Všechno, co jste chtěli vědět z teorie pravděpodobnosti, z teorie informace a ...

báli jste se zeptat

(1. část)

Jedinečnou funkcí statistiky je, že umožňuje vědci číselně vyjádřit nejistotu v jeho závěrech. (G. W. Snedecor)

(pro potřeby přednášky Úvod do strojového učení, PFL054)

Statistika se těší pochybnému vyznamenání tím, že je nejvíce nepochopeným vědním oborem. Neznamená to však, že je nejméně známá. Nepochopení nějaké věci totiž předpokládá, že se o ní něco ví, nebo přinejmenším se myslí, že se ví. O statistice však panuje všeobecné mínění, že z každého, kdo se naučil ve škole trochu počítat, lze bez obtíží udělat statistika prostě tím, že se mu tak říká. (H. Levinson)

- Náhodný pokus
- Nastal jev A
- *Pravděpodobnost* má modelovat relativní četnost

- Výsledek není předem znám
- Pravdivost tvrzení o výsledku pokusu

ZÁKLADNÍ POJMY

- **universum** (diskrétní, spojité) Ω
- jev jistý Ω, jev nemožný $\emptyset$
- sjednocení jevů $\cup_{i=1..n} A_i$
- průnik jevů $\cap_{i=1..n} A_i$
- jev opačný $A^c = \Omega - A$
- elementární jev $\omega \in \Omega$
- **algebra** $\mathcal{A}$: systém podmnožin Ω uzavřený na sjednocení, průnik, doplněk; Ω , $\emptyset \in \mathcal{A}$
- náhodný jev $A \in \mathcal{A}$

ZÁKLADNÍ POJMY (POKRAČOVÁNÍ)

- pravděpodobnost P reálná fce df na $\mathcal{A}$
  - $A \in \mathcal{A} \Rightarrow 1 \geq P(A) \geq 0, A \in \Omega$
  - **A,B vzájemně disjunktní** $\Rightarrow$ $P(A \cap B) = P(A) + P(B)$
  - $P(\Omega) = 1$ $(P(\varnothing) = 0)$

Klasický pravděpodobnostní prostor

- **konečný** prostor elementárních jevů, Ω
- algebra $\mathcal{A}$
  - $\emptyset \in \mathcal{A}$, $\boldsymbol{A} \in \mathcal{A} \Rightarrow \boldsymbol{A}^{\mathbf{c}} \in \mathcal{A}$
  - $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{A} \Rightarrow \boldsymbol{A} \cup \boldsymbol{B} \in \mathcal{A}$
  - $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{A} \Rightarrow \boldsymbol{A} \cap \boldsymbol{B} \in \mathcal{A}$
- pravděpodobnost P
- $P(\boldsymbol{A}) = |\boldsymbol{A}| / |\Omega|$ (na konečné množině Ω zavedena pravděpodobnost)

Jaká je pravděpodobnost, že při házení třemi mincemi najednou padnou právě 2 panny?

Ω = ?, $\boldsymbol{A}$ = ?, P($\boldsymbol{A}$) = ?

- Ω = {OOO, OOP, OPO, OPP, POO, POP, PPO, PPP}
- A ={PPO, POP, OPP}
- P(A) = 3/8

- přechod od **konečného** prostoru elementárních jevů k prostoru **spočetnému**

Kolgomorova definice pravděpodobnosti

pravěpodobnostní prostor

- prostor elementárních jevů, Ω
- σ–algebra, $\mathcal{A}$
  - $\emptyset \in \mathcal{A}$, $\boldsymbol{A} \in \mathcal{A} \Rightarrow \boldsymbol{A}^{\boldsymbol{c}} \in \mathcal{A}$
  - $\boldsymbol{A}_{\boldsymbol{i}} \in \mathcal{A} \Rightarrow \cup_{\boldsymbol{i=1..\infty}} \boldsymbol{A}_{\boldsymbol{i}} \in \mathcal{A}$
  - $(\boldsymbol{A}_{\boldsymbol{i}} \in \mathcal{A} \Rightarrow \cap_{\boldsymbol{i=1..\infty}} \boldsymbol{A}_{\boldsymbol{i}} \in \mathcal{A})$

Kolgomorova df psti (pokračování)

- **P:** $\mathcal{A} \rightarrow [0,1]$
  - **P (*A*)** ≥ 0, ***A*** $\in \Omega$
  - **P**$(\Omega) = 1$ (**P**$(\varnothing) = 0$)
  - ***A*$_1$, *A*$_2$,... vz. disjunktní množiny** $\in \mathcal{A}$,
    - **P**$(\cup_{i=1..\infty} A_i) = \Sigma_{i=1..\infty}$**P**$(A_i)$
  - **P** = ?

Složená pravděpodobnost, nezávislost jevů,

- Jevy A, B jsou nezávislé $\Leftrightarrow$
  $P(A,B)=P(A)*P(B)$

- **Složená pravděpodobnost** $P(A,B)$
- **Podmíněná pravděpodobnost** $P(A|B)$
  - úplně závislé jevy $P(A|B) = 1$
  - závislé $P(A|B) = ?$
  - nezávislé $P(A|B) = P(A)$

- **Bayesův vzorec**
  - $P(A|B) = P(A,B) / P(B)$

Bayesův inverzní vzorec

- $P(A|B) = P(A)*P(B|A)/P(B)$

Náhodná veličina

- $\omega \in \Omega$; $\boldsymbol{X}: \Omega \to \mathcal{R}$
- $P[\boldsymbol{X} = x] = P(\{\omega \in \Omega; X(\omega) = x\}$
- $P[\boldsymbol{X} = x]$ rozdělení náhodné veličiny $\boldsymbol{X}$
- diskrétní, spojitá
- střední hodnota náhodné veličiny

$$E[\omega] = 1/\Omega \sum_{\omega \in \Omega} X(\omega) = \sum_x x\, P[\boldsymbol{X} = x]$$

Statistik je ten, kdo s hlavou v rozpálené troubě a s nohama v nádobě s ledem na dotaz, jak se cítí, odpoví: "V průměru se cítím dobře.„ (anonym)

Teorie informace

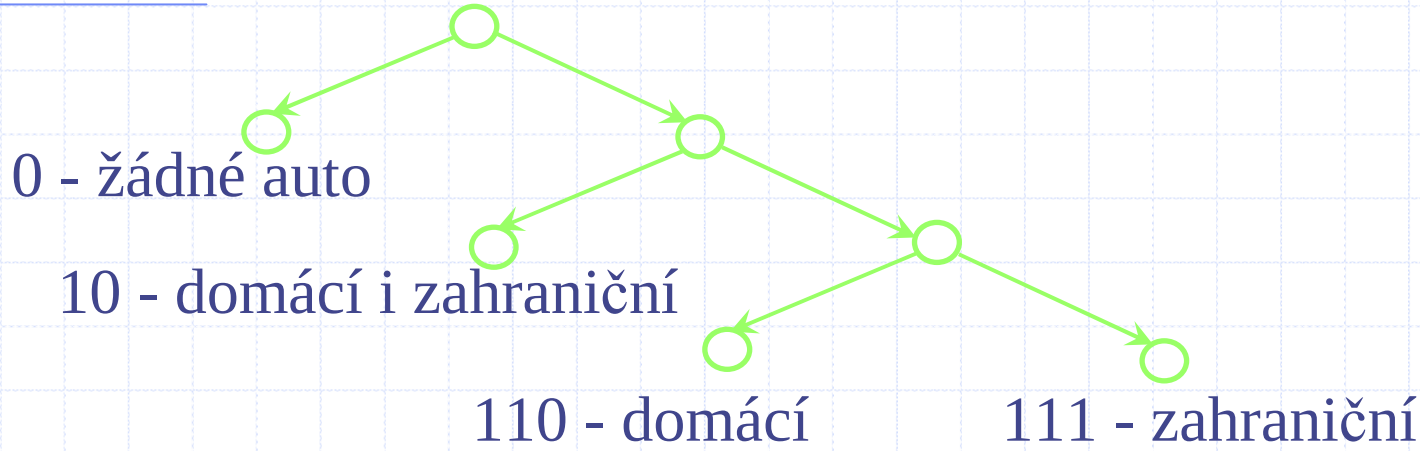
- TEORIE KÓDOVÁNÍ: 0 - žádné auto, 1 - domácí, 2 - zahraniční
  3 - domácí a zahraniční
- vysílání signálů na křižovatce podle dané situace
- při binárním kódování 0(**00**), 1(**01**), 2(**10**), 3(**11**)
- situace
  - stejně pravděpodobné *např*. (0.25)
  - nestejně pravděpodobné

    např. **0** (0.5), **1** (0.125), **2** (0.125), **3** (0.25)

EFEKTIVNÍ KÓDOVÁNÍ: častější zprávy kratší kód

tedy: 0(0), 1(110), 2(111), 3(10)

- jednoznačně rozpoznat začátek a konec kódu



- „Kolik“ **informace** získáme, známe-li výsledek pokusu?
- „Jak velkou“ **nejistotu** přináší neznalost výsledku pokusu?

Axiomatická definice entropie

entropie - míra stupně neurčitosti pokusu X

výsledky pokusu	X_1	X_2	...	Xn
pravděpodobnosti	$p(X_1)$	$p(X_2)$	...	$p(X_n)$

$$\mathbf{H(X) =_{ozn.} \Phi_n(p_1, p_2,\ldots,p_n)}$$

1. Hodnota fce $\Phi_n(p_1, p_2,\ldots,p_n)$ se nezmění při libovolné permutaci čísel $p_1, p_2,\ldots,p$

2. Fce $\Phi_2(p_1, p_2)$ je *spojitá*

3. $\Phi_n(p_1, p_2,\ldots,p_n) = \Phi_{n-1}(p_1+p_2,\ldots,p_n) + (p_1+p_2)\Phi_2(p_1/p_1+p_2, p_2/p_1+p_2)$

4. $\Phi_n(1/n,1/n,\ldots,1/n) = f(n)$ s rostoucím n roste

ad vlastnost č. 3

- $n=3,\ H(X) = \Phi(p_1,p_2,p_3)$

I. X_1, X_2 ⟹ $X \Longrightarrow Y$,

II. X_3

- $n=2,\ p(Y_1) = p_1 + p_2\ ,\ p(X_3) = p_3$

$H(Y) = \Phi(p_1+p_2,p_3)$

$H(X) \geq H(Y)$

$Y \Longrightarrow Y'$,

- $n=2,\ p(X_1) = p_1/(p_1+ p_2)$,

$p(X_2) = p_2/(p_1+ p_2)$

ad vlastnost č.3

$$H(Y') = \Phi(p_1/(p_1+p_2),\ p_2/(p_1+p_2))$$

$$\mathbf{H(X) = H(Y) + (p_1+p_2)\,H(Y')}$$

$$\mathbf{\Phi(p_1,p_2,p_3) = \Phi(p_1+p_2,p_3) + (p_1+p_2)\Phi(p_1/(p_1+p_2),\ p_2/(p_1+p_2))}$$

Axiomatická definice entropie *(pokračování)*

Jediná funkce, která splňuje podmínky 1.- 4., má tvar:
(bez důkazu)

$$\Phi_n(p_1, p_2,\ldots,p_n) = c(-p_1 \log p_1 - p_2 \log p_2 - \ldots - p_n \log p_n)$$

($c \log_a p = \log_b p$, kde $b^c = a$)

Entropie

- X - diskrétní náhodná veličina

  $\mathbf{H(X) = -\Sigma_{x \in F}\, p(x) \log_2 p(x)} \quad \mathbf{(H(X) \approx H(p))}$

- entropie vs kódování
  - *entropie je dolní mez průměrného počtu bitů potřebných k zakódování zprávy*
  - *entropie jako míra nejistoty obsahu zprávy (s délkou kódu nejistota roste)*

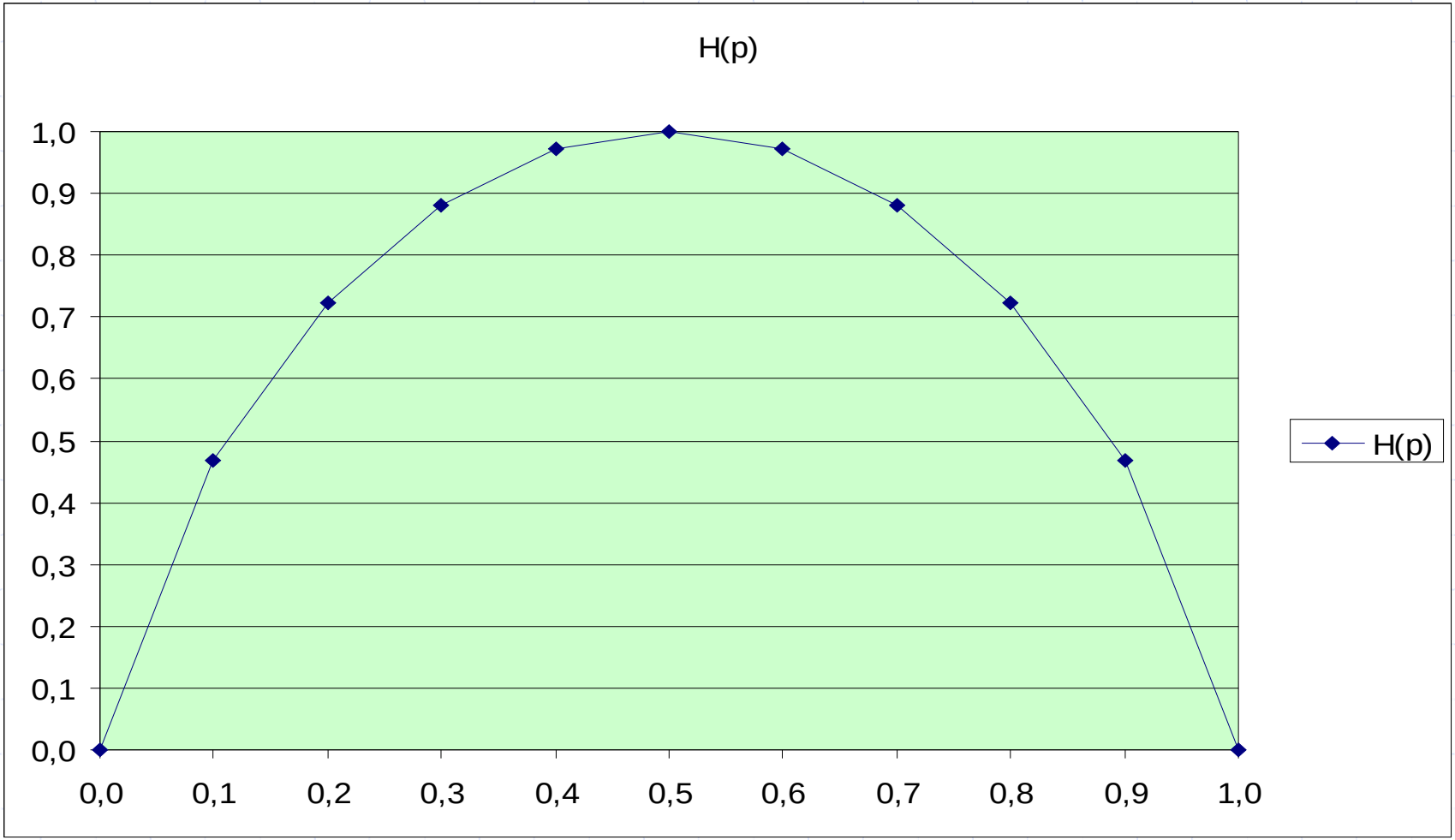
Vlastnosti entropie

- $H(X) \geq 0$
- $H_b(X) = (\log_b a)H(X)$
- p,q
  - $-\sum_{x \in F} p(x)\log_2 p(x) \leq -\sum_{x \in F} p(x)\log_2 q(x)$

(Jensenova nerovnost)

H(p) vs p

- *X = 1 s pravděpodobností **p**,*
- *X = 0 s pravděpodobností **1-p***



Shannonova hra

“nápodoba českého textu”

česká abeceda - *42* písmen(bez rozlišení ú a ů, plus mezera)

A. urna **1** se 42 lístečky - vybírání a vkládání zpět

“ďj mrgučxýďyaýweaožá”

B. urna **2** - lístečky podle četností písmen

“žia ep atndi zéuořmp”

C. urny 1-42 - 42 uren s dvojicemi písmen (c^i,c^j), počty dle $p(c^i/c^j)$

“lí di oneprá sguluvicechupsv”

Shannonova hra - výsledky

	H_A	H_B	H_C
čestina	5,39	4,67	3,87
ruština	5	4,35	3,52
angličtina	4,76	4,03	3,32
němčina	4,76	4,10	

Složená a podmíněná entropie

- H(X,Y) – množství informace pro předpovídání výsledků obou pokusů zároveň

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

- $H(Y/X) = \Sigma_{x \in F} p(x) H(Y/X = x)$
  $= -\Sigma_{x \in F} p(x) \Sigma_{Y \in G} p(y/x) \log p(y/x)$
  $= -\Sigma_{x \in F} \Sigma_{y \in G} p(x) p(x/y) \log p(y/x)$
  $= -\Sigma_{x \in F} \Sigma_{y \in G} p(x,y) \log p(y/x)$

- $H(X) \geq H(X/Y)$, $H(X) + H(Y) \geq H(X,Y)$

Chain rule

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

  $= - \sum_{x \in F} \sum_{y \in G} p(x,y) \log p(x)p(y/x)$

  $= - \sum_{x \in F} \sum_{y \in G} p(x,y) \log p(x) - \sum_{x \in F} \sum_{y \in G} p(x,y) \log p(y/x)$

  $= - \sum_{x \in F} p(x) \log p(x) - \sum_{x \in F} \sum_{y \in G} p(x,y) \log p(y/x)$

  $= H(X) + H(Y/X)$

- $H(X,Y/Z) = H(X/Z) + H(Y/X,Z)$

- $H(Y/X) \neq H(X/Y)$ ačkoli

  $H(X) - H(X/Y) = H(Y) - H(Y/X)$

Křížová entropie

- "správný" model známe/neznáme????
- aproximace - jak kvalitní? $\Rightarrow$

$\Rightarrow$ **Křížová entropie**

$$H(p,q) =_{def} -\Sigma_{x \in F} p(x) \log q(x)$$

$\Rightarrow$ **Křížová entropie na slovo**

$$(1/n)H(X) =_{def} -(1/n) \Sigma_{x \in F} p(x) \log q(x)$$

$\Rightarrow$ **Křížová entropie jazyka**

$$H(L, q) = \lim_{n \to \infty} (1/n) \Sigma_{x \in F} p(x) \log q(x)$$

Relativní entropie, vzájemná informace, perplexita

- **Relativní entropie**

(Kullback-Leibler vzdálenost)

$0 \leq \Sigma_{x \in F} p(x) \log_2 p(x) - \Sigma_{x \in F} p(x) \log_2 q(x) = H(p,q) - H(p)$

$\Sigma_{x \in F} p(x) \log(p(x)/q(x)) =_{def} D(p||q)$

- **Vzájemná informace**

$I(X;Y) = \Sigma_{x \in F} \Sigma_{y \in G} p(x,y) \log(p(x,y)/p(x)p(y)) =$

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

- **Perplexita**

$Perp(X) = 2^{H(X)}$

Relativní entropie *(pokračování)*

$m(X,Y)$

1. $m(X,Y) \geq 0,\ m(X,Y) = 0 \Leftrightarrow X = Y$

2. $m(X,Y) = m(Y,X)$

3. $m(X,Y) \leq m(X,Z) + m(Z,Y)$

D(p||q) ... splňuje 1., ale nesplňuje 2. a 3.

např.

p(1) = 1/4, p(2) = 3/4, r(1) = r(2) = 1/2, q(1) = 3/4, q(2) = 1/4

Proto **lépe**: $d(p,q) = (\Sigma_x(p(x) - q(x))^2)^{1/2}$

Perplexita - *příklad*

Předpověď dalšího slova w_t na základě *t-1* předchozích slov

$w_1 w_2 \ldots w_{t-1}$

$$H(w_t^i / w_1 w_2 \ldots w_{t-1}) =$$

$$= -\Sigma_{i=1.N} P(w_t^i / w_1 w_2 \ldots w_{t-1}) \log_2 P(w_t^i / w_1 w_2 \ldots w_{t-1})$$

předpoklad: $P(w_t^i / w_1 w_2 \ldots w_{t-1}) = 1/N$

$$H(w_t^i / w_1 w_2 \ldots w_{t-1}) = -\Sigma_{i=1.N} 1/N \log_2 1/N = \log_2 N$$

$$Perp(w_t^i / w_1 w_2 \ldots w_{t-1}) = N$$

Vzájemná informace vs entropie

- $I(X;Y) = \Sigma_{x,y} p(x,y) \log (p(x,y)/p(x)p(y))$

  $= \Sigma_{x,y} p(x,y) \log (p(x/y)/p(x))$

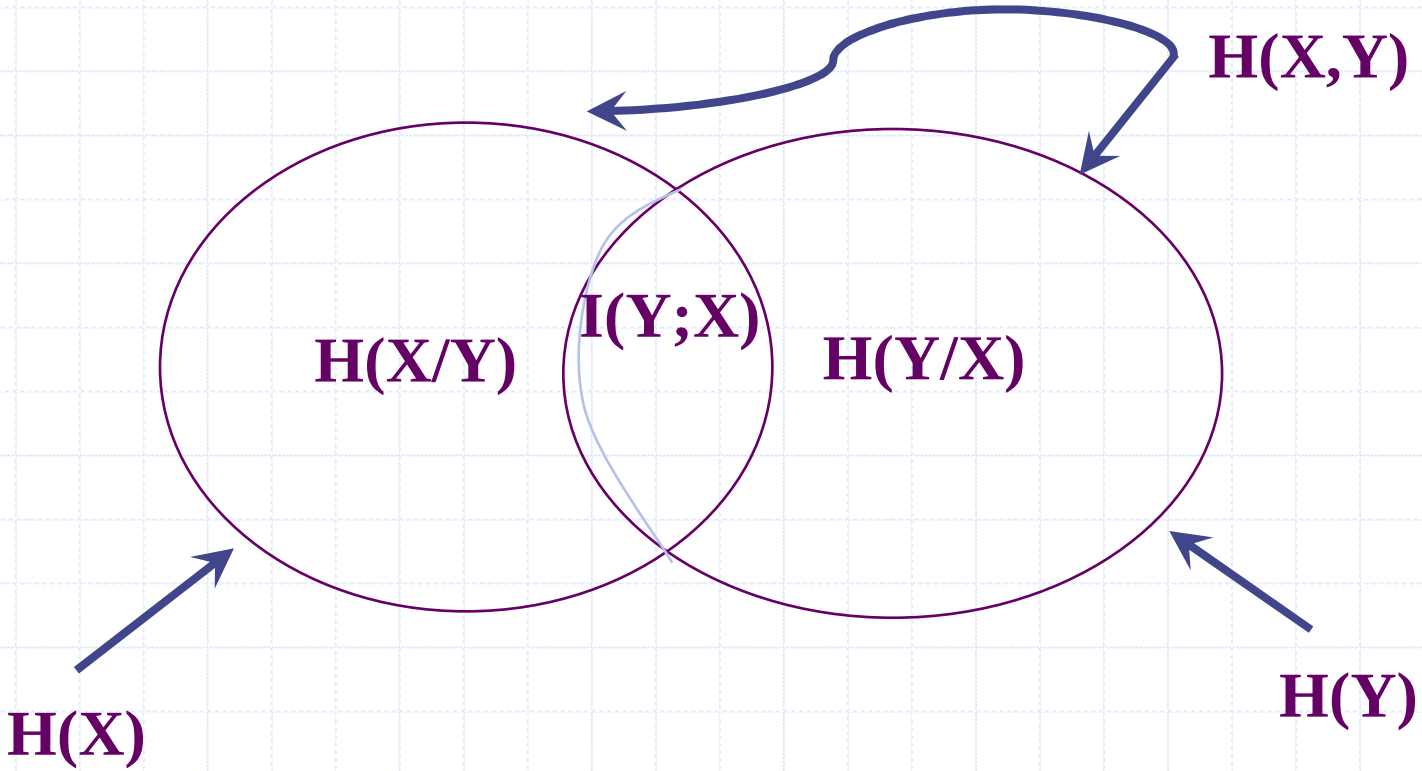
  $= - \Sigma_{x,y} p(x,y) \log p(x) + \Sigma_{x,y} p(x,y) \log p(x/y)$

  $= - \Sigma_{x} p(x) \log p(x) - (- \Sigma_{x,y} p(x,y) \log p(x/y))$

  $= H(X) - H(Y/X)$

- $I(X;Y) = H(Y) - H(X/Y)$
- $I(X;Y) = H(X) + H(Y) - H(Y/X)$
- $I(X;X) = H(X) - H(X/X) = H(X)$

Diagram vzájemná informace vs entropie



Chain rule *(pokračování)*

- $\mathbf{H(X_1, X_2,\ldots,X_n) = \Sigma_{i=1..n}\, H(X_i/X_{i-1}, \ldots,X_1)}$

- $\mathbf{I(X_1, X_2,\ldots,X_n;Y) = \Sigma_{i=1..n}\, I(X_i;Y/X_{i-1}, \ldots,X_1)}$

$$I(X_1, X_2,\ldots,X_n;Y) = H(X_1, X_2,\ldots,X_n) - H(X_1, X_2,\ldots,X_n/Y)$$
$$= \Sigma_{i=1..n}\, H(X_i/X_{i-1}, \ldots,X_1) - \Sigma_{i=1..n}\, H(X_i/X_{i-1}, \ldots,X_1,Y)$$
$$= \Sigma_{i=1..n}\, I(X_i;Y/X_{i-1}, \ldots,X_1)$$

- $\mathbf{D(p(x,y)\,||\,q(x,y)) = D(p(x)\,||\,q(x)) + D(p(y/x)\,||\,q(y/x))}$