

Divergence entropie

Oddíl 3.2

Podle skript *Teorie informace – text k přednášce*
Štěpán Holub a Michal Kupsa
SkriptaKompletMoodle.pdf, strany 22–25

Obsah přednášky

1. Divergence: definice a význam

Jak porovnat dvě rozdělení na stejné množině?

2. Konvexita a Jensenova nerovnost

Nástroj, ze kterého plynou všechny nerovnosti této kapitoly.

3. Nezápornost divergence

$\mathcal{D}(P \parallel Q) \geq 0$ a rovnost jen pro $P = Q$.

4. Důsledky pro entropii

Horní mez $\log n$, nezápornost vzájemné informace, podmiňování.

Co zůstalo otevřené z oddílu 3.1

Ze součtových vztahů (tvrzení 3.14) známe rovnosti

$$\mathcal{I}(X : Y) = \mathcal{H}(X) - \mathcal{H}(X | Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X, Y).$$

Zatím ale neumíme dokázat žádnou z těchto **nerovností**:

- $\mathcal{I}(X : Y) \geq 0$,
- $\mathcal{H}(X | Y) \leq \mathcal{H}(X)$,
- $\mathcal{H}(X, Y) \leq \mathcal{H}(X) + \mathcal{H}(Y)$.

V čem je potíž

Bodový obsah $\mathfrak{I}_{X:Y}(\omega)$ může být záporný. Znaménko střední hodnoty proto nelze určit člen po členu.

Plán

Zavedeme divergenci dvou rozdělání a dokážeme, že je nezáporná. Všechny tři nerovnosti z toho vyplynou.

Divergence entropie

Definice 3.16

Nechť P a Q jsou rozdělení na **stejně** konečné množině A .

Divergence entropie rozdělení P vzhledem k rozdělení Q je

$$\mathcal{D}(P \parallel Q) = \sum_{a \in s(P)} P(a) \log \frac{P(a)}{Q(a)}, \quad \text{pokud } s(P) \subseteq s(Q).$$

Jinak klademe $\mathcal{D}(P \parallel Q) = +\infty$.

- Sčítáme přes nosič P a vážíme hodnotami P .
- Podmínka $s(P) \subseteq s(Q)$ zaručuje $Q(a) > 0$ ve všech členech.
- Nekonečno nastane, když P připouští hodnotu, kterou Q vylučuje.
- Jednotlivé členy mohou být záporné: stačí $P(a) < Q(a)$.

Jiné názvy: relativní entropie, Kullbackova–Leiblerova divergence.

Význam: „pomýlená entropie“

Pro $s(P) \subseteq s(Q)$ rozepíšeme logaritmus podílu:

$$\begin{aligned}\mathcal{D}(P \parallel Q) &= - \sum_{a \in s(P)} P(a) \log Q(a) + \sum_{a \in s(P)} P(a) \log P(a) \\ &= \underbrace{\sum_{a \in s(P)} P(a) (-\log Q(a))}_{\text{pomýlená entropie}} - \mathcal{H}(P).\end{aligned}$$

- $\mathcal{H}(P)$: výsledku a přisuzujeme správný obsah $-\log P(a)$.
- Pomýlená entropie: výsledky mají rozdělení P , ale přisuzujeme jim obsah $-\log Q(a)$ podle rozdělení Q .

Divergence je rozdíl těchto dvou průměrů

Odpovídá situaci, kdy kódování vhodné pro Q použijeme na data s rozdělením P . Divergence měří, kolik bitů na jeden výsledek tím v průměru ztrácíme.

Příklad: kód pro nesprávné rozdělení

Na množině $A = \{a, b, c\}$ uvažujme

$$P = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right), \quad Q = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right).$$

	a	b	c	průměr vzhledem k P
$-\log P$	1	2	2	$\mathcal{H}(P) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = 1,5$
$-\log Q$	2	2	1	$\frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 1 = 1,75$

$$\mathcal{D}(P \parallel Q) = 1,75 - 1,5 = 0,25 \text{ bitu.}$$

Kontrola z definice:

$$\mathcal{D}(P \parallel Q) = \frac{1}{2} \log 2 + \frac{1}{4} \log 1 + \frac{1}{4} \log \frac{1}{2} = \frac{1}{2} + 0 - \frac{1}{4} = \frac{1}{4}.$$

Nejčastějšímu výsledku a jsme přisoudili zbytečně velký obsah. Úspora u výsledku c to nevyváží.

Divergence není symetrická

Rozdělení z příkladu 3.21

Pro $P = (\frac{1}{5}, \frac{4}{5})$ a $Q = (\frac{2}{3}, \frac{1}{3})$ vychází

$$\mathcal{D}(P \parallel Q) = \frac{1}{5} \log \frac{3}{10} + \frac{4}{5} \log \frac{12}{5} \doteq 0,663,$$

$$\mathcal{D}(Q \parallel P) = \frac{2}{3} \log \frac{10}{3} + \frac{1}{3} \log \frac{5}{12} \doteq 0,737.$$

Krajní příklad

Pro $P = (\frac{1}{2}, \frac{1}{2})$ a $Q = (1, 0)$ je

$$\mathcal{D}(P \parallel Q) = +\infty, \quad \mathcal{D}(Q \parallel P) = 1 \cdot \log \frac{1}{1/2} = 1.$$

- Obecně $\mathcal{D}(P \parallel Q) \neq \mathcal{D}(Q \parallel P)$. Záleží na tom, které rozdělení je skutečné a které jen předpokládané.
- Divergence proto **není vzdálenost** v obvyklém smyslu. Společné mají nezápornost a to, že nula znamená shodu.

Konvexní a konkávní funkce

Definice 3.17

Funkce $f : I \rightarrow \mathbb{R}$ je **konvexní** na intervalu I , pokud pro každé $x, y \in I$ a každé $t \in (0, 1)$ platí

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y).$$

Graf tedy leží pod každou svou sečnou.

Funkce je **striktně konvexní**, pokud pro každé $x \neq y$ platí ostrá nerovnost.

Funkce f je **(striktně) konkávní**, je-li $-f$ (striktně) konvexní.

Kritérium

Je-li $f''(x) > 0$ na otevřeném intervalu I , je f na I striktně konvexní.

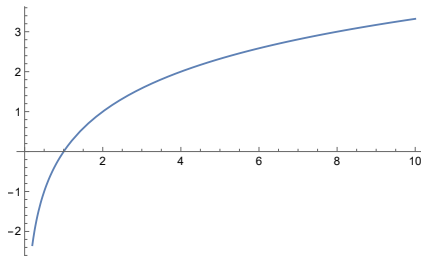
Logaritmus je striktně konkávní

Pro $\log = \log_2$ na $(0, \infty)$ platí

$$(\log x)'' = -\frac{1}{x^2 \ln 2} < 0.$$

Proto:

- $\log x$ je **striktně konkávní**,
- $-\log x$ je **striktně konvexní** a klesající.



Graf funkce $\log x$.

Příklad

Pro $x = 1$, $y = 4$ a $t = \frac{1}{2}$ je

$$\log\left(\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 4\right) = \log 2,5 \doteq 1,322 > \frac{1}{2} \log 1 + \frac{1}{2} \log 4 = 1.$$

Logaritmus průměru je větší než průměr logaritmů.

Jensenova nerovnost

Tvrzení 3.18 a 3.19, ve skriptech bez důkazu

I. Pro čísla. Nechť $f : I \rightarrow \mathbb{R}$ je konvexní, $x_1, \dots, x_n \in I$, $t_1, \dots, t_n \geq 0$ a $\sum_i t_i = 1$. Pak

$$f\left(\sum_{i=1}^n t_i x_i\right) \leq \sum_{i=1}^n t_i f(x_i).$$

II. Pro náhodnou veličinu. Nechť ξ je reálná diskrétní veličina s hodnotami v I a $f : I \rightarrow \mathbb{R}$ je konvexní. Pak

$$f(\mathbb{E}\xi) \leq \mathbb{E}f(\xi).$$

Rovnost pro striktně konvexní f

V I. je pak množina $\{x_i : t_i > 0\}$ jednoprvková. V II. je ξ triviální: nosič P_ξ má jediný prvek.

Pro konkávní funkce platí opačné nerovnosti.

► Důkaz

Divergence je nezáporná

Tvrzení 3.20

Pro každá dvě rozdělení P, Q na konečné množině A platí

$$\mathcal{D}(P \parallel Q) \geq 0, \quad \text{přičemž} \quad \mathcal{D}(P \parallel Q) = 0 \iff P = Q.$$

Význam

Pomýlená entropie je vždy alespoň tak velká jako skutečná:

$$\sum_{a \in s(P)} P(a)(-\log Q(a)) \geq \mathcal{H}(P).$$

Nejmenší průměrný obsah dostaneme, když výsledkům přisuzujeme obsah podle jejich skutečného rozdělení.

To předjímá pozdější výsledky o kódování: nejefektivnější jsou kódy odpovídající skutečnému informačnímu obsahu.

Důkaz 3.20: nerovnost

Je-li $s(P) \not\subseteq s(Q)$, je $\mathcal{D}(P \parallel Q) = +\infty$ a $P \neq Q$. Necht' tedy $s(P) \subseteq s(Q)$. Podíl v logaritmu obrátíme a použijeme Jensenovu nerovnost pro konvexní funkci $-\log$:

$$\begin{aligned}\mathcal{D}(P \parallel Q) &= \sum_{a \in s(P)} P(a) \left(-\log \frac{Q(a)}{P(a)} \right) \\ &\geq -\log \left(\sum_{a \in s(P)} P(a) \cdot \frac{Q(a)}{P(a)} \right) = -\log \left(\sum_{a \in s(P)} Q(a) \right) \\ &\geq -\log \left(\sum_{a \in s(Q)} Q(a) \right) = -\log 1 = 0.\end{aligned}$$

- **První nerovnost:** Jensen s vahami $P(a)$ a čísly $x_a = Q(a)/P(a)$.
- **Druhá nerovnost:** součet přes $s(P)$ je nejvýše součet přes $s(Q)$ a funkce $-\log$ je klesající. □

Důkaz 3.20: kdy nastává rovnost

Je-li $P = Q$, jsou všechny logaritmy nulové a $\mathcal{D}(P \parallel P) = 0$.

Obráceně, necht' $\mathcal{D}(P \parallel Q) = 0$. Pak v obou nerovnostech nastala rovnost.

- **Jensen, striktní konvexita:** podíl je na nosiči konstantní,

$$\frac{Q(a)}{P(a)} = \alpha \quad \text{pro všechna } a \in s(P).$$

- **Druhá nerovnost:** funkce $-\log$ je ryze klesající, takže

$$\sum_{a \in s(P)} Q(a) = 1.$$

Dohromady

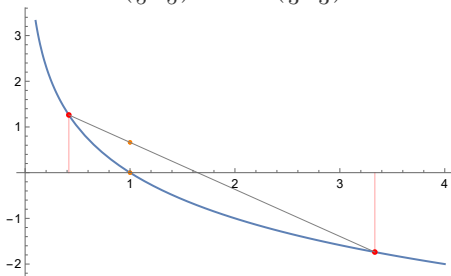
$$\alpha = \sum_{a \in s(P)} P(a) \cdot \alpha = \sum_{a \in s(P)} Q(a) = 1.$$

Tedy $Q = P$ na $s(P)$. Tam se hodnoty Q už sečtou na 1, takže jinde je Q nulové stejně jako P . □

Geometrický pohled na důkaz

Příklad 3.21

Pro $P = (\frac{1}{5}, \frac{4}{5})$, $Q = (\frac{2}{3}, \frac{1}{3})$ má podíl Q/P hodnoty $\frac{10}{3}$ a $\frac{5}{12}$.



Průměr podílu (váhy P):

$$\frac{1}{5} \cdot \frac{10}{3} + \frac{4}{5} \cdot \frac{5}{12} = 1.$$

Průměr $-\log$ podílu:

$$\mathcal{D}(P \parallel Q) \doteq 0,663.$$

Graf funkce $-\log x$ a sečna.

- Průměr hodnot $-\log$ leží **na sečně** nad bodem 1.
- Hodnota funkce v průměru je $-\log 1 = 0$ a leží **na grafu**.
- Sečna konvexní funkce leží nad grafem. Svislá vzdálenost obou bodů je právě $\mathcal{D}(P \parallel Q)$.

Entropie a velikost abecedy

Tvrzení 3.22

Nechť X má hodnoty v n -prvkové množině A . Pak

$$\mathcal{H}(X) \leq \log n.$$

Rovnost nastává, právě když je P_X rovnoměrné na A .

Důkaz

Nechť $U_n(a) = \frac{1}{n}$ pro každé $a \in A$. Pak $s(P_X) \subseteq A = s(U_n)$ a

$$\begin{aligned} 0 \leq \mathcal{D}(P_X \parallel U_n) &= \sum_{a \in s(P_X)} P_X(a) \log \frac{P_X(a)}{1/n} \\ &= \log n \sum_{a \in s(P_X)} P_X(a) + \sum_{a \in s(P_X)} P_X(a) \log P_X(a) \\ &= \log n - \mathcal{H}(X). \end{aligned}$$

Rovnost nastává, právě když $P_X = U_n$.

□

Divergence od rovnoměrného rozdělení

Z důkazu máme užitečný vzorec

$$\mathcal{H}(X) = \log n - \mathcal{D}(P_X \parallel U_n).$$

Entropie je tím menší, čím víc se P_X liší od rovnoměrného rozdělení.

Příklady pro $n = 2$

P_X	$(\frac{1}{2}, \frac{1}{2})$	$(\frac{3}{4}, \frac{1}{4})$	$(1, 0)$
$\mathcal{H}(X)$	1	$\doteq 0,811$	0
$\mathcal{D}(P_X \parallel U_2)$	0	$\doteq 0,189$	1

Souvislost s oddílem 3

Rovnoměrné rozdělení na m hodnotách má entropii $\log m$. Nyní víme, že žádné jiné rozdělení na těchto hodnotách tak velkou entropii nemá.

Vzájemná informace je divergence

Tvrzení 3.23

Pro rozdělení P_X na A a P_Y na B definujeme **součinnové rozdělení** na $A \times B$:

$$(P_X \cdot P_Y)(a, b) = P_X(a)P_Y(b).$$

Takové sdružené rozdělení by veličiny měly, kdyby byly nezávislé.

$$\mathcal{I}(X : Y) = \mathcal{D}(P_{X,Y} \parallel P_X \cdot P_Y).$$

Proto

$$\mathcal{I}(X : Y) \geq 0, \quad \mathcal{I}(X : Y) = 0 \iff X \text{ a } Y \text{ jsou nezávislé.}$$

Vzájemná informace jako „divergence od nezávislosti“

Měří, jak moc se skutečné sdružené rozdělení liší od toho, které by odpovídalo nezávislým veličinám se stejnými marginály.

Důkaz tvrzení 3.23

Nosiče. Je-li $P_{X,Y}(a,b) > 0$, pak $P_X(a) \geq P_{X,Y}(a,b) > 0$ a stejně $P_Y(b) > 0$. Tedy

$$s(P_{X,Y}) \subseteq s(P_X \cdot P_Y)$$

a divergence je dána součtem z definice 3.16.

Rovnost. Porovnáme s definicí 3.12:

$$\begin{aligned}\mathcal{I}(X : Y) &= \sum_{(a,b) \in s(P_{X,Y})} P_{X,Y}(a,b) \log \frac{P_{X,Y}(a,b)}{P_X(a)P_Y(b)} \\ &= \mathcal{D}(P_{X,Y} \parallel P_X \cdot P_Y).\end{aligned}$$

Důsledek. Podle tvrzení 3.20 je $\mathcal{I}(X : Y) \geq 0$. Rovnost nastává, právě když

$$P_{X,Y}(a,b) = P_X(a)P_Y(b) \quad \text{pro všechna } (a,b) \in A \times B,$$

což je definice nezávislosti X a Y .



Průběžný příklad: divergence od nezávislosti

Příklad 3.13

Skutečné $P_{X,Y}$			
	b_0	b_1	P_X
a_0	1/2	0	1/2
a_1	1/4	1/4	1/2
P_Y	3/4	1/4	

Součinnové $P_X \cdot P_Y$			
	b_0	b_1	
a_0	3/8	1/8	1/2
a_1	3/8	1/8	1/2
	3/4	1/4	

$$\mathcal{I}(X : Y) = \frac{1}{2} \log \frac{1/2}{3/8} + \frac{1}{4} \log \frac{1/4}{3/8} + \frac{1}{4} \log \frac{1/4}{1/8} \doteq 0,311.$$

- Stejný výpočet jako v oddílu 3.1, jen jinak čtený.
- Prostřední člen je záporný, celek kladný.
- V opačném pořadí je $\mathcal{D}(P_X \cdot P_Y \parallel P_{X,Y}) = +\infty$: součinnové rozdělení připouští dvojici (a_0, b_1) , skutečné ne.

Subaditivita a podmiňování

Tvrzení 3.24

Pro náhodné veličiny X, Y platí

$$\mathcal{H}(X, Y) \leq \mathcal{H}(X) + \mathcal{H}(Y), \quad \mathcal{H}(X | Y) \leq \mathcal{H}(X).$$

Rovnosti nastávají, právě když jsou X a Y nezávislé.

Důkaz

Ze součtových vztahů (tvrzení 3.14) a z $\mathcal{I}(X : Y) \geq 0$:

$$\mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X, Y) = \mathcal{I}(X : Y) \geq 0,$$

$$\mathcal{H}(X) - \mathcal{H}(X | Y) = \mathcal{I}(X : Y) \geq 0.$$

V obou řádcích nastává rovnost, právě když $\mathcal{I}(X : Y) = 0$, tedy při nezávislosti. □

Tím jsou dokázány všechny tři nerovnosti z úvodu.

Jak nerovnosti číst

- $\mathcal{H}(X, Y) \leq \mathcal{H}(X) + \mathcal{H}(Y)$. Popsat dvojici najednou není dražší než popsat každou složku zvlášť. Ušetříme právě $\mathcal{I}(X : Y)$.
- $\mathcal{H}(X | Y) \leq \mathcal{H}(X)$. Znalost Y nejistotu o X v průměru nezvýší.
- $\mathcal{I}(X : Y) \geq 0$. Společná část informace není záporná, i když bodový obsah $\mathfrak{I}_{X:Y}$ záporný být může.

V průběžném příkladu

$$\underbrace{\mathcal{H}(X, Y)}_{1,5} < \underbrace{\mathcal{H}(X) + \mathcal{H}(Y)}_{\doteq 1,811}, \quad \underbrace{\mathcal{H}(X | Y)}_{\doteq 0,689} < \underbrace{\mathcal{H}(X)}_1.$$

Pozor na jednotlivé podmínky

Pro pevnou hodnotu podmínky nerovnost platit nemusí: v oddílu 3.1 jsme viděli $\mathcal{H}(Y | X = a_1) = 1 > \mathcal{H}(Y) \doteq 0,811$.