

Nedoločenost angleškega jezika

Uvod

Claude E. Shannon je leta 1948 v prelomnem članku *A Mathematical Theory of Communication* [1] postavil temelje moderne informacijske teorije. V njem je vpeljal nedoločenost (entropijo) kot mero informacije. V članku *Prediction and Entropy of printed English* [2], iz leta 1951, je omenjeno mero uporabil pri analizi angleških besedil. Hotel je ugotoviti, koliko informacije, v povprečju, nosi posamezna črka. Problema se je lotil analitično in eksperimentalno. Rezultati so pokazali, da črka angleške abecede nosi v povprečju približno 1,5 bita informacije [3].

Podoben poskus želimo ponoviti tudi mi. Ugotoviti želimo približek povprečne informacije, ki jo nosijo črke angleške abecede. Izhajamo iz nedoločenosti naključne spremenljivke X , ki predstavlja črko angleške abecede. Ob predpostavki, da so zaporedne črke neodvisne in poznamo njihove verjetnosti, velja:

$$H(X) = - \sum_i^I p(x_i) \log_2 p(x_i) . \quad (1)$$

Če z naključnima spremenljivkama X_p in X_n označimo dve zaporedni črki v besedilu in zgornja predpostavka drži, lahko za nedoločenost para črk zapišemo

$$H(X_p, X_n) = H(X_p) + H(X_n) . \quad (2)$$

Ker so črke v besedilu med seboj odvisne velja $H(X_p, X_n) < H(X_p) + H(X_n)$. Povprečno informacijo $H(X_n|X_p) < H(X_n)$, ki jo dobimo, ko izvemo črko X_n , lahko izračunamo kot

$$H(X_n|X_p) = H(X_p, X_n) - H(X_p) . \quad (3)$$

Ta količina nam predstavlja povprečno informacijo na črko, ki jo dobimo, če predhodno črko že poznamo. Pri obravnavi nam X_p lahko predstavlja poljuben niz črk (pare, trojice, ...)

$$\mathbf{X_p} = (X_1, X_2, \dots, X_{n-1}) . \quad (4)$$

Ker X_n predstavlja naslednjo črko v istem besedilu, lahko zapis poenostavimo

$$\mathbf{X_{pn}} = (X_1, X_2, \dots, X_p, X_n) , \quad \text{sledi} \quad (5)$$

$$H(X_n|\mathbf{X_p}) = H(\mathbf{X_{pn}}) - H(\mathbf{X_p}) \quad (6)$$

$$= - \sum_i^I p(\mathbf{x_{pni}}) \log_2 p(\mathbf{x_{pni}}) + \sum_j^J p(\mathbf{x_{pj}}) \log_2 p(\mathbf{x_{pj}}) . \quad (7)$$

- I predstavlja število vseh možnih nizov dolžine n ,
- J predstavlja število vseh možnih nizov dolžine $n - 1$,
- $p(\mathbf{x_{pni}})$ je verjetnost pojavitve niza $\mathbf{x_{pni}}$ dolžine n ,
- $p(\mathbf{x_{pj}})$ je verjetnost pojavitve niza $\mathbf{x_{pj}}$ dolžine $n - 1$.

V limiti, ko se n približuje neskončnosti, se $H(X_n|\mathbf{X_p})$ približuje pravi vrednosti (povprečni informaciji na znak).

Naloga

Napišite funkcijo z imenom `naloga1` v programskem jeziku Octave, ki izračuna približek povprečne informacije na znak $H(X_n|\mathbf{X}_p)$ in redundanco $R = 1 - H(X_n|\mathbf{X}_p)/\log_2 M$ za dano število poznanih predhodnih črk. Vhodna argumenta funkcije sta stolpični vektor (niz) z besedilom in število poznanih predhodnih črk (celoštevilska vrednost na intervalu $[0,3]$). Besedilo lahko vsebuje vse črke angleške abecede (male in velike), števila in poljubna ločila ter presledke, tabulatorje, ipd. Iz vhodnih podatkov odstranite vse ne-alfanumerične znake in vse črke spremenite v velike tiskane.

Prototip funkcije:

```
function [H,R] = naloga1(besedilo,p)
% besedilo - stolpicni vektor znakov (char)
% p - stevilo poznanih predhodnih znakov; 0, 1, 2 ali 3.
% p = 0 pomeni, da racunamo povprecno informacijo na znak
% abecede brez poznanih predhodnih znakov: H(X1)
% p = 1 pomeni, da racunamo povprecno informacijo na znak
% abecede pri enem poznanim predhodnemu znaku: H(X2/X1)
% p = 2: H(X3/X1,X2)
% p = 3: H(X4/X1,X2,X3)
%
% H - skalar; povprecna informacija na znak abecede
% z upostevanjem stevila poznanih predhodnih znakov p
% R - skalar; redundanca znaka abecede z upostevanjem
% stevila poznanih predhodnih znakov p

H = [];
R = [];
end
```

Testni primeri

Na učilnici se nahajajo tri testni primeri besedil, za katere imate podane tudi povprečno informacijo na znak za predpisano število poznanih predhodnih znakov. Podatki so podani v obliki datotek `.mat`, ki jih naložite v Octave s pomočjo ukaza `load ime_datoteke.mat`. Po izvedenem ukazu se bodo v okolju Octave pojavile spremenljivke z besedili in rezultati. Priloženo imate tudi funkcijo `test_naloga1`, ki jo lahko uporabite za preverjanje pravilnosti rezultatov, ki jih vrača vaša funkcija. Pri testiranju vaših funkcij upoštevajte naslednje omejitve:

- rezultat je pravilen, če se od danega razlikuje za manj kot 10^{-3} ,
- izvajanje funkcije je časovno omejeno na dve minuti.

Namigi

Predprocesiranje:

Vhod: Danes je lep dan. Naredil bom domaco nalogo.

Izhod: DANESJELEPDANNAREDILBOMDOMACONALOGO

Uporabne funkcije: `char`, `upper`, `isalnum`

Računanje povprečne informacije na znak:

Najprej je potrebno izračunati verjetnosti posameznih nizov črk glede na to, kateri približek računamo oz. koliko predhodnih znakov poznamo. Poslužimo se lahko funkcij: `unique`, `histc`, `circshift`. Slednja nam poenostavi računanje verjetnosti nizov črk. Ko imamo izračunane verjetnosti, uporabimo formulo za entropijo. Ali jo lahko izračunamo tudi brez uporabe zanke 'for'?

Literatura

- [1] C. E. Shannon: A mathematical theory of communication. Bell system technical journal, zv. 27, 1948.
- [2] C. E. Shannon: Prediction and entropy of printed English. Bell Systems Technical Journal, zv. 30, str. 50–64, 1951.
- [3] D.G. Luenberger: Information Science, Princeton University, str. 43-44, 2006.