MI-SPOL-17 (MI-SPI) ●✓

Základy teorie informace a kódování, entropie

Kódování: zobrazení $C: X \to D^*$ ← diskrétní náhodná veličina

$D^* = \bigcup_{k=1}^{\infty} D^k$

D^* - množina konečných řetězců symbolů $\in$ - dané abecedy D

obraz $C(x)$ - kódové slovo → jeho délka $l(x)$

Střední délka kódu: $L(C)$ náhodné veličiny X s $p(x)$ rozdělením

$$L(C) = \sum_{x \in X} l(x)\, p(x) = E\, l(X)$$

$L(C)$ nemůže být menší než entropie daného textu $H(X)$

→ pokud se rovnají, máme optimální kód

Nesingulární kódy

→ pokud je zobrazení C prosté = musím jednoznačně dekódovat daný kód symbolu

- nemusí ale platit pro zprávu

$$x \neq x' \Rightarrow C(x) \neq C(x')$$

Jednoznačně dekódovatelné kódy

$\left[C^*(x_1 x_2 \dots x_n) = C(x_1) C(x_2) \dots C(x_n) \right]$

→ pokud je C^* nesingulární

Instantní (prefixový) kód

→ žádné kódové slovo není prefixem jiného

→ sobě rozeznáme kódové slovo, musím na další

Huffmanovo kódování

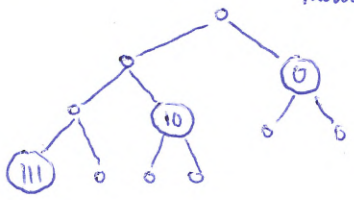
- agregace vždy nejmíně pravděpodobných → zpětné přiřazení hodnot
- je optimální: pro Huff. kód C^* a libovolný unik. dek. kód C' $L(C^*) \leq L(C')$

Kraft - McMillan inequality

pro libovolný jednoz. dec. kód musí délky kódových slov splnit nerovnost:

$$\sum_i D^{-l_i} \leq 1$$

$|D$ - činná abeceda$|$ — i — délky kódových slov



\+ pro každou n-tici délek splňující, existuje instantní kód

→ McMillan rozšířil nerovnost i na uniq. dekódovatelný

Optimální kódy

$L(C) \geq H_D(X)$ ← entropie — platí pouze pro instantní (prefixový kód)!

- rovnost nastává pokud $D^{-l_i} = p_i$ pro všechna i

kódová slova — odpovídající délky

$$H_D(X) \leq L(C^*) < H_D(X) + 1$$

→ optimální kód se vzdálí od dolní meze max o 1

Entropie – míra neuspořádanosti

(pravděpodobnostní funkce $p(x) \rightarrow \forall x \in \mathbb{R} : p(x) = P(X = x)$)

Entropie $H(X) =$ $\boxed{H(X) = -\sum_{x \in X} p(x) \log_2 p(x)}$ – zdání pouze na rozdělení, ne na číselných hodnotách

- báze $H_b(X)$ může měnit základ logaritmu $\log_b p(x)$

Entropie je střední hodnota míry nevědomosti (vlastní informace)

$$I(x) = -\log_2 p(x)$$ ← vždy nezáporná, rovná 0 u jistých jevů

$$H(X) = \mathbb{E}(I(x))$$

- je nejvyšší pro rovnoměrné rozdělení (nevíme nic)

= očekávaná míra nevědomosti náhodné veličiny H

optimální počet binárních otázek: mezi $H(X)$ a $H(X)+1$ (viz optimální kódy)

Sdružená entropie $H(X,Y)$ – entropie sdruženého rozdělení několika náh. veličin

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$

– analogicky se definuje pro náhodné vektory

Podmíněná entropie $H(Y|X)$ – podm. ent. náhodných veličin X, Y se sdruženým rozdělením $p(x,y)$

$$H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)$$

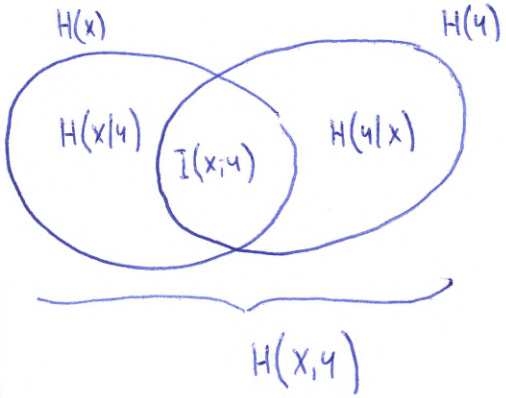
Řetězcové pravidlo

$$H(X,Y) = H(X) + H(Y|X)$$

⇒ Úplná část informace je v Y navíc oproti X

Relativní entropie (Kullback – Leiblerova vzdálenost)

$$D(p\|q) = \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)}$$



Vzájemná informace

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

$$= D(p(x,y)\,\|\,p(x)p(y))$$

– míra informace, kterou sdílí obě veličiny X a Y

→ vzdálenost od nezávislosti

$$I(X,Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$