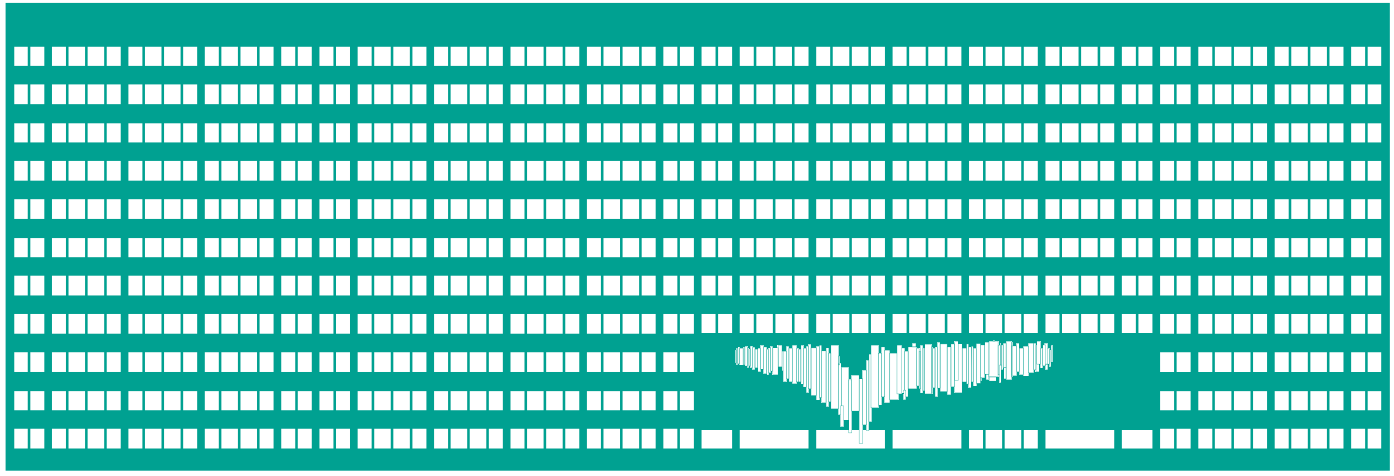
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

Analýza a komprese signálu

Teorie informací

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA404

michal.vasinek@vsb.cz

28.2.2023

Obsah

- Entropie - neurčitost = informace
- Entropie - fundamentální limit v kompresi dat
- Relativní entropie
- Vzájemná informace

Motivace

Otázka

Předpokládejte, že máme čtyři různé symboly (nebo libovolné jiné objekty). Kolik bitů potřebujeme abychom je dokázali od sebe odlišit?

Motivace

Question

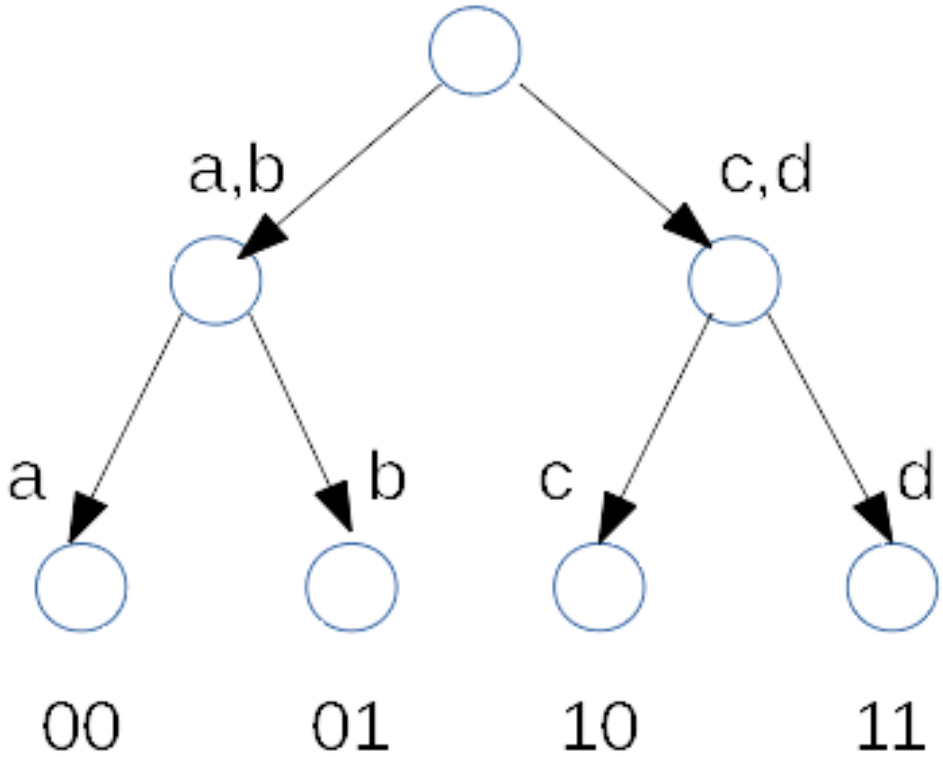
Předpokládejte, že máme čtyři různé symboly (nebo libovolné jiné objekty). Kolik bitů potřebujeme abychom je dokázali od sebe odlišit?

Zkusme řešení pomocí binárního dotazování (ANO/NE otázek):

- Vyskytuje se symbol v množině {a,b}. Pokud Ano pak si poznačme 0, jinak si poznačíme 1.
- Pokud byla odpověď Ano, pak víme, že symbol je buď **a** nebo **b**, takže další otázka bude: Je to **a**? Pokud Ano poznačíme si 0, pokud ne poznačíme si 1.
- Pokud byla odpověď Ne, pak víme, že symbol je buď **c** nebo **d**, takže další otázka bude: Je to **c**? Pokud Ano poznačíme si 0, pokud ne poznačíme si 1.

Pro rozlišení mezi symboly tedy potřebujeme dvě binární čísla, jinými slovy 2 bity.

Pro rozlišení mezi symboly tedy potřebujeme dvě binární čísla, jinými slovy 2 bity.



Obrázek: Binární dotazování.

Kontrolní otázka

Otázka - 1 bonusový bod

Co když různých symbolů bude 8, 16, 32 etc., kolik pak budeme potřebovat bitů pro jejich odlišení?

Definice

Abeceda

Abeceda je množina symbolů. Obvykle značíme Σ (sigma), její velikost(počet různých symbolů) σ. Například abeceda DNA je: $\Sigma = \{A, C, G, T\}$ a její velikost je $\sigma = 4$.

Náhodná proměnná

Náhodná proměnná je funkce, která přiřadí hodnotu náhodné údalosti. Obvykle značíme velkými písmeny, např. X, Y. Příkladem může být náhodná proměnná reprezentující hod kostkou. V kompresi dat náhodná proměnná reprezentuje událost, že přečteme nějaký symbol.

Definice - pokračování

Pravděpodobnostní funkce

Pravděpodobnostní funkce je funkce, která vrátí pravděpodobnost, že diskrétní náhodná proměnná nabývá právě nějaké konkrétní hodnoty x. $p(x) = Pr\{X = x\}, x \in \Sigma$.

Zpráva

Zpráva je sekvencí náhodných proměnných: $X_1 X_2 \ldots X_n$. Délka zprávy je n. Příkladem zprávy může být libovolný soubor na vašem disku, zobrazená webová stránka, video na youtube... .

Logaritmy

Pravidlo o součinu

$$\log ab = \log a + \log b$$

Pravidlo o podílu

$$\log \frac{a}{b} = \log a - \log b$$

Všechny logaritmy, které budeme uvažovat budou mít základ roven 2: $\log_2$.

Entropie náhodné proměnné

Entropie reprezentuje míru neurčitosti náhodné proměnné.

- Používá se k měření množství informace, kterou v průměru budeme potřebovat k popisu náhodné proměnné.
- V kompresi dat nám entropie udává fundamentální limit, kterého jsme schopni dosáhnout kompresními algoritmy. Nelze žádnou zprávu zkomprimovat lépe, než na úroveň entropie. Pokud by náš kompresní algoritmus dokázal zkomprimovat data lépe, než je limit daný entropií, pak ztratíme schopnost data zpětně zrekonstruovat.

Entropie

- Předpokládejme, že každý symbol má stejnou pravděpodobnost a velikost abecedy je σ.
- $Pr(X = x) = \frac{1}{\sigma}$ - pravděpodobnost jednoho konkrétního symbolu.
- $\log \sigma$ - **minimální** počet bitů potřebných pro zakódování jednoho symbolu.
- Abychom obdrželi průměrný počet bitů na symbol musíme sečíst počty bitů za každý symbol a podělit σ.

Entropie - stejně pravděpodobné symboly

Entropie $H(X)$ je průměrný počet bitů, který potřebujeme pro reprezentaci jednoho symbolu.

$$H(X) = \sum_{x \in \Sigma} \frac{1}{\sigma} \log \sigma = \frac{1}{\sigma} \sum_{x \in \Sigma} \log \sigma = \log \sigma$$

Entropie - Příklad

Diskrétní náhodná proměnná X. Dvě možné hodnoty buď 0 nebo 1.

$$X = \begin{cases} 1 & \text{s pravděpodobností } p, \\ 0 & \text{s pravděpodobností } 1-p. \end{cases}$$

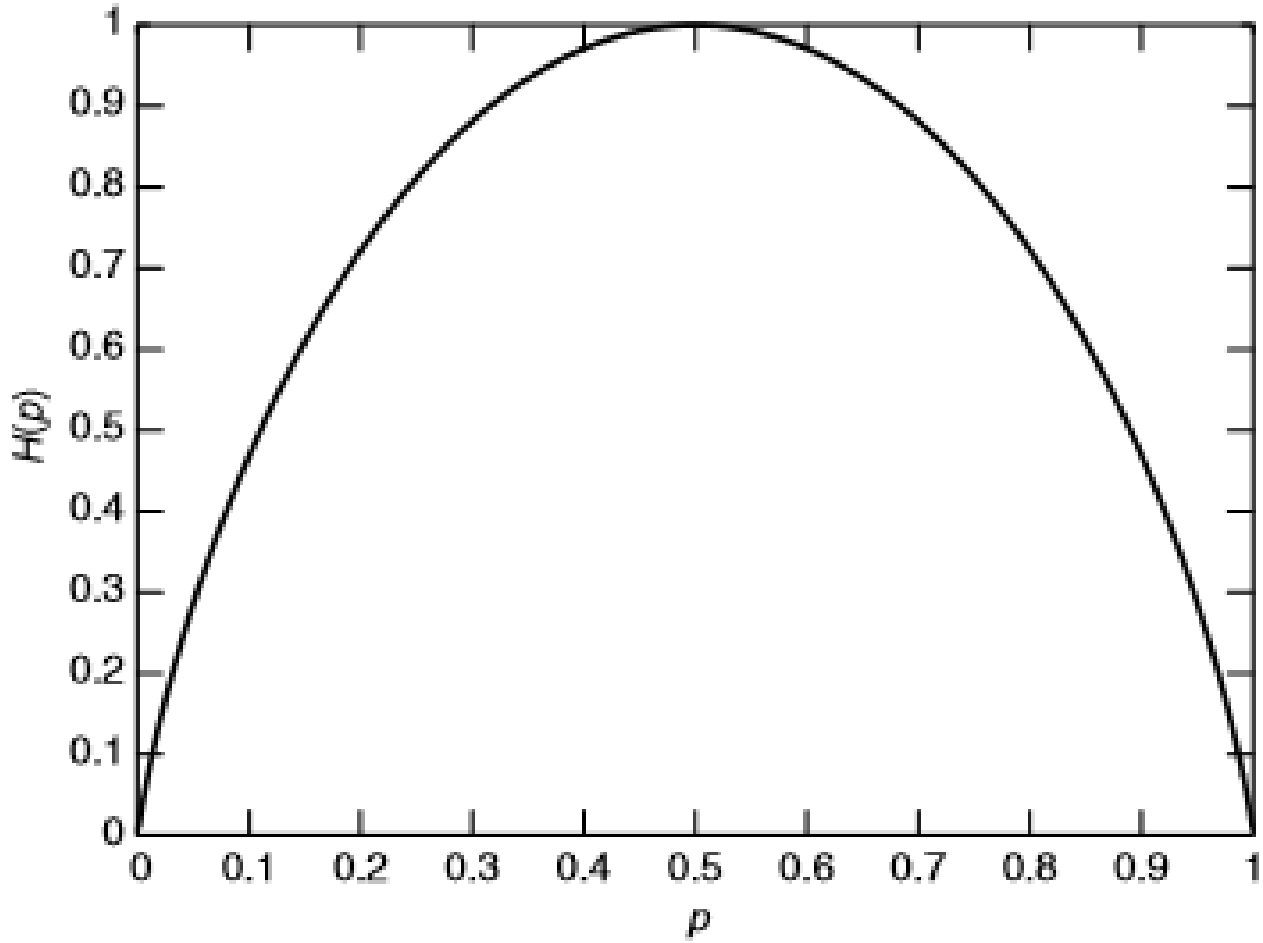
Potom

$$H(X) = -p \log p - (1-p) \log (1-p)$$

- $p(0) = p(1) = 0.5$ then

$$H(X) = -0.5 \log 0.5 - (1-0.5) \log (1-0.5) = 1 \text{bit}$$

Entropie - Příklad 2



Obrázek: $H(p)$ vs. p.

Další příklad

- $p(a) = 0.5$
- $p(b) = 0.25$
- $p(c) = 0.25.$

S použitím binárního dotazování:

- Je symbol **a**. Pokud ano zapíšeme si 0 jinak si zapíšeme 1.
- Pokud byla odpověď Ano, pak víme, že symbolem bylo **a**.
- Pokud byla odpověď Ne, pak víme, že symbol je buď **b** nebo **c**, tudíž se můžeme zeptat: je symbolem **b**? Pokud Ano, pak si zapíšeme 0 jinak 1.

Další příklad

- $p(a) = 0.5 => -\log 0.5 = 1$
- $p(b) = 0.25 => -\log 0.25 = 2$
- $p(c) = 0.25 => -\log 0.25 = 2$

V průměru máme:

$$-0.5 \log 0.5 - 0.25 \log 0.25 - 0.25 \log 0.25 = -0.5 - 0.25*2 - 0.25*2 = 1.5 \text{bits}$$

Entropie je dána vztahem:

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

Jedná se o jednu z nejslavnějších rovnic vědy: Shannonova rovnice.

Entropie - příklad 2

Nechť

$$X = \begin{cases} a & \text{s pravděpodobností } \frac{1}{2}, \\ b & \text{s pravděpodobností } \frac{1}{4}. \\ c & \text{s pravděpodobností } \frac{1}{8}, \\ d & \text{s pravděpodobností } \frac{1}{8}. \end{cases}$$

Entropie je:

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{7}{4} \text{ bits}$$

Shannonuv experiment - entropie jazyka

_____rese _at_____k____ompr____c____a_) _____p_a__va__
____itaco__c__da_ s
cil_m_z_e_s_____e__ch__bje__(_ed_o_ka__ba____ _ri
so_c_snem ________a_i__n_

Shannonuv experiment - entropie jazyka

Ko_ _rese dat (t_ _é _omp_ _m_ _ _ d_t_ je_ _ _racova_i
po_itac_vych dat s ci_em zmensi_ _eji_h_o_ _ _m
(j_dno_ka:_bajt) _ri _ _u_asnem_zacho_ani_ _n_ormaci v
_at_ch_o_s_zenyc_._ _kol_ _ _k_mprese da_ _je _ _en_ _ _
datov_ _tok_

Shannonuv experiment - entropie jazyka

Komprese dat (take komprimace dat) je zpracovani pocitacovych dat s cilem zmensit jejich objem (jednotka: bajt) pri soucasnem zachovani informaci v datech obsazenych. Ukolem komprese dat je zmensit datovy tok pri jejich prenosu nebo zmensit potrebu zdroju pri ukladani informaci.

Entropie - jiné odvození

Entropie - Definice

Počet bitů, které potřebujeme abychom rozlišili jednu zprávu od ostatních zpráv, pokud disponujeme znalostí o četnostech symbolů.

- Mějme velmi dlouhou zprávu o N symbolech nad abecedou $\Sigma = \{0, 1\}$.
- Spočítáme četnosti symbolů $f(0)$ a $f(1)$.

Otázka

Kolik různých zpráv můžeme vytvořit pokud použijeme $f(0)$ nul a $f(1)$ jedniček?

Otázka

Kolik různých zpráv můžeme vytvořit pokud použijeme $f(0)$ nul a $f(1)$ jedniček?

Permutace s opakováním:

$$\frac{N!}{f(0)!f(1)!}$$

Abychom mohli odlišit k různých objektů potřebujeme $\log k$ bitů. To samé platí pro různé zprávy stejné délky:

$$\log \frac{N!}{f(0)!f(1)!} = \log N! - log f(0)! - log f(1)!$$

Stirlingova aproximace

Problém

Jak spočítat logaritmus faktoriálu?

Odpověď: použijeme Stirlingovu approximaci.

$$\log k! = k \log k - k$$

Úkol

S použitím Strilingovy rovnice zjednoduště výraz:

$$\log \frac{N!}{f(0)!f(1)!} = \log N! - log f(0)! - log f(1)!$$

Entropie - odvození pomocí dlouhých zpráv

- $N = f(0) + f(1)$
- $p(x) = \frac{f(x)}{N} => f(x) = p(x)N$

$$\begin{aligned}
\log \frac{N!}{f(0)!f(1)!} &= \log N! - logf(0)! - logf(1)! \\
&= N \log N - N - f(0) \log f(0) + f(0) \\
&- f(1) \log f(1) + f(1) \\
&= N \log N - f(0) \log f(0) - f(1) \log f(1) \\
&= N \log N - Np(0) \log p(0)N - Np(1) \log p(1)N \\
&= N(\log N - p(0) \log p(0)N - p(1) \log p(1)N)
\end{aligned}$$

Entropie - odvození pomocí dlouhých zpráv

$$p(x) \log Np(x) = p(x) \log p(x) + p(x) \log N$$

$$-p(0) \log N - p(1) \log N = -(p(0) + p(1)) \log N = -\log N$$

Spojením dohromady obdržíme:

$$\begin{aligned} &= N(\log N - p(0) \log p(0)N - p(1) \log p(1)N) \\ &= N(-p(0) \log p(0) - p(1) \log p(1) \\ &= NH(X) \end{aligned}$$

Protože se zpráva skládá z N symbolů, podělením N obdržíme entropii vztaženou k jednomu symbolu, t.j. Shannonu entropii.

Empirická entropie

Empirická entropie je entropii vypočtenou na základě znalosti rozdělení symbolů v reálných datech(souborech, proudech dat, ...)

- Spočítejte četnosti symbolů $f(x)$ ve zprávě m o délce n.
- Spočítejte empirickou pravděpodobnost: $p(x) = f(x)/n$.
- Spočtěte entropii pomocí Shannonovy rovnice:

$$H(X) = -\sum_{x \in \Sigma} p(x) \log p(x)$$

Sdružená entropie

Sdružená entropie $H(X,Y)$ dvojice diskrétních náhodných proměnných (X,Y) se sdruženým rozdělením $p(x,y)$ je definována jako:

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

Například, nemusíme vždy uvažovat jen rozdělení jednotlivých symbolů, ale můžeme uvažovat dvojice symbolů, tzv. digramy či bigramy v textu

Podmíněná entropie

Typicky v angličtině se vyskytuje mnoho slov začínajících na th, například, the, this, there, that, atd. Můžeme nějak využít znalosti, že symbol t je často následován symbolem h?

Podmíněná entropie

Pokud $(X, Y) \sim p(x, y)$, podmíněná entropie $H(Y|X)$ je definována jako:

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

Podmíněná entropie - poznámky

- Kdykoli přečteme v anglickém textu symbol t, nejspíše bychom zkusili hádat, že následující symbol bude h.
- Pokud bychom náš kompresní algoritmus založili na podmíněné pravděpodobnosti $p(Y|X = t)$ spíše než na marginální pravděpodobnosti $p(X)$, budeme schopni navrhnout efektivnější kódy pro reprezentaci symbolu h. Připomeňme si, že potřebujeme $-\log p$ bitů pro reprezentaci symbolu.
- Toto můžeme dále zobecnit na: $H(Y|X_1X_2\ldots X_n)$. Zde $X_1X_2\ldots X_n$ je kontext(vyjádřený jako podmíněná pravděpodobnost) pomocí kterého budeme kódovat následující symbol.

Student teorie informace vstoupil první den na vysoké škole do podivného a bizarního světa. Jedinými zvuky bylo občasné vyřknutí čísla jedním z profesorů a následný smích. Jeden z profesorů řekl "52", následovala krátká pauza a pak se ozval smích. Někdo jiný řekne "713", to samé, všichni padnou smíchy k zemi.
"Co se to tu děje?"zeptal se svého učitele.
"Vyprávíme si vtipy,"řekl jeho učitel.
"Vyprávíte si vtipy?"
"Ano, víte, všichni tu pracujeme tak dlouho, že známe vtipy jeden druhého. Je jich tisíce. A tak jsme jako informační teoretici použili kompresi dat. Prostě jsme jim všem přiřadili čísla, od 0 do 999. Ušetří to spoustu času a námahy. Chtěli byste to zkusit? Stačí říct libovolné číslo od 0 do 999..."
Nebyl úplně přesvědčený. Ale zkusil to. Velmi tiše zašeptal "477".
Sotva se ozvalo zašustění.

Podíval se na svého učitele. "Co se děje?"zeptal se. "Zkus to znovu,"řekl učitel.
A tak to zkusil. "318"- opět totéž, nic, sotva šelest.
"Něco je špatně,"říká.
"No,"říká učitel, "nejde ani tak o ten vtip, jako o způsob, jakým ho vyprávíte!"
Student nakonec uspěl náhodou tím nejnečekanějším způsobem. Zavolal číslo mimo rozsah 0 až 999. "Mínus 105,"řekl.
Nejprve se ozval ohromený údiv, pak se rozesmál nejprve jeden profesor, pak druhý, pak další, až se všichni váleli po zemi a drželi se za boky.
Tento vtip nikdo předtím neslyšel.

Relativní entropie

- Uvažujme, že dostanete data jejichž rozdělení je $p(x)$.
- Ale já vám řeknu, že rozdělení je $q(x) \neq p(x)$.

Otázka

Jaký bude rozdíl v naší schopnosti zakódovat data, pokud namísto skutečného rozdělení $p(x)$ použijeme rozdělení $q(x)$?

Odpovědí je relativní entropie!

Relativní entropie

- Uvažujme, že dostanete data jejichž rozdělení je $p(x)$.
- Ale já vám řeknu, že rozdělení je $q(x) \neq p(x)$.

Použijete $-\log q(x)$ bitů k zakódování symbolu x, nicméně v našich datech se symbol x vyskytuje s pravdpodobností $p(x)$ takže váš kód bude nejlépe reprezentován pomocí:

$$-\sum_{x \in X} p(x) \log q(x)$$

$$\begin{aligned} D(p||q) &= -\sum_{x \in X} p(x) \log q(x) - \sum_{x \in X} p(x) \log p(x) \\ &= -\sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \end{aligned}$$

Vzájemná informace

Uvažujme dvě náhodné proměnné X a Y se sdruženým rozdělením $p(x, y)$ a marginálním rozdělením $p(x)$ and $p(y)$. Vzájemná informace $I(X; Y)$ je dána jako relativní entropie mezi sdruženým rozdělením a součinovým rozdělením $p(x)p(y)$.

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

- Sdružené rozdělení vs. marginální rozdělení - například pravděpodobnost, že se digram ab vyskytuje ve zprávě $m = abbabbbaaa \ldots$ je $p(a, b)$ zatímco marginální pravděpodonosti jsou $p(a)$ a $p(b)$

Vzájemná informace

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= \sum_{x,y} p(x,y) \log p(x|y) - \sum_{x,y} p(x,y) \log p(x) \\
&= \sum_{x,y} p(x,y) \log p(x|y) - \sum_{x} p(x) \log p(x) \\
&= H(X) - H(X|Y)
\end{aligned}
$$

Tudíž, dodatečná informace, kterou získáme znalostí náhodné proměnné Y snižuje nejistotu náhodné proměnné X.

DÄŻkuji za pozornost

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA404

michal.vasinek@vsb.cz

28.2.2023

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY