Úvod do kódovania – Úloha č. 1

Termín odovzdania 25. marec 2025, vo formáte pdf cez *Google Classroom*

(R) bude znamenať odkaz na knižku S. Romana, (G) P. Garretta a (H) R. Hilla; uvedené bude spravidla číslo príkladu/problému (kapitola.sekcia.problém), alebo číslo strany, kde sa o danej veci píše. Neočakávajú sa úplne kompletné a perfektné riešenia. Aj čiastkové riešenia s drobnými opomenutiami, logickými medzerami, či neporiadnym zápisom si môžu vyžadovať veľa práce a námahy a môžu dostať plný počet bodov. T.j. nemusíte týmito domácimi úlohami stráviť všetok svoj voľný čas počas nadchádzajúceho semestra. Na druhej strane, očakáva sa preukázanie výraznejších snáh, aby domáce úlohy splnili svoj účel – naučiť sa, resp. samostatne objaviť niečo netriviálne z preberaného materiálu.

Vždy je tu možnosť absolvovania konzultácií. Tiež môže pomôcť preskúmanie viacerých príkladov daného fenoménu pri hľadaní dôkazu všeobecného tvrdenia. Akceptované budú aj riešenia založené na počítačových simuláciách, pokiaľ budú primerane zdôvodnené a bude to v danom kontexte dávať zmysel.

Domáca úloha bude obsahovať príklady s celkovým ohodnotením prevyšujúcim 50 bodov, čo je maximum, ktoré sa v rámci jednej úlohy dá získať. To znamená, že si môžete zvoliť, ktorým príkladom sa budete venovať a ktoré nakoniec odovzdáte. Keďže sa dá očakávať, že nie všetky riešenia budú za plný počet bodov, má zmysel odovzdať príklady, ktorých celkové hodnotenie prevyšuje 50 bodov.

1. (8 bodov) Uvažujme náhodný dostatočne dlhý text v angličtine. Použite (nájdite zdroje) nejaké zmysluplné odhady distribúcie dĺžok slov v štandardnom anglickom texte (napr. 3.2 % má dĺžku 1 znak, 17.0 % má dĺžku 2 znaky, atď. – tieto čísla môžu byť vymyslené) pre odhad frekvencie symbolu medzery v texte, ktorý obsahuje len slová a medzery. Následne analyzujte dostatočne dlhý anglický text podľa vlastného výberu a nájdite frekvenciu výskytu medzery. Porovnajte empiricky zistenú frekvenciu s odhadovanou pravdepodobnosťou výskytu medzery na základe distribúcie dĺžok slov a vysvetlite prípadný rozdiel.

2. (12 bodov) Uvažujme náhodný dostatočne dlhý anglický text s relatívnymi frekvenciami $p_a, p_b, \ldots, p_z$ znakov anglickej abecedy podľa (G), resp. podľa (R). Porovnajte takýto text s textom generovaným náhodným generátorom symbolov z anglickej abecedy (generuje sa znak po znaku) s rovnakými pravdepodobnosťami $p_a, p_b, \ldots, p_z$.

a) Nájdite vzorec pre výpočet pravdepodobnosti $p_{\alpha\beta}$ výskytu blokov dĺžky 2 zložených z po sebe nasledujúcich znakov $\alpha\beta$ vyjadrený pomocou pravdepodobností p_α a p_β v náhodne generovanom texte. Porovnajte s relatívnymi frekvenciami (zopdpovedajúcimi pravdepodobnostiam $p'_{\alpha\beta}$) takýchto blokov v náhodnom dostatočne dlhom anglickom texte. (Relatívne frekvencie zistite empiricky alebo využitím nejakého serózneho zdroja) Vysvetlite rozdiel medzi zodpovedajúcimi pravdepodobnosťami a relatívnymi frekvenciami.

b) Použijúc vyššie spomenutý prístup, dá sa zistiť pôvod textu? Zdôvodnite.

c) Existuje n pre ktoré sa relatívne frekvencie, resp. pravdepodobnosti výskytu blokov dĺžky n v dvoch druhoch textov začnú približovať? Alebo budú rozdiely medzi pravdepodobnosťami s rastúcim n rásť? Zdôvodnite, podporte svoje argumenty prípadne nejakými výpočtami.

d) Ktorá entropia by mala byť pre náhodný dostatočne dlhý anglický text vyššia? Entropia, kde za zdroj považujeme bloky znakov dĺžky 2 a uvažujeme príslušné (empirické) pravdepodobnosti $p'_{\alpha\beta}$ alebo entropia zdroja, kde predpokladáme, že text prichádza náhodne znak po znaku? Je entropia takéhoto textového zdroja jednoznačne určená? Zdôvodnite.

3. (12 bodov) Uvažujme *Morseovu abecedu* ako binárny 'čiarka-bodka' kód.

a) Určite pravdepodobnosť prichádzajúcej čiarky vs. bodky v správe pozostávajúcej z náhodnej postupnosti znakov anglickej abecedy zakódovanej pomocou Morseovej abecedy, pričom predpokladáme, že všetky znaky anglickej abecedy (bez medzier) sa vyskutujú s rovnakou pravdepodobnosťou. Vypočítajte entropiu takéhoto binárneho zdroja.

b) Určite pravdepodobnosť prichádzajúcej čiarky vs. bodky v správe pozostávajúcej z 'priemerného' anglického textu. Vypočítajte entropiu v tomto prípade a porovnajte s výsledkom časti a).

c) Poskytnite 'matematické' zdôvodnenie rozdielu medzi týmito dvoma entropiami.

4. (8 bodov) Nech $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ je okamžite rozkódovateľný kód nad r-árnou abecedou $\mathcal{A}$. Dokážte, že nasledujúce podmienky sú ekvivalentné:

a) $\mathcal{C}$ je *maximálny* okamžite rozkódovateľný kód, t.j. žiadne ďalšie slovo $w \in \mathcal{A}^*$ sa nedá pridať do $\mathcal{C}$ bez straty vlastnosti okamžitej rozkódovateľnosti.

b) Každý konečný reťazec v $\mathcal{A}^*$ má ako prefix niektoré z kódových slov v $\mathcal{C}$ alebo je prefixom nejakého slova v $\mathcal{C}$.

c) Platí rovnosť

$$\sum_{i=1}^{n} \frac{1}{r^{len(c_i)}} = 1.$$

5. (8 bodov) Vyriešte nasledujúce cvičenia z prvej kapitoly Romanovej knihy:

a) Vypočítajte $H(\frac{1}{a}, \frac{1}{a}, \ldots, \frac{1}{a}, \frac{2}{a}, \ldots, \frac{2}{a}, \ldots \frac{2}{a})$, kde $\frac{1}{a}$ vystupuje vo funkcii entropie s-krát a $\frac{2}{a}$ t-krát, t.j. $s + 2t = a$.

b) Koľko informácie získame výberom zo štandardnej sady hracích kariet, keď výber každej hodnoty karty je rovnako pravdepodobný, ale čierna farba má dvojnásobnú pravdepodobnosť oproti červenej? Ako to súvisí s časťou a)?

6. (8 bodov) Majme prirodzené číslo $k > 1$. Rozhodnite aký je vzťah medzi entropiami $H(p_1, p_2, \ldots, p_n)$ a $H(\frac{p_1}{k}, \frac{p_1}{k}, \ldots, \frac{p_1}{k}, p_2, \ldots, p_n)$ (môžu sa rovnať, jedna môže byť vyššia ako druhá; odpoveď tiež môže závisieť od pravdepodobnostnej distribúcie $(p_1, p_2, \ldots, p_n)$ a čísla k). Zdôvodnite správnosť vašej odpovede a tiež poskytnite neformálne vysvetlenie.

7. (8 bodov) Predpokladajme, že máme zdroj $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ s pravdepodbnostnou distribúciou $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$. Chceli by sme zvýšiť entropiu tohto zdroja zmenou dvoch pravdepodobností (celkový súčet zostáva stále 1), t.j. pridať ϵ k jednej z pravdepodobností, $p_i + \epsilon$, a odpočítať rovnakú hodnotu od inej z pravdepodobností, $p_j - \epsilon$. Aký je najefektívnejší spôsob, ako toto docieliť? Alebo na výbere dvojice pravdepodobností p_i, p_j nezáleží? Zdôvodnite.

8. (12 bodov) Uvažujme *Morseovu abecedu* ako ternárny 'čiarka-bodka-medzera' kód. T.j. každá čiarka-bodka postupnosť kódujúca písmeno anglickej abecedy je nasledovaná medzerou.

a) Určite pravdepodobnostnú distribúciu pre priestor {čiarka, bodka, medzera} v náhodnej postupnosti písmen anglickej abecedy, pričom sa každé písmeno vyskytuje s rovnakou pravdepodobnosťou. Vypočítajte entropiu takéhoto zdroja.

b) Porovnajte entropie z 3a) a 8a) a poskytnite matematické vysvetlenie rozdielu medzi nimi.

c) Určite pravdepodobnosť čiarky ako funkcie priemernej dĺžky slova ℓ (bez ohľadu na jazyk) ak sú všetky slová (t.j. reťazce znakov dĺžky ℓ) rovnako pravdepodobné.

9. (8 bodov) Pozrite sa na Vetu 1.2.8 v Romanovej knihe a jej dôkaz. Jej znenie tvorí nerovnosť s parametrami n, k a λ. Existujú nejaké ich hodnoty, ktoré povedú k rovnosti? Aký veľký môže byť rozdiel medzi pravou a ľavou stranou nerovnosti? Svoje odpovede primerane zdôvodnite.