RELATIVNA ENTROPIJA KOT MERA PRESENEČENJA

ŽIGA VIRK

Fakulteta za računalništvo in informatiko
Univerza v Ljubljani

Math. Subj. Class. (2010): 94A17, 94A24

Skozi interpretacijo teorije informacij razložimo, kako merimo količino informacij, povprečno količino informacij (entropijo) ter presenečenja, ki izhajajo iz dolžin sporočil. Mera presenečenja je poznana pod različnimi imeni: relativna entropija, Kullback-Leiblerjeva divergenca ...

RELATIVE ENTROPY AS A MEASURE OF SURPRISE

Using the interpretation developed by the coding theory we explain how we measure the amount of information, average content of information (entropy) and surprisal arising from the lengths of codes. The measure of surprisal is known under various names: relative entropy, Kullback-Leibler divergence ...

Uvod

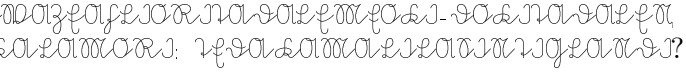
V današnjih časih smo nenehno v stiku z veliko količino podatkov. Nekateri podatki nam podajo veliko informacij, drugi malce manj. V tem članku bomo predstavili osnove teorije informacij (informatika in teorija informacij sta različni področji), skozi katere bomo lahko merili količino informacij, povprečno količino informacij (entropijo) in presenečenje, izhajajoče iz kodiranja. Kot motivacijo si oglejmo naslednji dialog.

OSEBA A:  ?

PREVOD 1: Kako si se imel za vikend?

OSEBA B:  .

PREVOD 2: Dobro.

OSEBA A:  ?

PREVOD 3: Je deževalo?

OSEBA B:  .

PREVOD 4: Malce je deževalo dopoldne. Ravno dovolj, da mi je opralo vesoljsko plovilo. Popoldne ob kosilu pa je že sijalo sonce.

Čeprav osebi govorita v nam neznanem jeziku, so nekateri prevodi (predpostavimo, da so prevodi ustrezni) bolj presenetljivi kot drugi. Izstopata prevoda 3 in 4. Prevoda 1 in 2 sta primerljive dolžine z originalnim zapisom. Prevod 3 je presenetljiv, saj je precej krajši od originalnega zapisa. Kot tak nakazuje na to, da je podajanje informacij v originalnem zapisu precej neučinkovito. V nasprotnem smislu je presenetljiv prevod 4: kratek originalen zapis s štirimi simboli vsebuje tri povedi. V tem primeru bi podvomili o kvaliteti prevoda ali pa ugibali, da je morda standarden način pozdrava (matematično gledano bi to pomenilo, da se v neznanem jeziku pojavlja z visoko verjetnostjo). Karkoli je že razlog, presenečenji, ki izhajata iz prevodov 3 in 4, sta povezani s pričakovano dolžino prevoda in s tem povezano količino informacij. Namen tega članka je, da omenjene opazke formuliramo v matematični obliki. Pri tem bomo namesto prevajanja obravnavali kodiranje. Obširna moderna obravnava omenjene matematične osnove in interpretacija v kontekstih teorije informacij ter biološke raznolikosti je podana v [6].

Začetki merjenja informacij segajo v štirideseta leta dvajsetega stoletja, ko je Shannon [9] definiral **količino informacije** v kontekstu verjetnosti. Ob tem bi opozorili, da so zaradi interdisciplinarne in široko aplikativne narave teorije koncepti večkrat dobili različna imena. Količina informacij je poznana pod naslednjimi izrazi: information content, self-information, surprisal (v našem kontekstu bo presenečenje nekaj drugega) ter Shannon information. Na osnovi te količine je Shannon definiral **entropijo** slučajne spremenljivke kot povprečno količino informacij. Koncept entropije je bil takrat seveda že poznan, zato se entropiji v našem kontekstu včasih reče informacijska entropija ali Shannonova entropija. Entropija je na enak ali podoben način definirana v termodinamiki in kvantni fiziki. Kratek pregled literature pokaže še številne druge kontekste, v katerih se entropija pojavlja v taki obliki.

Na osnovi Shannonovega dela in razumevanja entropije v kontekstu informacij je Huffman kot študent razvil algoritem za optimalno kodiranje jezikov [3]. Njegov pristop je poznan pod imenom **Huffmanovo kodiranje** in je osnova za kompresijske metode brez izgube informacij. Med drugim se izboljšave Huffmanovega kodiranja uporabljajo pri računalniških formatih .JPEG [8] in .MP3 [4] datotek.

Zadnji koncept, ki ga bomo predstavili v članku, je **relativna entropija**. Le-ta meri presenečenje v smislu zgornjega primera. Formalno sta jo definirala Kullback in Leibler [5], matematično pa predstavlja količino različnosti dveh slučajnih spremenjivk na n točkah. Natančno definicijo bomo podali v zadnjem poglavju. Na tem mestu le omenimo, da gre za nesimetrično količino, zato se je ne omenja z izrazom metrika. V literaturi se omenja pod izrazoma razdalja (distance) ali divergenca (divergence). Kljub nesimetričnosti se je relativna entropija izkazala za pomembno količino na različnih po-

dročjih. V tuji literaturi je poznana pod različnimi imeni: Kullback-Leibler information, Kullback-Leibler distance, Kullback-Leibler divergence, directed divergence, information divergence, information deficiency, amount of information, discrimination information, relative information, gain ali information ali information gain, discrimination distance in error. Na koncu bomo omenili še pomen in uporabo relativne entropije v zadnjem času.

Količina informacij

Količino podatkov merimo z biti. Spominska celica v klasičnem (ne-kvantnem) računalniku praviloma zavzame vrednost 0 ali 1. Količina podatkov shranjena v eni taki celici predstavlja 1 bit podatkov. Z enim bitom lahko opišemo dve različni stanji, z n biti pa 2^n stanj. Količina podatkov pove, koliko različnih stanj lahko s podatki opišemo. Bite tradicionalno združujemo v byte, ki so osnova za večje količine podatkov: MB, GB ...

Količina podatkov na splošno ni enaka količini informacij. Medtem ko količina podatkov meri njihovo razsežnost v spominu, je informacija odvisna od konteksta. Izjavo »V Ljubljani ob polnoči ne sije sonce« lahko kot podatke zapišemo z nekaj sto biti v standardnih kodiranjih, težko pa bi rekli, da vsebuje kakšno informacijo. Za opredelitev količine informacij potrebujemo kontekst. V matematičnem jeziku bo kontekst diskretna slučajna spremenljivka X z izidi $x_1, x_2, \ldots, x_m$ in pripadajočimi verjetnostmi $p_1, p_2, \ldots, p_m$. Količino informacij lahko definiramo podobno kot količino podatkov. Če n bitov podatkov opiše 2^n različnih stanj, n bitov informacije opiše 2^n različnih enako verjetnih dogodkov $x_1, x_2, \ldots, x_{2^n}$ (katerih verjetnost je torej 2^{-n}).

Definicija 1. Količina informacij, vsebovana v dogodku verjetnosti p, je enaka $\log_2(1/p) = -\log_2 p$ bitov. Enota je bit oz. shannon.

Osnova 2 v logaritmu izhaja iz dejstva, da en bit podatkov opiše dve različni stanji. Občasno se za osnovo uporablja kakšno drugo število $a > 1$. Pri tem se zaradi lastnosti logaritmov količina informacij spremeni za faktor, odvisen le od a. Pri osnovi e se enoti informacije včasih reče nat, pri osnovi 10 pa digit oz. hartley.

Primeri količin informacij v dogodkih:

1. Izid meta poštenega kovanca: 1 bit.
2. Ob 23:00 v Ljubljani ne bo sijalo sonce: 0 bitov, če upoštevamo, da se izid zgodi z verjetnostjo 1.
3. V izidu *ura je 14:23* (brez podatka o sekundah) je manj informacij kot v izidu *ura je 14:23:15*.

4. Izid meta poštene standardne kocke: $\log_2 6 = 1 + \log_2 3$ bitov.
5. Sodost-lihost izida meta poštene standardne kocke: 1 bit.
6. Ostanek izida meta poštene standardne kocke pri deljenju s 3: $\log_2 3$ bitov.

Zadnji trije primeri nakazujejo, da naša definicija zadošča pričakovani lastnosti: če sta dogodka A in B neodvisna (torej je $P(A \cap B) = P(A)P(B)$), potem je količina informacije podana z dogodkom $A \cap B$ enaka vsoti količin informacij posameznih delov. Sledeči izrek pove, da ta lastnost do osnove a natanko določi funkcijo količine informacij (in nedvoumno utemelji definicijo 1).

Izrek 2. *Naj bo $f\colon [0,1] \to [0,\infty)$ funkcija, ki zadošča naslednjim pogojem:*

1. *$f(1) = 0$, oz. gotovi dogodki ne podajo nič informacij.*
2. *f je strogo padajoča v p, oz. redkejši dogodki podajo več informacij.*
3. *$f(p \cdot q) = f(p) + f(q)$.*

Tedaj je

$$f(p) = -\log_a p$$

za neki $a > 1$.

Dokaz. Naj bo $x = f(1/2) > 0$. Z induktivno uporabo predpostavke 3 dobimo enakost

$$f((1/2)^m) = x \cdot m, \quad \text{za vse } m \in \mathbb{N}. \tag{1}$$

Za naravno število k velja

$$x = f(1/2) = f\left(\left((1/2)^{1/k}\right)^k\right) = kf\left((1/2)^{1/k}\right)$$

in torej $f\left((1/2)^{1/k}\right) = x/k$. Če v tem primeru ponovno induktivno uporabimo predpostavko 3, dobimo

$$f\left((1/2)^{m/k}\right) = x \cdot m/k, \quad \text{za vse } m, k \in \mathbb{N}. \tag{2}$$

Izberimo $t \in [0,\infty)\setminus\mathbb{Q}$ ter konvergentni zaporedji racionalnih števil iz $[0,\infty)$ z limito t: zaporedje s_i naj monotono narašča proti t, zaporedje z_i pa naj monotono pada proti t. Po predpostavki 2 dobimo

$$x \cdot t = \lim_{i\to\infty} f\left((1/2)^{z_i}\right) \le f\left((1/2)^t\right) \le \lim_{i\to\infty} f\left((1/2)^{s_i}\right) = x \cdot t.$$

Torej velja

$$f\left((1/2)^t\right) = x \cdot t, \quad \text{za vse } t \in [0,1]. \tag{3}$$

Iz predpostavk 1 in 2 sledi, da je $x > 0$. Tedaj je $a = 2^{1/x} > 1$, velja $x = 1/\log_2 a$ in po enačbi (3) sledi

$$f(p) = f\left((1/2)^{\log_{1/2} p}\right) = x \cdot \log_{1/2} p = \frac{1}{\log_2 a} \cdot (-\log_2 p) = -\log_a p,$$

za vsak $p \in [0,1]$. ∎

Entropija kot povprečna količina informacij

Entropija diskretne slučajne spremenjivke je povprečna količina informacij vsebovana v izidih.

Definicija 3. Naj bo X diskretna slučajna spremenljivka z izidi x_1, x_2, …, x_m in pripadajočimi verjetnostmi $p_1, p_2, \ldots, p_m$. **Entropija** X je enaka

$$H(X) = \sum_{i=1}^{m} p_i \log_2(1/p_i).$$

Izraz $0 \cdot \log_2(1/0)$ matematično ni definiran. V našem kontekstu bomo kljub temu uporabljali **dogovor** $0 \cdot \log_2(1/0) = 0$. Prvi razlog je dejstvo, da lahko spremenljivki X vedno umetno dodamo nov izid x_{m+1} z verjetnostjo 0. S tem spremenljivke praktično ne spremenimo, čeprav smo formalno spremenili (razširili) njen opis. Želimo, da omenjena sprememba ne vpliva na entropijo, tj., da dodaten člen $0 \cdot \log_2(1/0)$ v vsoti ne spremeni entropije. Drugi razlog je, da izraz $0 \cdot \log_2(1/0)$ v entropiji dejansko izhaja iz $p \cdot \log_2(1/p)$ pri $p = 0$ in

$$\lim_{p \searrow 0} p \cdot \log_2(1/p) = 0.$$

Namesto omenjenega dogovora bi lahko torej vsak izraz $p_i \log_2(1/p_i)$ v definiciji 3 zamenjali z

$$\lim_{p \searrow p_i} p \cdot \log_2(1/p).$$

Opazimo, da $H(X)$ lahko zavzame katerokoli vrednost na intervalu $[0, \log_2 m]$:

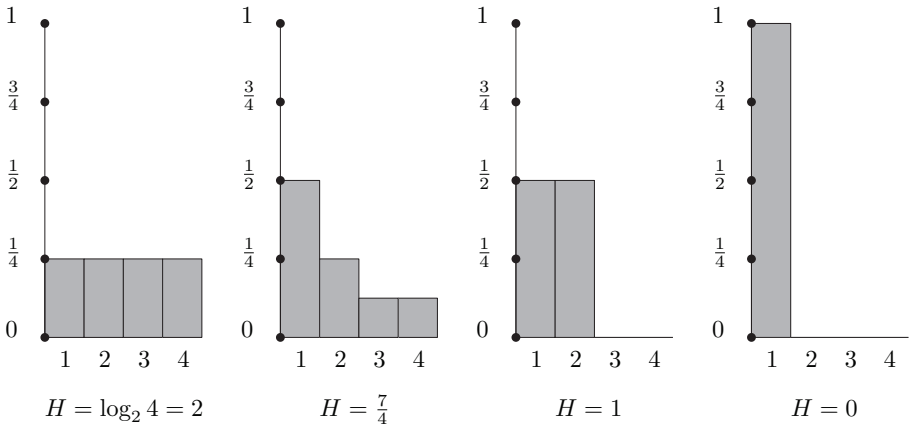
- Minimum: $H(X) = 0$ natanko tedaj, ko obstaja gotov izid x_i, tj., $p_i = 1$.
- Maksimum: $H(X) = \log_2 m$ natanko tedaj, ko X predstavlja enakomerno porazdelitev.

Večanje entropije pri fiksnem m torej predstavlja premikanje proti enakomerni porazdelitvi na m točkah.

Primeri entropije:

- Entropija meta poštenega kovanca je enaka $\log_2 2 = 1$.
- Entropija meta poštene kocke je enaka $\log_2 6$.
- Če kovanec ni pošten, je entropija manjša kot 1.

Več primerov je podanih na sliki 1.



Slika 1. Entropija nekaterih porazdelitev na štirih točkah.

Učinkovita kodiranja

Pogosto želimo, da je naše komuniciranje učinkovito. Informacije želimo izraziti oz. prenesti na čim bolj ekonomičen način. Med drugimi v ta namen lahko uporabljamo učinkovite tipkovnice (npr. tipa Dvorak), okrajšave (npr.), kratice (DMFA), bližnjice (Ctrl-C) ipd. Poleg tega se je matematični zapis (npr. računov) skozi zgodovino razvil v tako obliko, ki učinkovito poda veliko količino informacij. Zapis enačbe »Osnova naravnega logaritma na potenco korena prvega negativnega celega števil pomnoženega s kvocientom obsega in premera kroga je enaka prvemu negativnemu celemu številu« je rahlo manj učinkovit kot $e^{i\pi} = -1$. Učinkovitost komuniciranja je lepo povzel Pitagora: *»Ne izrecite malo z veliko besedami, raje povejte veliko v le nekaj besedah.«*

V tem poglavju si bomo ogledali učinkovita kodiranja v smislu učinkovitega zapisa informacij. Kot običajno naj bo naš kontekst diskretna slučajna spremenljivka X z izidi $x_1, x_2, \ldots, x_m$ in pripadajočimi verjetnostmi

$p_1, p_2, \ldots, p_m$. V okviru kodiranja izidi $x_1, x_2, \ldots, x_m$ predstavljajo abecedo (seznam črk oz. simbolov x_i), verjetnosti p_i pa so relativne frekvence, s katerimi se črke x_i pojavljajo v danem jeziku ali besedilu, sestavljenem iz zaporedja črk. Na primer, besedilo lahko predstavlja del zapisa DNK v obliki zaporedja iz črk A, C, G in T. V primeru besedila v slovenščini bi bila (kodirna) abeceda sestavljena iz velikih in malih črk, presledka in ločil (ter potencialno drugih uporabljenih znakov).

V našem kontekstu torej na podlagi podanega besedila generiramo slučajno spremenljivko X, ki predstavlja relativne frekvence črk. Kodiranje besedila pomeni, da vsaki črki priredimo dvojiško kodo (končno zaporedje ničel in enic), s katero bomo dotično črko predstavili v računalniškemu pomnilniku.

Definicija 4. Kodiranje slučajne spremenljivke X je injektivna preslikava, ki vsakemu izidu x_i priredi kodo (včasih omenjeno kot kodno besedo) y_i v obliki končnega dvojiškega zaporedja (tj. zaporedja, sestavljenega iz števil 0 in 1).

Kode morajo biti seveda take, da lahko iz njih enolično rekonstruiramo originalno besedilo. Znano je kodiranje ASCII, ki vsak znak običajno zakodira s sedmimi oz. osmimi biti. Rekonstrukcija besedila je v tem primeru enostavna, saj vsakih osem bitov predstavlja en simbol abecede. Za zapis besedila dolžine n bomo torej porabili $8n$ bitov. Bistveno vprašanje je, ali lahko to dolžino zmanjšamo brez izgube informacij. Izkaže se, da je odgovor pogosto da. Začnimo razlago s primerom.

črka	frekvenca	kodiranje 1	kodiranje 2	neustrezno kodiranje
a	0,5	0 0	0	0
b	0,25	0 1	1 0	0 1
c	0,25	1 0	1 1	1 0
d	0	1 1		11
povprečna dolžina kode		2	3/2	-

Tabela 1. Primer kodiranja.

Tabela 1 podaja abecedo (a, b, c, d) s pripadajočimi relativnimi frekvencami črk/vrednostmi. Ker so črke štiri, lahko vse enolično zapišemo z dvema bitoma, kar porodi kodiranje 1. V tem primeru je povprečna dolžina kode črke enaka 2. Kodiranje 2 predstavlja alternativno kodiranje, pri katerem je povprečna dolžina kode črke enaka

$$0{,}5 \cdot 1 + 0{,}25 \cdot 2 + 0{,}25 \cdot 2 = 3/2.$$

Z uporabo tega kodiranja bodo torej kodiranja v podani abecedi precej krajša brez izgube informacij. Ideja pri kodiranju 2 je, da bolj pogostim oz. verjetnim črkam priredimo krajšo kodo. Pri tem morajo biti črkam prirejene take kode, da lahko iz kodiranja enolično rekonstruiramo besedilo. To najlažje dosežemo, če so začetki kod različni:

Definicija 5. Kodiranje slučajne spremenljivke X s kodami $y_1, y_2, \ldots, y_m$ je **predponsko** (instantaneous oz. prefix-free), če se nobena koda y_i ne pojavi kot začetno zaporedje kakšne druge kode.

Primer kodiranja, ki ni predponsko, je podan v tabeli 1 pod stolpcem neustrezno kodiranje: koda 0 črke *a* je začetni odsek kode 0 1 črke *b*. Zapis 010 lahko predstavlja besedo *ac* ali *ba*. Rekonstrukcija v tem primeru ni enolična, s kodiranjem smo torej izgubili nekaj informacije.

Tabeli 2 in 3 podata še nekaj podobnih primerov kodiranja.

črka	p	kodiranje 1	kodiranje 2
a	0,5	0 0	0
b	0,25	0 1	1 0
c	0,125	1 0	1 1 0
d	0,125	1 1	1 1 1
povprečna dolžina kode		2	7/8

Tabela 2. Drugi primer kodiranja.

črka	p	kodiranje 1	kodiranje 2
a	1/3	0 0	0
b	1/3	0 1	1 0
c	1/3	1 0	1 1
povprečna dolžina kode		2	5/3

Tabela 3. Tretji primer kodiranja.

Pri tem se zastavi vprašanje: do kolikšne mere lahko skrajšamo povprečno dolžino kode? Izkaže se, da je spodnja meja podana z entropijo X (glej alinejo 2 v izreku 6).

Izrek 6. *Naj bo X diskretna slučajna spremenljivka (abeceda) z izidi (črkami) $x_1, x_2, \ldots, x_m$ in pripadajočimi verjetnostmi (relativnimi frekvencami) $p_1, p_2, \ldots, p_m$. Podano naj bo predponsko kodiranje, ki črki x_i priredi kodo y_i dolžine L_i. Tedaj velja:*

1. $\sum_{i=1}^{m}(1/2)^{L_i} \leq 1$.
2. *Povprečna dolžina kode je večja ali enaka entropiji:* $\sum_{i=1}^{m} p_i L_i \geq H(X)$.

Dokaz. 1. Vsakemu številu t na intervalu $[0,1)$ lahko priredimo dvojiški zapis (upoštevajoč le decimalke):

$$t = 0{,}b_1b_2b_3\ \ldots \qquad \text{oziroma} \qquad t = \sum_{i=1}^{\infty} b_i \cdot 2^{-i}.$$

Pri tem se za števila z dvema zapisoma (npr. $0{,}011111\ldots = 0{,}1$) omejimo le na tisti zapis, ki se ne konča z neskončnim zaporedjem enic. Vsaki kodi y_i priredimo interval

$$J_i = \{t \in [0,1) \mid \text{dvojiški zapis } t \text{ se začne z } y_i\}.$$

Na primer:

- Če je $y_1 = 1$, potem je $J_1 = [1/2, 1)$, saj so dvojiški zapisi vseh števil na J_1 oblike $0{,}1^*$ (zapis $0{,}1^*$ pomeni, da je prva decimalka 1, druge decimalke pa so poljubne).
- Če je $y_2 = 011$, potem je $J_2 = [3/8, 4/8)$, saj so dvojiški zapisi vseh števil na J_2 oblike $0{,}011^*$ (zapis $0{,}011^*$ pomeni, da so prve tri decimalke $0, 1$ in 1, ostale decimalke pa so poljubne).

Intervali J_i so disjunktni zaradi predponskosti kodiranja in vsebovani v $[0,1)$. Po definiciji so njihove dolžine $(1/2)^{L_i}$. Skupna dolžina ne more presegati dolžine intervala $[0,1)$, kar pomeni $\sum_{i=1}^{m}(1/2)^{L_i} \leq 1$.

2. Pri dokazu alineje 2 bomo uporabili konkavnost funkcije $f(x) = \log_2 x$ (za neenakost med vrsticama 2 in 3) ter alinejo 1 (na zadnjem koraku).

$$\begin{aligned}
H(X) - \sum_{i=1}^{m} p_i L_i &= \sum_{i=1}^{m} p_i \log_2(1/p_i) + \sum_{i=1}^{m} p_i \log_2\left((1/2)^{L_i}\right) \\
&= \sum_{i=1}^{m} p_i \log_2 \frac{(1/2)^{L_i}}{p_i} \\
&\leq \log_2 \left(\sum_{i=1}^{m} p_i \cdot \frac{(1/2)^{L_i}}{p_i} \right) \\
&= \log_2 \left(\sum_{i=1}^{m} (1/2)^{L_i} \right) \\
&\leq \log_2 1 = 0
\end{aligned}$$

Sledi $H(X) \leq \sum_{i=1}^{m} p_i L_i$. ∎

Opomba 7. Alineja 1 izreka 6 se imenuje Kraft–McMillanova neenakost, glej str. 46 v [6].

Kodiranje slučajne spremenljivke X, pri katerem za dolžine kod L_i izidov x_i velja $L_i = \log_2(1/p_i)$, se imenuje **idealno kodiranje**. V praksi takšno kodiranje obstaja natanko tedaj, ko so p_i potence števila $1/2$. Po definiciji entropije za idealna kodiranja velja, da je njihova povprečna dolžina kode enaka entropiji. Kodiranji številka 2 v prvih dveh tabelah sta idealni kodiranji, kodiranje v tretji tabeli pa ne. Naslednji izrek pove, kako lahko konstruiramo kodiranje slučajne spremenljivke X, pri katerem se povprečna dolžina kode spodnji meji približa do enega bita natančno.

Izrek 8. *Naj bo X diskretna slučajna spremenljivka (abeceda) z izidi (črkami) $x_1, x_2, \ldots, x_m$ in pripadajočimi verjetnostmi (relativnimi frekvencami) $p_1, p_2, \ldots, p_m$. Tedaj obstaja predponsko kodiranje, ki črki x_i priredi kodo y_i dolžine $L_i = \lceil \log_2(1/p_i) \rceil$. Poleg tega velja $\sum_{i=1}^{m} p_i L_i < H(X) + 1$.*

Dokaz. Brez škode za splošnost privzemimo, da je $p_1 \geq p_2 \geq \ldots \geq p_m$. Za vsak i je število $q_i = (1/2)^{L_i}$ najmanjša celoštevilska potenca števila $1/2$, ki je manjša od p_i. Števila $q_1, q_2, \ldots, q_i$ so celoštevilski večkratniki $(1/2)^{L_i}$, zato enako velja za njihove poljubne vsote. Črki x_i priredimo interval

$$J_i = [q_1 + q_2 + \cdots + q_{i-1}, q_1 + q_2 + \cdots + q_{i-1} + q_i).$$

Upoštevajmo dogovor iz prejšnjega dokaza in naj y_i predstavlja prvih L_i števk v dvojiškem zapisu števila $q_1 + q_2 + \cdots + q_{i-1}$ (druge števke so enake 0, saj je omenjeno število celoštevilski večkratnik $(1/2)^{L_i}$). Interval J_i sovpada s števili iz $[0, 1)$, katerih prvih L_i števk v dvojiškem zapisu se ujema z y_i. Ker so intervali J_i disjunktni, je dobljeno kodiranje predponsko.

Iz definicije L_i sledi, da je $L_i < \log_2(1/p_i) + 1$. Tedaj je

$$\sum_{i=1}^{m} p_i L_i < \sum_{i=1}^{m} p_i \left(\log_2(1/p_i) + 1\right) = H(X) + 1. \qquad \blacksquare$$

Primer 9. Oglejmo si demonstracijo zadnjega dokaza na primeru, podanem v tabeli 2. Verjetnosti p_i porodijo vrednosti $L_i = \log_2(1/p_i)$, saj so vse vrednosti p_i celoštevilske potence $1/2$. Pripadajoči intervali so

$$J_1 = [0, 1/2),\ J_2 = [1/2, 3/4),\ J_3 = [3/4, 7/8),\ J_4 = [7/8, 1).$$

- Koda y_1 sestoji iz ene ($L_1 = 1$) števke dvojiškega zapisa števila 0, tj., $y_1 = 0$.

- Koda y_2 sestoji iz dveh ($L_2 = 2$) števk dvojiškega zapisa števila $1/2$, tj., $y_2 = 10$.
- Koda y_3 sestoji iz treh ($L_3 = 3$) števk dvojiškega zapisa števila $3/4$, tj., $y_1 = 110$.
- Koda y_4 sestoji iz treh ($L_4 = 3$) števk dvojiškega zapisa števila $7/8$, tj., $y_4 = 111$.

Dobili smo kodiranje 2 iz tabele 2.

Predstavljeni rezultati se nanašajo na kodiranja, pri katerih vsaki črki priredimo svojo kodo. V tem primeru smo videli, da lahko povprečno dolžino kode črke zmanjšamo pod $H(X) + 1$, a ne pod $H(X)$. Izkaže se, da bi v primeru, ko bi namesto črk kodirali nize črk, povprečno dolžino kode črke lahko poljubno približali $H(X)$. Sorodna ideja se pojavlja pri pisavah, ki z znaki zapisujejo zloge namesto črk.

Relativna entropija kot mera presenečenja

V tem poglavju naj bo X diskretna slučajna spremenljivka (abeceda) z izidi (črkami) $x_1, x_2, \ldots, x_m$ in pripadajočimi verjetnostmi (relativnimi frekvencami) $p_1, p_2, \ldots, p_m$. Mislimo si lahko, da izhaja iz frekvenc črk v nekem daljšem besedilu. Prav tako naj bo Y diskretna slučajna spremenljivka (abeceda) z izidi (črkami) $x_1, x_2, \ldots, x_m$ in pripadajočimi verjetnostmi (relativnimi frekvencami) $q_1, q_2, \ldots, q_m$. Pri tem si mislimo, da gre za neko drugo besedilo z istimi črkami.

Povprečna dolžina kod črk v idealnem kodiranju X (ki je sicer dosegljiva le v primeru, ko so p_i potence $1/2$) je

$$\sum_{i=1}^{m} p_i \log_2(1/p_i) = H(X).$$

Če bi slučajno spremenljivko X zakodirali v idealnem kodiranju za slučajno spremenljivko Y, bi bila povprečna dolžina črk enaka

$$\sum_{i=1}^{m} p_i \log_2(1/q_i).$$

Gibbsova neenakost (gre za »zvezno« verzijo neenakosti iz alineje 2 izreka 6, dokaz je podan v okviru naloge 2.26 na strani 37 v [7]) pravi, da je ta

količina večja ali enaka entropiji:

$$\sum_{i=1}^{m} p_i \log_2(1/q_i) \geq H(X).$$

Relativna entropija je razlika med tema dvema količinama in predstavlja presežno povprečno dolžino kod črk, ki pri kodiranju besedila X nastane zaradi uporabe kodiranja, optimiziranega za neko drugo besedilo.

Definicija 10. Relativna entropija med diskretnima slučajnima spremenljivkama X in Y je

$$D(X \parallel Y) = \sum_{i=1}^{m} p_i \log_2(1/q_i) - \sum_{i=1}^{m} p_i \log_2(1/p_i) = \sum_{i=1}^{m} p_i \log_2(p_i/q_i).$$

V uvodu smo podali primera presenetljivih in nepresenetljivih prevodov, pri čemer je količina presenečenja izhajala iz nepričakovane dolžine prevodov. Relativna entropija v kontekstu kodiranja formalizira to idejo presenečenja. Seveda je prevajanje besedil bolj kompleksno kot kodiranje. Po drugi strani pa smo definirali relativno entropijo slučajne spremenljivke. S tem smo podali pojem, ki ni vezan le na kodiranje.

Omenimo nekaj lastnosti relativne entropije:

- $D(X \parallel Y)$ običajno ni enako $D(Y \parallel X)$.
- $D(X \parallel Y) \in [0, \infty]$.
- $D(X \parallel Y) = 0$ natanko tedaj, ko je $X = Y$.
- $D(X \parallel Y) = \infty$ natanko tedaj, ko obstaja i, da velja $0 = q_i < p_i$. V tem primeru idealno kodiranje za Y ne vsebuje kode za x_i, zato z njim ne moremo zakodirati besedila, ki to črko vsebuje.

Pomen relativne entropije

Relativna entropija je imela osrednji pomen pri razvoju teorije kodiranja, njena številna imena pa nakazujejo, da se uporablja na veliko področjih. Ker sama po sebi ni metrika, se je pojavila potreba, da bi na njeni podlagi definirali čim bolj sorodne količine, ki zadoščajo pogojem za metriko. Tako se je razvija Fisherjeva informacijska metrika, ki je, okvirno rečeno, Riemannova metrika podana kot koren infinitezimalne relativne entropije.

Fisherjeva metrika predstavlja osnovo informacijske geometrije. Podobna metrika je koren Jensen-Shannonove divergence, definirana kot

$$\sqrt{JS(X \parallel Y)} = \sqrt{(D(X \parallel Y) + D(Y \parallel X))/2}.$$

Dejstvo, da je omenjena količina metrika, je bilo dokazano šele v tem tisočletju [2]. Primerjava teh metrik je podana v [1].

Z vzponom analize podatkov se je relativna entropija ustalila kot ena izmed glavnih mer različnosti porazdelitev, saj pogosto nastopa pri evalvaciji ali konstrukciji optimizacijskih funkcij na podatkih. Relativna entropija na prostoru porazdelitev na n točkah porodi zanimivo geometrijo, ki se v zadnjem času uporablja tudi v okviru topološke analize podatkov [1].

LITERATURA

[1] H. Edelsbrunner, Ž. Virk in H. Wagner, *Topological data analysis in information space*, In Proc. 35th Ann. Sympos. Comput. Geom., 2019.

[2] D. M. Endres in J. E. Schindelin, *A new metric for probability distributions*, IEEE Transactions on Information Theory **49** (2003), 7, 1858–1860.

[3] D. Huffman, *A method for the construction of minimum-redundancy codes*, Proceedings of the IRE **40** (1952), 9, 1098–1101.

[4] N. Kehtarnavaz in N. Kim, *Digital signal processing system-level design using LabVIEW*, Elsevier, 2011.

[5] S. Kullback in R. A. Leibler, *On information and sufficiency*, Annals of Mathematical Statistics **22** (1951), 1, 79–86.

[6] T. Leinster, *Entropy and diversity: The axiomatic approach*, Cambridge University Press, Cambridge, 2021.

[7] D. J. C. MacKay, *Information theory, inference and learning algorithms*, Cambridge University Press, Cambridge, 2003.

[8] V. van der Meer in J. van den Bos, *JPEG file fragmentation point detection using Huffman code and quantization array validation*, In The 16th International Conference on Availability, Reliability and Security (ARES 2021), Association for Computing Machinery, New York, NY, USA, Article 46, 1–7.

[9] C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal **27** (1948), 3, 379–423.