

(13) Informácia a entropia

1. Príklad 1: hádzanie mincou, počet otázok na uhádnutie výsledku

Predstavme si, že sa hádže 4 krát mincou a my máme za úlohu uhádnuť výsledok, pričom sa môžeme pýtať, či sa nachádza v nami zvolenej množine a dostaneme odpoveď áno/nie.

Potrebuje na to 4 otázky, postupne nimi zistíme výsledky jednotlivých hodov (je na prvom mieste hlava? je na druhom mieste hlava? atď.)

Každý výsledok má rovnakú pravdepodobnosť, ktorá sa rovná $1/2^4$, preto na odhalenie daného výsledku x potrebujeme $-\log_2(P(X=x))$ otázok.

2. Príklad 2: mená I., počet otázok na uhádnutie výsledku

Náš protihráč si zvolí meno s danými pravdepodobnosťami, my sa ho znovu snažíme uhádnuť, pričom sa môžeme pýtať, či sa nachádza v nami zvolenej množine a dostaneme odpoveď áno/nie.

Meno	Pravdepodobnosť
Adam	1/2
Boris	1/4
Cyril	1/8
Daniel	1/8

Jednotlivými otázkami budeme deliť pravdepodobnosť na polovicu:

- Je to Adam?
 - Ak áno, máme odpoveď pomocou 1 otázky.
 - Ak nie, pýtame sa: je to Boris?
 - Ak áno, máme odpoveď pomocou 2 otázok.
 - Ak nie, pýtame sa: je to Cyril?
 - Ak áno, máme odpoveď pomocou 3 otázok.
 - Ak nie, vieme, že je to Daniel a odpoveď sme získali pomocou 3 otázok.

Stredná hodnota počtu otázok je: $\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 =$

$= P(\text{Adam}) \times [-\log_2(P(\text{Adam}))] + P(\text{Boris}) \times [-\log_2(P(\text{Boris}))] + P(\text{Cyril}) \times [-\log_2(P(\text{Cyril}))] + P(\text{Daniel}) \times [-\log_2(P(\text{Daniel}))]$

3. Príklad 3: hádame zviera, počet otázok na uhádnutie výsledku

<https://math.mit.edu/~sheffield/2019600/Lecture33.pdf>

x	$P\{X = x\}$	$-\log P\{X = x\}$
Dog	1/4	2
Cat	1/4	2
Cow	1/8	3
Pig	1/16	4
Squirrel	1/16	4
Mouse	1/16	4
Owl	1/16	4
Sloth	1/32	5
Hippo	1/32	5
Yak	1/32	5
Zebra	1/64	6
Rhino	1/64	6

Znovu delíme na polovicu:

- Je z množiny {dog, cat}? (pp. množiny je 1/2, pp. zvyšku tiež)
 - Ak áno, pýtame sa: "Je to dog?".
 - Ak áno, máme odpoveď dog na dve otázky.
 - Ak nie, je to cat a tiež sme získali odpoveď na dve otázky.
 - Ak nie, pýtame sa: "Je z množiny {cow, pig, squirrel}?" (pp. množiny je 1/4, pp. zvyšku tiež)
 - Ak áno, pýtame sa: "Je to cow?" (pp. je 1/8, pp. zvyšku tiež)
 - Ak áno, odpoveď cow sme získali na 3 otázky.
 - Ak nie, pýtame sa: "Je to pig?"
 - Ak áno, odpoveď pig sme získali na 4 otázky.
 - Ak nie, vieme, že je to squirrel a odpoveď sme získali na 4 otázky.
 - Ak nie, pýtame sa "Je z množiny {mouse, owl}?" (pp. množiny je 1/8, pp. zvyšku tiež)
 - Bez ohľadu na odpoveď, mouse aj owl získame na 4 otázky

atď.

Odpoveď, ktorá má pravdepodobnosť $p = (1/2)^k$, nájdeme na k otázok, teda na $-\log_2(p)$ otázok. Stredná hodnota počtu otázok, ktoré potrebujeme, je súčet členov $-p \cdot \log_2(p)$.

4. Príklad 4: mená II., generovanie pomocou mince + kódovanie výsledkov

Majme znovu tú istú tabuľku:

Meno	Pravdepodobnosť
Adam	1/2
Boris	1/4
Cyril	1/8
Daniel	1/8

Chceme mená kódovať pomocou postupnosti núl a jednotiek tak, aby sme ich (v priemere pri veľkom počte opakovaní) potrebovali čo najmenej.

Určite to vieme dosiahnuť s dvoma bitmi – 4 možnosti kódujeme ako 00, 01, 10, 11.

Dá sa to aj lepšie:

- Adam: 0
- Boris: 10
- Cyril: 110
- Daniel: 111

Navyše majú tieto kódy tú vlastnosť, že nepotrebujú medzery, napríklad význam 01011110010 je jednoznačne daný.

Stredná hodnota počtu bitov, ktoré potrebujeme, je súčet členov $-p \cdot \log_2(p)$, čo je 1,75 (a teda menej ako 2)

Chceme generovať mená podľa tejto tabuľky tak, že pri tom využijeme hádzanie mincou. Chceli by sme ňou hádzať v priemere pri veľkom počte opakovaní čo najmenej krát.

- Hodíme prvý krát.
 - Ak padne znak, bude to znamenať meno Adam
 - Ak padne hlava, hádzeme druhý raz.
 - Ak padne znak, bude to znamenať meno Boris
 - Ak padne hlava, hádzeme tretí raz.
 - Ak padne znak, bude to znamenať meno Cyril.
 - Ak padne hlava, bude to meno Daniel.

Ak má meno pravdepodobnosť $p = (1/2)^k$, vygenerujeme ho na k hodov mincou, teda na $-\log_2(p)$ hodov. Stredná hodnota počtu hodov, ktoré potrebujeme, je súčet členov $-p \cdot \log_2(p)$. Pritom znovu napríklad ZHZHHZHHHZZHHZ je jednoznačne dané.

5. Definícia a výpočet entropie

Nech X je diskrétna náhodná premenná nadobúdajúca hodnoty $x_1, x_2, \dots$ s pravdepodobnosťami $p_1, p_2, \dots$. Potom **entropia** $H(X)$ je definovaná ako:

$$H(X) = - \sum_i p_i \log_2 p_i$$

Entropia $H(X)$ meria "neurčitost" náhodnej premennej.

Všimnime si, že nezávisí od konkrétnych hodnôt, ktoré sa nadobúdajú, ale iba od ich pravdepodobností. Nemeria teda rozptýlenosť hodnôt, ako to robí disperzia, ale množstvo informácie.

Napríklad majme náhodné premenné, ktoré nadobúdajú hodnoty $+n$ a $-n$ s rovnakou pravdepodobnosťou. Ich entropia je rovnaká, ale disperzia s rastúcim n rastie.

Na základe tejto informácie teda môžeme očakávať, že napríklad:

- Pri alternatívnom rozdelení (1 s pravdepodobnosťou p , inak 0) bude maximálne entropia pre $p=1/2$ – vtedy máme o výsledku najmenej informácie.
- Ak náhodná premenná nadobúda daný konečný počet hodnôt, entropia je maximálna, ak sa nadobúdajú s rovnakými pravdepodobnosťami.

Obidve vlastnosti platia, dokážeme prvú:

Majme alternatívne (Bernoulliho) rozdelenie:

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

Entropia je potom

$$H(X) = - \sum_{x=0}^1 P(X = x) \log_2 P(X = x) = -[p \log_2 p + (1 - p) \log_2 (1 - p)].$$

Derivácia podľa p :

$$\frac{dH}{dp} = -\log_2 p + \log_2 (1 - p) = \log_2 \left(\frac{1 - p}{p} \right) = \log_2 \left(\frac{1}{p} - 1 \right)$$

Derivácia je kladná pre $p \in (0, \frac{1}{2})$, nulová pre $p = \frac{1}{2}$, záporná pre $p \in (\frac{1}{2}, 1)$. Teda v bode $p = \frac{1}{2}$ má maximum (po tomto bod rastie, potom klesá),

Ako dôležitá charakteristika náhodných premenných sa uvádza napríklad aj pri pravdepodobnostných rozdeleniach na Wikipédii:

en.wikipedia.org/wiki/Geometric_distribution			
contrast, the following form of the geometric distribution is used for modeling the number of failures until the first success:		μ	σ
	Excess kurtosis	$6 + \frac{p^2}{1-p}$	$6 + \frac{p^2}{1-p}$
	Entropy	$\frac{-(1-p) \log(1-p) - p \log p}{p}$	$\frac{-(1-p) \log(1-p) - p \log p}{p}$
	MGF	$\frac{pe^t}{1 - (1-p)e^t}$	$\frac{p}{1 - (1-p)e^t}$