Kullback-Leiblerova divergence

Jan Kracík
jan.kracik@vsb.cz

Divergence v teorii pravděpodobnosti

- vyjadřují míru nepodobnosti (odchylku, “vzdálenost”) mezi pr. distribucemi
- různé třídy divergencí (f-divergence, Bregmanovy divergence, . . .)
- nejčastěji užívaná je Kullback-Leiblerova divergence

Uvažujme diskrétní náhodnou veličinu X s hodnotami v $\mathcal{X} = \{1, 2, \ldots, n\}$.
Pro libovolné pravděpodobnostní funkce $p(x), q(x)$ n. v. X definujeme *Kullback-Leiblerovu (KL) divergenci* vztahem

$$D(p(x)||q(x)) = \begin{cases} \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} \\ \qquad \text{jestliže } (\forall x \in \mathcal{X} : q(x) = 0 \Rightarrow p(x) = 0) \\ \\ +\infty \qquad \text{v ostatních případech} \end{cases},$$

s využitím konvence $0 \ln 0 = 0, 0 \ln \frac{0}{0} = 0$.

Obecnější definice KL divergence

Uvažujme pravděpodobnostní prostor $(\Omega, \mathcal{A})$, na něm pravděpodobnostní míry P, Q. KL divergenci pravděpodobností P, Q definujeme následovně:

$$D(P||Q) = \begin{cases} \int_\Omega \ln \frac{dP}{dQ} dP & \text{jestliže } P << Q \\ \\ +\infty & \text{jinak} \end{cases},$$

kde $\frac{dP}{dQ}$ je Radonova-Nikodymova derivace P vzhledem ke Q, $P << Q$ značí absolutní spojitost míry P vzhledem ke Q, tj.

$$P << Q \Leftrightarrow (\forall A \in \mathcal{A} : Q(A) = 0 \Rightarrow P(A) = 0).$$

Vlastnosti KL divergence

Označme $\mathcal{F}$ množinu všech pravděpodobnostních funkcí n.v. X.

- nezápornost: $\forall p, q \in \mathcal{F} : D(p||q) \geq 0$
  ...z Jensenovy nerovnosti
- $D(p||q) = 0 \Leftrightarrow (\forall x \in \mathcal{X} : p(x) = q(x))$

Vlastnosti KL divergence

Označme $\mathcal{F}$ množinu všech pravděpodobnostních funkcí n.v. X.

- nezápornost: $\forall p, q \in \mathcal{F} : D(p||q) \geq 0$
  ...z Jensenovy nerovnosti
- $D(p||q) = 0 \Leftrightarrow (\forall x \in \mathcal{X} : p(x) = q(x))$
- není symetrická: $D(p||q) \neq D(q||p)$
- nesplňuje trojúhelníkovou nerovnost:
  $D(p||r) \not\leq D(p||q) + D(q||r)$

Vlastnosti KL divergence

Označme $\mathcal{F}$ množinu všech pravděpodobnostních funkcí n.v. X.

- nezápornost: $\forall p, q \in \mathcal{F} : D(p||q) \geq 0$
  ...z Jensenovy nerovnosti
- $D(p||q) = 0 \Leftrightarrow (\forall x \in \mathcal{X} : p(x) = q(x))$
- není symetrická: $D(p||q) \neq D(q||p)$
- nesplňuje trojúhelníkovou nerovnost:
  $D(p||r) \not\leq D(p||q) + D(q||r)$
- konvexní v obou argumentech: $\forall p, q, r \in \mathcal{F}, \forall \alpha \in [0, 1]$:

$$\begin{aligned} D(\alpha p + (1-\alpha)q||r) &\leq \alpha D(p||r) + (1-\alpha)D(q||r) \\ D(p||\alpha q + (1-\alpha)r) &\leq \alpha D(p||q) + (1-\alpha)D(p||r) \end{aligned}$$

Informační monotonie

Uvažujme disjunktní dělení množiny $\mathcal{X}$: $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$, tj. $M_i \subset \mathcal{X}, \cup_{i=1}^{m} M_i = \mathcal{X}, M_i \cap M_j = \emptyset$ pro $i \neq j$.
Definujme n.v. Y s hodnotami v $\mathcal{Y} = \{1, 2, \ldots, m\}$

$$Y = y \Leftrightarrow X \in M_y.$$

Informační monotonie

Uvažujme disjunktní dělení množiny $\mathcal{X}$: $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$,
tj. $M_i \subset \mathcal{X}, \cup_{i=1}^{m} M_i = \mathcal{X}, M_i \cap M_j = \emptyset$ pro $i \neq j$.
Definujme n.v. Y s hodnotami v $\mathcal{Y} = \{1, 2, \ldots, m\}$

$$Y = y \Leftrightarrow X \in M_y.$$

Pro $p(x), q(x)$ dostaneme odpovídající pr.funkce n.v. Y

$$\begin{aligned} \tilde{p}(y) &= P(X \in M_y) = \sum_{x \in M_y} p(x), \\ \tilde{q}(y) &= Q(X \in M_y) = \sum_{x \in M_y} q(x). \end{aligned}$$

Informační monotonie

Uvažujme disjunktní dělení množiny $\mathcal{X}$: $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$, tj. $M_i \subset \mathcal{X}, \cup_{i=1}^{m} M_i = \mathcal{X}, M_i \cap M_j = \emptyset$ pro $i \neq j$.
Definujme n.v. Y s hodnotami v $\mathcal{Y} = \{1, 2, \ldots, m\}$

$$Y = y \Leftrightarrow X \in M_y.$$

Pro $p(x), q(x)$ dostaneme odpovídající pr.funkce n.v. Y

$$\begin{aligned} \tilde{p}(y) &= P(X \in M_y) = \sum_{x \in M_y} p(x), \\ \tilde{q}(y) &= Q(X \in M_y) = \sum_{x \in M_y} q(x). \end{aligned}$$

Pro $D(p||q)$ a $D(\tilde{p}||\tilde{q})$ potom platí:

$$D(\tilde{p}||\tilde{q}) \leq D(p||q).$$

Důkaz: Pro $D(p||q) = +\infty$ triviální. Předpokládejme, že $D(p||q) < +\infty$.

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} = \sum_{y \in \mathcal{Y}} \sum_{x \in M_y} p(x) \ln \frac{p(x)}{q(x)}$$

Důkaz: Pro $D(p||q) = +\infty$ triviální. Předpokládejme, že $D(p||q) < +\infty$.

$$
\begin{aligned}
D(p||q) &= \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} = \sum_{y \in \mathcal{Y}} \sum_{x \in M_y} p(x) \ln \frac{p(x)}{q(x)} \\
&= \sum_{y \in \mathcal{Y}} \sum_{x \in M_y} \tilde{p}(y) \frac{p(x)}{\tilde{p}(y)} \left(\ln \frac{\frac{p(x)}{\tilde{p}(y)}}{\frac{q(x)}{\tilde{q}(y)}} + \ln \frac{\tilde{p}(y)}{\tilde{q}(y)} \right)
\end{aligned}
$$

Důkaz: Pro $D(p||q) = +\infty$ triviální. Předpokládejme, že $D(p||q) < +\infty$.

$$
\begin{aligned}
D(p||q) &= \sum_{x\in\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} = \sum_{y\in\mathcal{Y}} \sum_{x\in M_y} p(x) \ln \frac{p(x)}{q(x)} \\
&= \sum_{y\in\mathcal{Y}} \sum_{x\in M_y} \tilde{p}(y) \frac{p(x)}{\tilde{p}(y)} \left(\ln \frac{\frac{p(x)}{\tilde{p}(y)}}{\frac{q(x)}{\tilde{q}(y)}} + \ln \frac{\tilde{p}(y)}{\tilde{q}(y)} \right) \\
&= \sum_{y\in\mathcal{Y}} \tilde{p}(y) D(p(x|X \in M_y)||q(x|X \in M_y)) + D(\tilde{p}||\tilde{q})
\end{aligned}
$$

Důkaz: Pro $D(p||q) = +\infty$ triviální. Předpokládejme, že $D(p||q) < +\infty$.

$$
\begin{aligned}
D(p||q) &= \sum_{x\in\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} = \sum_{y\in\mathcal{Y}} \sum_{x\in M_y} p(x) \ln \frac{p(x)}{q(x)} \\
&= \sum_{y\in\mathcal{Y}} \sum_{x\in M_y} \tilde{p}(y) \frac{p(x)}{\tilde{p}(y)} \left(\ln \frac{\frac{p(x)}{\tilde{p}(y)}}{\frac{q(x)}{\tilde{q}(y)}} + \ln \frac{\tilde{p}(y)}{\tilde{q}(y)} \right) \\
&= \sum_{y\in\mathcal{Y}} \tilde{p}(y) D(p(x|X \in M_y)||q(x|X \in M_y)) + D(\tilde{p}||\tilde{q}) \\
&\geq D(\tilde{p}||\tilde{q})
\end{aligned}
$$

□

Důkaz: Pro $D(p||q) = +\infty$ triviální. Předpokládejme, že $D(p||q) < +\infty$.

$$
\begin{aligned}
D(p||q) &= \sum_{x\in\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} = \sum_{y\in\mathcal{Y}} \sum_{x\in M_y} p(x) \ln \frac{p(x)}{q(x)} \\
&= \sum_{y\in\mathcal{Y}} \sum_{x\in M_y} \tilde{p}(y) \frac{p(x)}{\tilde{p}(y)} \left(\ln \frac{\frac{p(x)}{\tilde{p}(y)}}{\frac{q(x)}{\tilde{q}(y)}} + \ln \frac{\tilde{p}(y)}{\tilde{q}(y)} \right) \\
&= \sum_{y\in\mathcal{Y}} \tilde{p}(y) D(p(x|X \in M_y)||q(x|X \in M_y)) + D(\tilde{p}||\tilde{q}) \\
&\geq D(\tilde{p}||\tilde{q})
\end{aligned}
$$

□

Kdy bude platit $D(p||q) = D(\tilde{p}||\tilde{q})$?

Uvažujme obecnější úlohu: Mějme statistický model $f(x|\theta), \theta \in \Theta$. Zobrazení $T : \mathcal{X} \to \mathcal{T}$ je *postačující statistikou* modelu $f(x|\theta), \theta \in \Theta$, platí-li

$$\forall \theta \in \Theta : f(x|t, \theta) = f(x|t).$$

N.v. T nese stejnou informaci o θ jako n.v. X.

Uvažujme obecnější úlohu: Mějme statistický model $f(x|\theta), \theta \in \Theta$. Zobrazení $T : \mathcal{X} \to \mathcal{T}$ je *postačující statistikou* modelu $f(x|\theta), \theta \in \Theta$, platí-li

$$\forall \theta \in \Theta : f(x|t, \theta) = f(x|t).$$

N.v. T nese stejnou informaci o θ jako n.v. X.

Označme $\tilde{f}(t|\theta), \theta \in \Theta$ statistický model indukovaný zobrazením T.

Uvažujme obecnější úlohu: Mějme statistický model $f(x|\theta), \theta \in \Theta$. Zobrazení $T : \mathcal{X} \to \mathcal{T}$ je *postačující statistikou* modelu $f(x|\theta), \theta \in \Theta$, platí-li

$$\forall \theta \in \Theta : f(x|t, \theta) = f(x|t).$$

N.v. T nese stejnou informaci o θ jako n.v. X.

Označme $\tilde{f}(t|\theta), \theta \in \Theta$ statistický model indukovaný zobrazením T. Pak platí

$$\forall \theta_1, \theta_2 \in \Theta : D(f(x|\theta_1)||f(x|\theta_2)) = D(\tilde{f}(t|\theta_1)||\tilde{f}(t|\theta_2)).$$

Uvažujme obecnější úlohu: Mějme statistický model $f(x|\theta), \theta \in \Theta$. Zobrazení $T : \mathcal{X} \to \mathcal{T}$ je *postačující statistikou* modelu $f(x|\theta), \theta \in \Theta$, platí-li

$$\forall \theta \in \Theta : f(x|t, \theta) = f(x|t).$$

N.v. T nese stejnou informaci o θ jako n.v. X.

Označme $\tilde{f}(t|\theta), \theta \in \Theta$ statistický model indukovaný zobrazením T. Pak platí

$$\forall \theta_1, \theta_2 \in \Theta : D(f(x|\theta_1)||f(x|\theta_2)) = D(\tilde{f}(t|\theta_1)||\tilde{f}(t|\theta_2)).$$

Kullback-Leiblerova divergence (f-divergence obecně) jsou invariantní vůči zobrazením postačujícími statistikami. Z tohoto pohledu jsou f-divergence přirozeným nástrojem pro porovnání pravděpodobnostních distribucí na rozdíl od např. metrik indukovaných L_p normami na hustotách.

KL divergence pro náhodné vektory

Uvažujme náhodný vektor X, Y s hodnotami v $\mathcal{X} \times \mathcal{Y}$. Pro KL divergenci pr. funkcí $p(x, y), q(x, y)$ platí:

$$\begin{aligned} D(p(x,y)||q(x,y)) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y) \ln \frac{p(x,y)}{q(x,y)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x|y)p(y) \left(\ln \frac{p(x|y)}{q(x|y)} + \ln \frac{p(y)}{q(y)} \right) \\ &= D(p(x,y)||q(x|y)) + D(p(y)||q(y)), \end{aligned}$$

kde

$$D(p(x,y)||q(x|y)) = E_{p(y)}\left[D(p(x|y)||q(x|y)\right]$$

je *podmíněná KL divergence*.

Maximálně věrohodný odhad (MLE)

Uvažujme statistický model $f(x|\theta), \theta \in \Theta$, data $x_1, \ldots, x_t$ realizace i.i.d. n.v. s rozdělením $f(x|\theta_0)$, pro nějaké (neznámé) $\theta_0 \in \Theta$.

Maximálně věrohodný odhad (MLE)

Uvažujme statistický model $f(x|\theta), \theta \in \Theta$, data $x_1, \ldots, x_t$ realizace i.i.d. n.v. s rozdělením $f(x|\theta_0)$, pro nějaké (neznámé) $\theta_0 \in \Theta$.

Logaritmická věrohodnostní funkce (pro data $x_1, \ldots, x_t$):

$$l(\theta|x_1, \ldots, x_t) = \ln f(x_1, \ldots, x_t|\theta) = \ln \prod_{\tau=1}^{t} f(x_\tau|\theta).$$

Maximálně věrohodný odhad (MLE)

Uvažujme statistický model $f(x|\theta), \theta \in \Theta$, data $x_1, \ldots, x_t$ realizace i.i.d. n.v. s rozdělením $f(x|\theta_0)$, pro nějaké (neznámé) $\theta_0 \in \Theta$.

Logaritmická věrohodnostní funkce (pro data $x_1, \ldots, x_t$):

$$l(\theta|x_1, \ldots, x_t) = \ln f(x_1, \ldots, x_t|\theta) = \ln \prod_{\tau=1}^{t} f(x_\tau|\theta).$$

Maximálně věrohodný odhad $\hat{\theta}_{MLE}$ je pak definován jako

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} l(\theta|x_1, \ldots, x_t).$$

KL divergence a MLE

Označme $r(x)$ empirickou pravděpodobnostní funkci z $x_1, \ldots, x_t$, tj.

$$r(x) = \frac{1}{t} \sum_{\tau=1}^{t} \delta(x, x_\tau),$$

kde $\delta(i,j) = 1$ pro $i = j$, $\delta(i,j) = 0$ pro $i \neq j$.
Log. věrohodnostní funkci pak lze vyjádřit jako

$$\begin{aligned} l(\theta|x_1, \ldots, x_t) &= \ln \prod_{\tau=1}^{t} f(x_\tau|\theta) = \sum_{\tau=1}^{t} \ln f(x_\tau|\theta) \\ &= t \sum_{x \in \mathcal{X}} r(x) \ln f(x|\theta) = -t(H(r) + D(r(x)||f(x|\theta))), \end{aligned}$$

kde $H(r) = E_r[-\ln r(x)]$ je informační entropie pr. funkce $r(x)$ (nezávislá na θ).

Odtud plyne

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} l(\theta | x_1, \ldots, x_t) = \arg\min_{\theta \in \Theta} D(r(x) || f(x|\theta)).$$

Maximálně věrohodný odhad tedy odpovídá projekci empirické distribuce na model ve smyslu minimalizace KL divergence.

Klasický předpoklad, že data pochází z rozdělení z určitého parametrického modelu bývá zpravidla v praxi nerealistický. Jaký smysl má potom úloha statistického odhadování?

Klasický předpoklad, že data pochází z rozdělení z určitého parametrického modelu bývá zpravidla v praxi nerealistický.
Jaký smysl má potom úloha statistického odhadování?
Pro log. věrohodnost platí

$$l(\theta|x_1,\ldots,x_t) = t\left(\frac{1}{t}\sum_{\tau=1}^{t}\ln f(x_\tau|\theta)\right).$$

Podle zákona velkých čísel s rostoucím počtem dat, tj. $t \to \infty$,

$$\frac{1}{t}\sum_{\tau=1}^{t}\ln f(x_\tau|\theta) \xrightarrow{a.s.} E[\ln f(x|\theta)],$$

kde střední hodnota je brána vzhledem ke skutečné distribuci generující data.

Obdobnou úvahou pak dojdeme k závěru, že

$$\hat{\theta}_{MLE} \xrightarrow{a.s.} \arg\min_{\theta \in \Theta} D(g(x)||f(x|\theta)),$$

kde $g(x)$ je pravděpodobnostní funkce, z níž jsou data generována.

Maximálně věrohodný odhad tedy konverguje k distribuci z modelu, která je nejblíže realitě.

Obdobnou úvahou pak dojdeme k závěru, že

$$\hat{\theta}_{MLE} \xrightarrow{a.s.} \arg\min_{\theta \in \Theta} D(g(x)||f(x|\theta)),$$

kde $g(x)$ je pravděpodobnostní funkce, z níž jsou data generována.

Maximálně věrohodný odhad tedy konverguje k distribuci z modelu, která je nejblíže realitě.

Poznámka k nesymetrii: Distribuce generující data se objevuje jako první argument v KL-divergenci.

Další příklad využití KL divergence:

- vyjadřuje pokles síly testu při chybně specifikovaných hypotézách
- důkazy konvergence iteračních algoritmů, např. EM algoritmus
- součást metod pro volbu modelu (AIC)
- geometrický význam (zvláště pro exponeniální rodiny), za určitých okolností splňuje obdobu Pythagorovy věty

Děkuji za pozornost.

Kullback-Leiblerova (KL) divergence hustot pravděpodobnosti $f(x), g(x)$ náhodné veličiny X je definována vztahem

$$D(f(x)||g(x)) = \begin{cases} \int_{\mathbb{R}} f(x) \ln \frac{f(x)}{g(x)} dx & \text{pokud } g(x) = 0 \Rightarrow f(x) = 0 \\ +\infty & \text{jinak} \end{cases},$$

s využitím konvence $0 \ln 0 = 0, 0 \ln \frac{0}{0} = 0$.

Kullback-Leiblerova (KL) divergence hustot pravděpodobnosti $f(x), g(x)$ náhodné veličiny X je definována vztahem

$$D(f(x)||g(x)) = \begin{cases} \int_{\mathbb{R}} f(x) \ln \frac{f(x)}{g(x)} dx & \text{pokud } g(x) = 0 \Rightarrow f(x) = 0 \\ +\infty & \text{jinak} \end{cases},$$

s využitím konvence $0 \ln 0 = 0, 0 \ln \frac{0}{0} = 0$.

Poznámky:

- definice zahrnuje diskrétní i spojité n.v. s abs. spojitým rozdělením
- lze definovat obecněji pro libovolné pravděpodobnostní míry na pravděpodobnostním prostoru s využitím Radonovy-Nikodymovy derivace