathcal{X}} \frac{(dP-dQ)^2}{dQ} = \int_{\mathcal{X}} \frac{dP^2}{dQ} - 1.$$

We note that we could have chosen $f(x) \triangleq x^2 - 1$ as well to get the same χ^2 divergence.

Exercise 2.6. Consider two functions $f,h : (0,\infty) \to \mathbb{R}_+$ differing in a linear term, i.e. $h(x) - f(x) = c(x-1)$ for all $x \in (0,\infty)$ and some $c \in \mathbb{R}$. Show that $D_h = D_f$.

Exercise 2.7. Show that $D(P\|Q) \leqslant \ln(1 + \chi^2(P\|Q))$ for all $P,Q \in \mathcal{M}(\mathcal{X})$.

Example 2.8 (Squared Hellinger distance). The map $x \mapsto f(x) \triangleq (1-\sqrt{x})^2$ results in squared Hellinger distance which is defined for any $P,Q \in \mathcal{M}(\mathcal{X})$ as

$$H^2(P,Q) \triangleq \mathbb{E}_Q\Big(1 - \sqrt{\frac{dP}{dQ}}\Big)^2 = \int_{\mathcal{X}} (\sqrt{dQ} - \sqrt{dP})^2 = 2 - 2\int_{\mathcal{X}} \sqrt{dPdQ}.$$

The quantity $B(P,Q) \triangleq \int_{\mathcal{X}} \sqrt{dPdQ}$ is known as the *Bhattacharyya coefficient* or *Hellinger affinity*. Hellinger distance $H : \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}_+$ is defined as $H(P,Q) \triangleq \sqrt{H^2(P,Q)}$ for all $P,Q \in \mathcal{M}(\mathcal{X})$.

Example 2.9 (Le Cam divergence (distance)). The map $x \mapsto f(x) \triangleq \frac{(1-x)^2}{2x+2}$ results in Le Cam divergence (distance) which is defined for any $P,Q \in \mathcal{M}(\mathcal{X})$ as

$$\mathrm{LC}(P,Q) \triangleq \mathbb{E}_Q \frac{(1-\frac{dP}{dQ})^2}{2(1+\frac{dP}{dQ})} = \frac{1}{2}\int_{\mathcal{X}} \frac{(dQ-dP)^2}{dQ+dP}.$$

Example 2.10 (Jensen-Shannon divergence). The map $x \mapsto f(x) \triangleq x \ln \frac{2x}{x+1} + \ln \frac{2}{x+1}$ results in Jensen-Shannon divergence which is defined for any $P,Q \in \mathcal{M}(\mathcal{X})$ as

$$\begin{aligned}\mathrm{JS}(P,Q) &\triangleq \mathbb{E}_P \ln \frac{2\frac{dP}{dQ}}{1+\frac{dP}{dQ}} + \mathbb{E}_Q \ln \frac{2}{1+\frac{dP}{dQ}} = \mathbb{E}_P \ln \frac{dP}{\frac{1}{2}d(P+Q)} + \mathbb{E}_Q \ln \frac{dQ}{\frac{1}{2}d(P+Q)} \\ &= D(P\|\frac{1}{2}(P+Q)) + D(Q\|\frac{1}{2}(P+Q)).\end{aligned}$$

Exercise 2.11. Show the following maps $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}_+$ define a metric on the space of probability distributions $\mathcal{M}(\mathcal{X})$.
(a) Total variation distance TV.
(b) Hellinger distance H.
(c) Square root of Le Cam divergence $\sqrt{\mathrm{LC}}$.
(d) Square root of Jensen-Shannon divergence $\sqrt{\mathrm{JS}}$.

3 Conditional divergence

Definition 3.1 (Conditional divergence). Consider measurable spaces $(\mathcal{X},\mathcal{F})$ and $(\mathcal{Y},\mathcal{G})$ and a pair of Markov kernels $P_{Y|X}:\mathcal{X}\to\mathcal{M}(\mathcal{Y})$ and $Q_{Y|X}:\mathcal{X}\to\mathcal{M}(\mathcal{Y})$, and also a probability measure P_X on $\mathcal{X}$. Assuming $(\mathcal{Y},\mathcal{G})$ is standard Borel measurable space, i.e. $\mathcal{G}\triangleq\mathcal{B}(\mathcal{Y})$, we define

$$D(P_{Y|X}\|Q_{Y|X}\mid P_X)\triangleq\mathbb{E}_{x\sim P_X}[D(P_{Y|X=x}\|Q_{Y|X=x})].$$

We observe that as usual in Lebesgue integration it is possible that a conditional divergence is finite even though $D(P_{Y|X=x}\|Q_{Y|X=x})=\infty$ for some x in a P_X-negligible set.

Theorem 3.2 (Chain rule). *For any pair of measures $P_{X,Y}$ and $Q_{X,Y}$ we have*

$$D(P_{X,Y}\|Q_{X,Y})=D(P_{Y|X}\|Q_{Y|X}\mid P_X)+D(P_X\|Q_X),$$

regardless of the versions of conditional distributions $P_{Y|X}$ and $Q_{Y|X}$ one chooses.

Proof. Recall that $P_{X,Y}=P_XP_{Y|X}$ and $Q_{X,Y}=Q_XQ_{Y|X}$. If $P_X\not\ll Q_X$ then $P_{X,Y}\not\ll Q_{X,Y}$ and both sides of chain rule equation are infinity. Thus, we can assume $P_X\ll Q_X$ without any loss of generality, and define relative density $\lambda_P\triangleq\frac{dP_X}{dQ_X}\in\mathbb{R}_+^{\mathcal{X}}$. We next define a kernel $R_{Y|X}:\mathcal{X}\to\mathcal{M}(\mathcal{Y})$ that is a mixture of kernels $R_{Y|X}\triangleq\frac{1}{2}P_{Y|X}+\frac{1}{2}Q_{Y|X}$, such that $P_{Y|X}\ll R_{Y|X}$ and $Q_{Y|X}\ll R_{Y|X}$. We write the corresponding relative densities for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$, as

$$f_P(y\mid x)\triangleq\frac{dP_{Y|X=x}}{dR_{Y|X=x}}(y),\qquad f_Q(y\mid x)\triangleq\frac{dQ_{Y|X=x}}{dR_{Y|X=x}}(y).$$

Defining $R_{X,Y}\triangleq Q_XR_{Y|X}$, we observe that $P_{X,Y}\ll R_{X,Y}$ and $Q_{X,Y}\ll R_{X,Y}$, and we can write down the corresponding relative densities or all $(x,y)\in\mathcal{X}\times\mathcal{Y}$, as

$$\frac{dP_{X,Y}}{dR_{X,Y}}(x,y)=\lambda_P(x)f_P(y\mid x),\qquad\frac{dQ_{X,Y}}{dR_{X,Y}}(x,y)=f_Q(y\mid x).$$

From the linearity of expectation, we can write the following equality

$$D(P_{X,Y}\|Q_{X,Y})=\mathbb{E}_{P_{X,Y}}\ln\frac{dP_{X,Y}}{dQ_{X,Y}}=\mathbb{E}_{P_{X,Y}}\ln\frac{\lambda_P(X)f_P(Y\mid X)}{f_Q(Y\mid X)}=\mathbb{E}_{P_{X,Y}}\ln\lambda_P(X)+\mathbb{E}_{P_{X,Y}}\ln\frac{f_P(Y\mid X)}{f_Q(Y\mid X)}.$$

The result follows from the observation that $E_{P_{X,Y}}\ln\lambda_P(X)=E_{P_X}\ln\lambda_P(X)=D(P_X\|Q_X)$, and the definition of conditional divergence which implies that

$$\mathbb{E}_{P_{X,Y}}\ln\frac{f_P(Y\mid X)}{f_Q(Y\mid X)}=\mathbb{E}_{x\sim P_X}\mathbb{E}_{P_{Y|X=x}}\ln\frac{dP_{Y|X=x}}{dQ_{Y|X=x}}=D(P_{X|Y}\|Q_{X|Y}\mid P_X).$$

□

4 Data processing inequality

Theorem 4.1 (Data processing inequality). *Consider two input distributions $P_X,Q_X\in\mathcal{M}(\mathcal{X})$ and a common Markov kernel $P_{Y|X}:\mathcal{X}\to\mathcal{M}(\mathcal{Y})$ such that the joint distributions are $P_{X,Y}=P_XP_{Y|X}$ and $Q_{X,Y}=Q_XP_{Y|X}$, and the corresponding output marginal distributions $P_Y\triangleq\int_{\mathcal{X}}dP_X(x)P_{Y|X=x}$ and $Q_Y\triangleq\int_{\mathcal{X}}dQ_X(x)P_{Y|X=x}$. Then $D(P_Y\|Q_Y)\leqslant D(P_X\|Q_X)$.*

Proof. The result follows from the chain rule of KL divergence. That is,

$$D(P_{X,Y}\|Q_{X,Y})=D(P_{X|Y}\|Q_{X|Y}\mid P_Y)+D(P_Y\|Q_Y)=D(P_{Y|X}\|Q_{Y|X}\mid P_X)+D(P_X\|Q_X).$$

Since $Q_{Y|X}=P_{Y|X}$, and KL divergence is always positive, we get the result. □