

Informacija in kodi – vaja 1

Entropija in informacijska analiza

Asistent: doc. dr. Klemen Grm

Entropija

- Mera nedoločenosti naključnih sistemov
 - Naključni dogodki
 - Naključne spremenljivke
 - Vezane naključne spremenljivke
- Večja entropija: bolj nedoločen sistem

Lastnosti entropije ene spremenljivke

- Matematična formulacija:

$$H(X) = -K \sum_{i=1}^n p(x_i) \log_d(p(x_i))$$

- Enote:

- K=1, d=2: biti/znak
- K=1, d=e: nati/znak

- Tudi:

$$H(X) = H(p(x_1), p(x_2), \dots, p(x_n))$$

Lastnosti entropije ene spremenljivke

- $H(X) \geq 0$
- Enakost velja, če za eno od stanj X velja $p(x_i) = 1$

$$H(0,1) = -(0 \log_2 0 + 1 \log_2 1) = 0$$

- $\left(\lim_{x \rightarrow 0} x \log x = 0 \right)$
- Intuicija: “naključni” sistem z verjetnostjo 1 ni nedoločen

Lastnosti entropije ene spremenljivke

- Entropija je neodvisna od permutacij stanj

$$H(X) = -K \sum_{i=1}^m p(x_i) \log_d(p(x_i))$$

$$= -K(p(x_1) \log_d p(x_1) + p(x_2) \log_d p(x_2) + \cdots + p(x_n) \log_d p(x_n))$$

$$= -K(p(x_2) \log_d p(x_2) + p(x_1) \log_d p(x_1) + \cdots + p(x_n) \log_d p(x_n))$$

Lastnosti entropije ene spremenljivke

- Entropijo maksimizira naključna spremenljivka z enakomerno verjetnostno porazdelitvijo stanj

$$H(0,1) = -(0 \log_2 0 + 1 \log_2 1) = 0$$

$$H(0.5, 0.5) = -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1$$

$$H(0.25, 0.75) = -(0.25 \log_2 0.25 + 0.75 \log_2 0.75) \approx 0.811$$

Lastnosti entropije ene spremenljivke

- Ob enakomerni porazdelitvi ima spremenljivka z več stanji večjo entropijo

$$H\left(\frac{1}{2}, \frac{1}{2}\right) = 1 \text{ bit/znak}$$

$$H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = 2 \text{ bita/znak}$$

[illegible]

Lastnosti entropije ene spremenljivke

- Stanje z verjetnostjo 0 ne vpliva na entropijo

$$H(0.5, 0.5, 0) = H(0.5, 0.5) = 1 \text{ bit/znak}$$

- V praksi: pazi pri računanju logaritma 0

Entropija odvisnih naključnih spremenljivk

- $(X, Y) \sim P_{XY}$
- $P_{XY} = (p(x_i, y_j) \geq 0 : i = 1, 2, \dots, m, j = 1, 2, \dots, n)$
- Če sta X in Y neodvisni:
- $P(XY) = P(X) \times P(Y)$;
- $H(XY) = H(X) + H(Y)$

Entropija odvisnih naključnih spremenljivk

- Vezana entropija:

$$H(P_{XY}) = H(X, Y) = - \sum_{i=1}^m \sum_{j=1}^n p(x_i, y_j) \log_2 p(x_i, y_j)$$

- Entropija X pri znanem Y :

$$H(P_{X|Y=y_j}) = H(X|Y = y_j) = - \sum_{i=1}^m p(x_i|y_j) \log_2 p(x_i|y_j)$$

- Pogojna entropija X glede na Y :

$$H(X|Y) = \sum_{j=1}^n p(y_j) H(X|Y = y_j) = - \sum_{i=1}^m \sum_{j=1}^n p(x_i, y_j) \log_2 p(x_i|y_j)$$

Zgled

Znake informacijskega vira $V \sim \begin{pmatrix} v_1 & v_2 & v_3 & v_4 & v_5 \\ 0.3 & 0.25 & 0.2 & 0.15 & 0.1 \end{pmatrix}$

Kodiramo dvojiško po predpisu

$$\begin{aligned} v_1 &\rightarrow 10, \\ v_2 &\rightarrow 11, \\ v_3 &\rightarrow 00, \\ v_4 &\rightarrow 010, \\ v_5 &\rightarrow 011. \end{aligned}$$

Določi osnovno entropijo vira ter residualno entropijo potem, ko prejmemo en dvojiški znak oz. dva dvojiška znaka.

Rešitev

- Preden sprejmemo 1. binarni znak:
- $H(V) = H(0.3, 0.25, 0.2, 0.15, 0.1)$
- $= -(0.3 \log_2 0.3 + \dots + 0.1 \log_2 0.1) \approx 2.23 \text{ bita/znak}$
- Intuicija: več znakov, kot sprejmemo, nižja je nedoločenost sistema
- Analitična rešitev: entropija razbitja,
- $H(p_1, \dots, p_m, r_1, \dots, r_n) = H(p, r) + pH\left(\frac{p_1}{p}, \dots, \frac{p_m}{p}\right) + rH\left(\frac{r_1}{r}, \dots, \frac{r_n}{r}\right)$

Rešitev

$$V \sim \begin{pmatrix} v_1 & v_2 & v_3 & v_4 & v_5 \\ 0.3 & 0.25 & 0.2 & 0.15 & 0.1 \end{pmatrix}$$

$$\begin{aligned} v_1 &\rightarrow 10, \\ v_2 &\rightarrow 11, \\ v_3 &\rightarrow 00, \\ v_4 &\rightarrow 010, \\ v_5 &\rightarrow 011. \end{aligned}$$

- Ko sprejmemo prvi znak:
- $0 \Rightarrow (r_1, r_2, r_3) = (v_3, v_4, v_5); r = P(v_3 \vee v_4 \vee v_5) = 0.45$
- $1 \Rightarrow (p_1, p_2) = (v_1, v_2); p = P(v_1 \vee v_2) = 0.55$
- $H(V) = H(0.45, 0.55) + 0.55H\left(\frac{0.3}{0.55}, \frac{0.25}{0.55}\right) + 0.45H\left(\frac{0.2}{0.45}, \frac{0.15}{0.45}, \frac{0.1}{0.45}\right)$
- $H(V) \approx 0.993 + 0.55 \times 0.994 + 0.45 \times 1.54 \approx 2.23$
- $H_{1z} = H(V) - H(p, r) \approx 1.24 \text{ bita/znak}$

Rešitev

$$V \sim \begin{pmatrix} v_1 & v_2 & v_3 & v_4 & v_5 \\ 0.3 & 0.25 & 0.2 & 0.15 & 0.1 \end{pmatrix}$$

$$\begin{aligned} v_1 &\rightarrow 10, \\ v_2 &\rightarrow 11, \\ v_3 &\rightarrow 00, \\ v_4 &\rightarrow 010, \\ v_5 &\rightarrow 011. \end{aligned}$$

- Ko sprejmemo drugi znak:
- v_1, v_2, v_3 določeni
- Preostane nedoločenost med v_4, v_5
- $H_{2Z} = H(V) - H(0.3, 0.25, 0.2, 0.25) \approx 2.23 - 1.99 \approx 0.24 \text{ bita/znak}$
- Domača naloga: izrazi H_{2Z} z entropijo razbitja

Entropija nizov znakov

- Ocena na podlagi statistike podatkov

Zgled: niz znakov $(z_1, z_2, z_3, z_4, \dots, z_n)$

- Možne znake modeliramo kot naključno spremenljivko X , določimo zalogo vrednosti $Z(X)$.
- $H_1(X)$: ocenimo verjetnosti posameznih stanj iz $Z(X)$
- $H_2(X)$: znake obravnavamo paroma, gledamo vsa možna zaporedja
 - $(z_1, z_2), (z_2, z_3), (z_3, z_4), \dots$
- $H_3(X)$: zaporedja dolžine 3
 - $(z_1, z_2, z_3), (z_2, z_3, z_4), (z_3, z_4, z_5), \dots$




Zgled

- Niz *aaaabbbaaab*
- H_1 :
 - $p(X = a) = 0.7; p(X = b) = 0.3$
 - $H_1(X) = -(0.7 \log_2 0.7 + 0.3 \log_2 0.3) \approx 0.88 \text{ bit/znak}$
- H_2 :
 - Niz spremenimo v zaporedje parov:
 - [aa, aa, aa, ab, bb, ba, aa, aa, ab]
 - $p(X = aa) = \frac{5}{9}; p(X = ab) = \frac{2}{9}; p(X = ba) = \frac{1}{9}, p(X = bb) = \frac{1}{9}$
 - $H_2(X) = H\left(\frac{5}{9}, \frac{2}{9}, \frac{1}{9}, \frac{1}{9}\right) \approx 1.658 \text{ bit/par} = 0.83 \text{ bit/znak}$
- H_3 : Domača naloga

Smiselnost entropij višjih redov

- Binarne datoteke: 1 bajt = 8 bitov
- Nabor možnih bajtov: $2^8 = 256$
- Statistično signifikantna ocena: $dolžina > 10 \times (nabor\ znakov)$
- H_1 : nabor $2^8 = 256$, min. dolžina za smiselno oceno $\approx 2.5\ kB$
- H_2 : nabor $2^{16} = 65536$, min. dolžina za smiselno oceno $\approx 655\ kB$
- H_3 : nabor $2^{24} \approx 16.8 \times 10^6$, min. dolžina $\approx 168\ MB$
- $H_4, H_5 \dots$ domača naloga: izračunaj potrebno dolžino

Vaja 1

Name	Date modified	Type	Size
 besedilo.txt	3/2/2022 10:59 AM	Text Document	1,015 KB
 posnetek.wav	3/2/2022 10:36 AM	WAV Audio File (VLC)	41,048 KB
 slika.bmp	3/19/2020 6:52 PM	BMP File	1,797 KB

- Obdelava datotek
 - posnetek, slika: izgubna oz. brezizgubna kompresija
 - besedilo: kompresija, kompresija + šifriranje
- Določitev entropije 1. in višjih redov
- Določitev in karakterizacija smiselnih rezultatov

Vaja 1

bmp






png



jpeg



 slika.bmp
 slika.jpg
 slika.png

19. 03. 2020 18:52

Datoteka BMP

1.797 KB

19. 03. 2020 18:53

Datoteka JPG

36 KB

19. 03. 2020 18:51

Datoteka PNG

815 KB