Síla a významnost asociace mezi proměnnými v systému

Program

1. Entropie jako míra neuspořádanosti.
2. Entropie jako míra informace.
3. Entropie na rozkladu množiny elementárních jevů.
4. Vlastnosti entropie.
5. Podmíněná entropie.
6. Vlastnosti podmíněné entropie.
7. Vzájemná informace.
8. Optimální histogramování.
9. Výpočet entropie ze vzorku dat.
10. Kvalita asociace: kontingenční test.

Papoulis, A. *Probability, Random Variables, and Stochastic Processes*. McGraw Hill 1991, kap. 15.

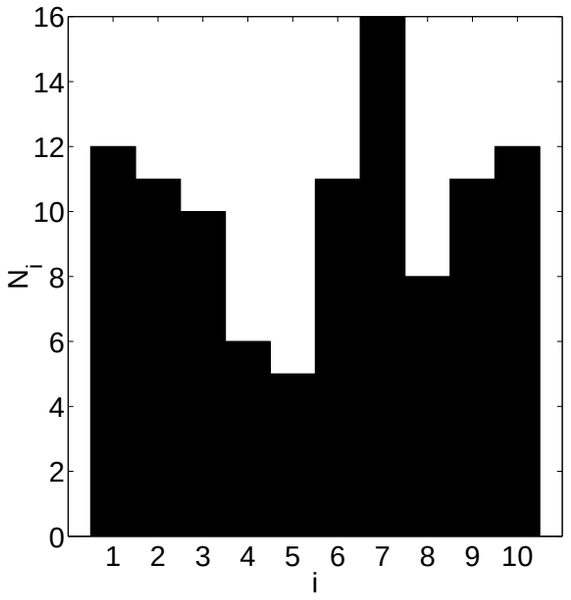
Anděl, J. *Statistické metody*. Praha: Matfyzpress, 1998; str. 157-167.

Duda, RO. – Hart, PE. – Stork, DG. *Pattern Classification*, 2nd ed. John Wiley & Sons, 2001; část 9.4.1.

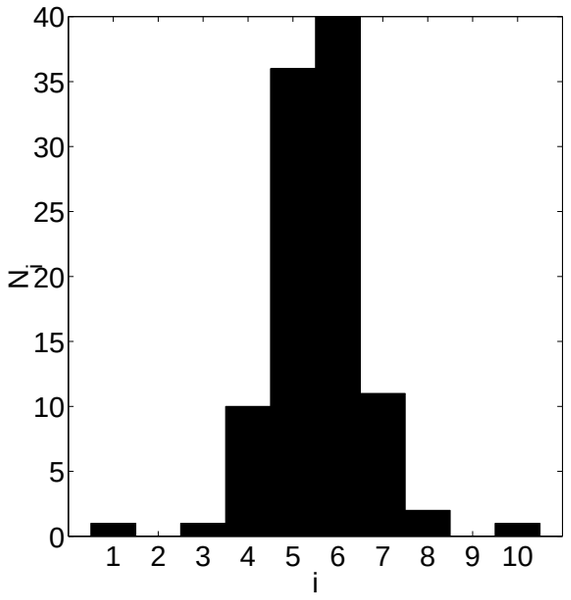
Entropie jako míra neuspořádanosti

Jev x_i: kulička padla do přihrádky i

$H = 3.2554$



$H = 2.0424$



$N = 100$ stejných objektů: $\{b_1, b_2, \ldots b_N\}$

$N!$ všech permutací

$N_i!$ permutací v každé přihrádce, $i = 1, 2, \ldots, m$

Celkový počet přerovnání objektů v histogramu je $W = \dfrac{N!}{\prod_{i=1}^{m} N_i!}$

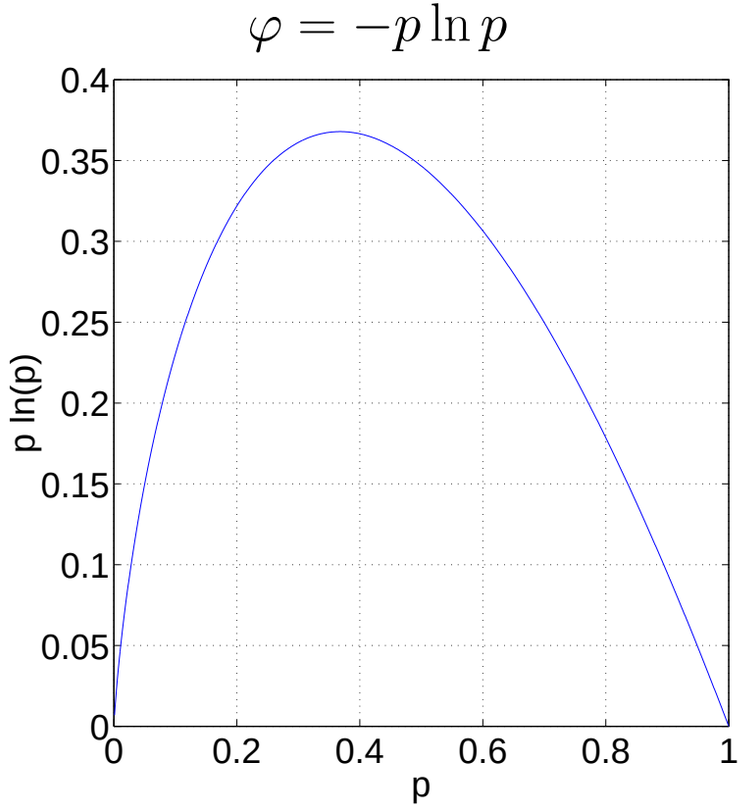
pokračování

Entropie

$$\lim_{N\to\infty} \frac{1}{N} \ln W = \cdots = \lim_{N\to\infty} -\sum_{i=1}^{m} \frac{N_i}{N} \ln \frac{N_i}{N} = -\sum_{i=1}^{m} p_i \ln p_i$$

Počet mikrostavů (rozdělení do m přihrádek) které dávají za vznik stejnému makrostavu (histogram).

- Entropie je malá pro úzká rozdělení
- Entropie je velká pro široká rozdělení



Entropie jako míra informace

Hledáme funkci $s(p)$:

1. spojitou
2. monotonně klesající s p
3. $s(1) = 0$ nulová nejistota
4. $s(p_A \cdot p_B) = s(p_A) + s(p_B)$

Hledaná funkce

$$s(p) = -s(\tfrac{1}{e}) \ln p, \qquad s(\tfrac{1}{e}) = 1 \quad \text{(definujeme)}$$

Střední míra informace (v Natech) na množině jevů $\mathbf{x} = \{x_k\}$

$$H(\mathbf{x}) = -\sum_k p(x_k) \ln p(x_k)$$

Vlastnosti rozkladu množiny elementárních jevů na třídy ekvivalence

Rozklad množiny elementárních jevů: $\{A_1, A_2, \ldots, A_n\}$

Definice

1. Rozklad je disjunktní pokrytí MEJ
2. Elementární rozklad $\mathcal{E} = \{\{e_1\}, \{e_2\}, \ldots, \{e_n\}\}$
3. Zjemnění rozkladu: Dány $\mathcal{A} = \{A_1, A_2, \ldots A_n\}, \quad \mathcal{B} = \{B_1, B_2, \ldots, B_m\}$, pak

$$\mathcal{B} \preceq \mathcal{A} \quad \text{právě když} \quad \forall i \, \exists! j : B_i \subseteq A_j \quad \text{(právě jedno } j\text{)}$$

4. Součin rozkladů $\mathcal{C} = \mathcal{A} \cdot \mathcal{B} = \{A_i \cap B_j, \, \forall i, j\}$ je největší společné zjemnění

$$\mathcal{A} \cdot \mathcal{B} \preceq \mathcal{A}, \quad \mathcal{A} \cdot \mathcal{B} \preceq \mathcal{B}$$

Vlastnosti

1. $\mathcal{E} \preceq \mathcal{A}$ pro každé $\mathcal{A}$
2. $\mathcal{A} \cdot \mathcal{B} = \mathcal{B} \cdot \mathcal{A}$ (komutativita)
3. $\mathcal{A} \cdot (\mathcal{B} \cdot \mathcal{C}) = (\mathcal{A} \cdot \mathcal{B}) \cdot \mathcal{C}$ (asociativita)
4. Jestliže $\mathcal{A}_1 \preceq \mathcal{A}_2$ a $\mathcal{A}_2 \preceq \mathcal{A}_3$ potom $\mathcal{A}_1 \preceq \mathcal{A}_3$ (tranzitivita)
5. Jestliže $\mathcal{B} \preceq \mathcal{A}$ potom $\mathcal{A} \cdot \mathcal{B} = \mathcal{B}$

Entropie na rozkladu množiny elementárních jevů

Rozklad MEJ $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$, zavedeme $P(A_i) \stackrel{\text{def}}{=} p_i$ (pouhá notace)

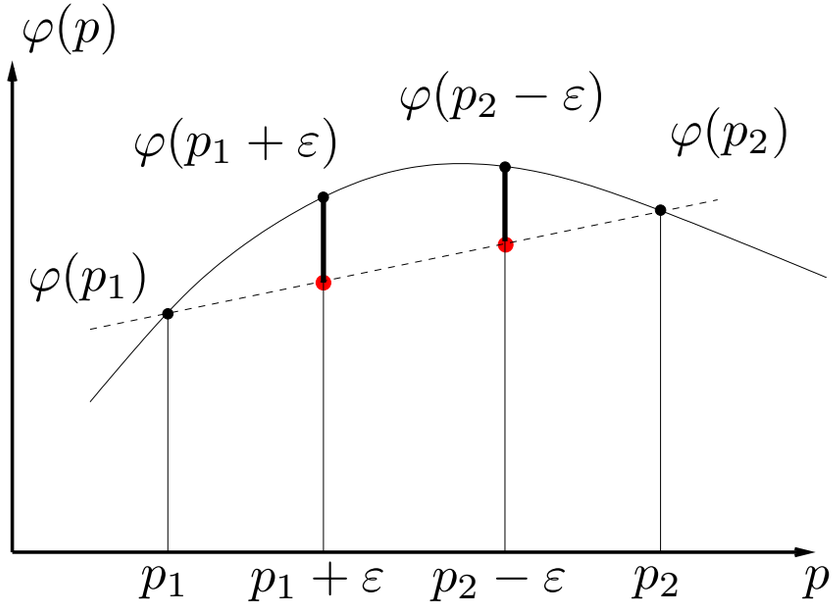
Entropie

$$H(\mathcal{A}) \stackrel{\text{def}}{=} -\sum_{i=1}^{n} p_i \ln p_i = \sum_{i=1}^{n} \varphi(p_i), \qquad \varphi(p_i) \stackrel{\text{def}}{=} -p_i \ln p_i$$

Pomocná věta:

$$\varphi(p_1 + p_2) \leq \varphi(p_1) + \varphi(p_2) \leq \varphi(p_1 + \varepsilon) + \varphi(p_2 - \varepsilon)$$

pokud $p_1 < p_1 + \varepsilon \leq p_2 - \varepsilon < p_2$



Vlastnosti

V1. Jestliže $\mathcal{B} \preceq \mathcal{A}$ potom $H(\mathcal{B}) \geq H(\mathcal{A})$

Zjemněním rozkladu se entropie zvýší. Pozn: zjemněním histogramu se entropie zvýší.

V1⇒V2. Pro každý rozklad $\mathcal{A}$ platí $H(\mathcal{A}) \leq H(\mathcal{E})$

Entropie každého rozkladu je menší nebo rovna entropii elementárního rozkladu.

V1⇒V3. Pro každý rozklad $\mathcal{A}$ a $\mathcal{B}$ platí

$$H(\mathcal{A}) \leq H(\mathcal{A} \cdot \mathcal{B}), \quad H(\mathcal{B}) \leq H(\mathcal{A} \cdot \mathcal{B}),$$

Pozn: $\mathcal{A} \cdot \mathcal{B} \preceq \mathcal{A}, \quad \mathcal{A} \cdot \mathcal{B} \preceq \mathcal{B}$

V4. Entropie rozkladu $\mathcal{A}$ je maximální pokud všechny jeho prvky mají stejnou pravděpodobnost: $p_i = P(A_i) = p$

Příklad: Statický systém v parlamentu

- podmnožina pěti vybraných poslanců $\{v_1, \ldots, v_5\}$
- někteří v_i hlasují podle v_j, $j \neq i$, ale nevíme kteří
- někteří hlasují nezávisle, nevíme kteří
- máme záznam 100 hlasování

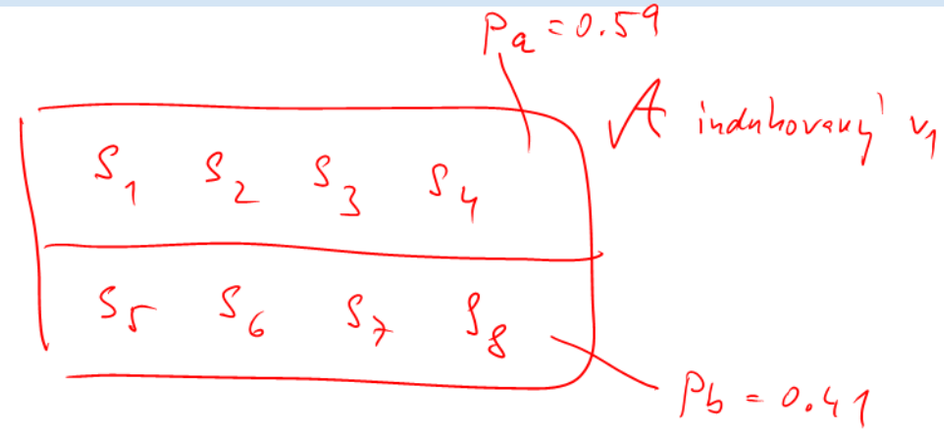
v_1	v_2	v_3	v_4	v_5	$p(\mathbf{s})$
0	0	0	0	0	0.02
0	0	0	0	1	0.04
0	0	0	1	0	0.02
0	0	0	1	1	0.05
0	0	1	0	0	0.03
0	0	1	0	1	0.03
0	0	1	1	0	0.09
0	0	1	1	1	0.03
0	1	0	0	0	0.03
0	1	0	0	1	0.05
0	1	0	1	0	0.04
0	1	0	1	1	0.01
0	1	1	0	0	0.03
0	1	1	0	1	0.03
0	1	1	1	0	0.04
0	1	1	1	1	0.05

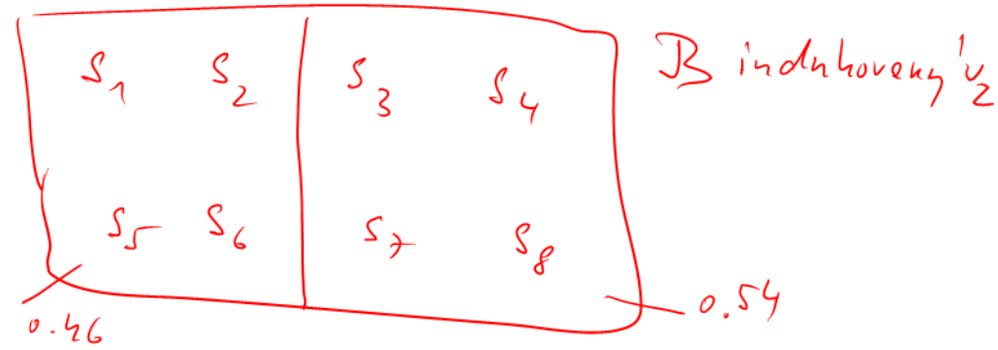
v_1	v_2	v_3	v_4	v_5	$p(\mathbf{s})$
1	0	0	0	0	0.01
1	0	0	0	1	0.03
1	0	0	1	0	0.03
1	0	0	1	1	0.03
1	0	1	0	0	0.01
1	0	1	0	1	0.03
1	0	1	1	1	0.01
1	1	0	0	0	0.03
1	1	0	0	1	0.03
1	1	0	1	0	0.03
1	1	0	1	1	0.04
1	1	1	0	0	0.01
1	1	1	0	1	0.04
1	1	1	1	0	0.03
1	1	1	1	1	0.05

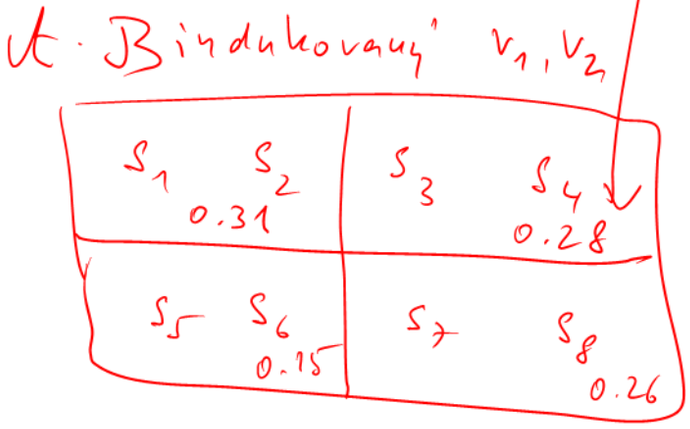
pokračování

	v_1	v_2	v_3	$p(v_1, v_2, v_3)$
s_1	0	0	0	0.13
s_2	0	0	1	0.18
s_3	0	1	0	0.13
s_4	0	1	1	0.15
s_5	1	0	0	0.10
s_6	1	0	1	0.05
s_7	1	1	0	0.13
s_8	1	1	1	0.13

p_a







$H(A) = -0.59 \ln 0.59 - 0.41 \ln 0.41 = 0.6769$ Nat

$H(E) = -0.13 \ln 0.13 - \ldots = 2.0342$

$H(B) = 0.6899$

$H(A \cdot B) = -0.31 \ln 0.31 - \ldots = 1.3543$

Podmíněná entropie

Příklad: Uvažujme sekvenci pokusů, v nichž nastal jev L (padla lichá při házení kostkou), v takové sekvenci je nejistota o elementárním rozkladu rovna $H(\mathcal{E} \mid L)$.

Uvažujme sekvenci, v níž nastal doplněk jevu S (sudá), v ní je nejistota o $\mathcal{E}$ rovna $H(\mathcal{E} \mid S)$.

Vážený součet $H(\mathcal{E} \mid L) \cdot P(L) + H(\mathcal{E} \mid S) \cdot P(S)$ je podmíněná entropie $\mathcal{E}$, pozorujeme-li rozklad $\{L, S\}$.

Def. Nechť rozklady $\mathcal{A}$ a $\mathcal{B}$ jsou $\mathcal{A} = \{A_1, A_2, \ldots, A_{N_A}\}, \quad \mathcal{B} = \{B_1, B_2, \ldots, B_{N_B}\}$.
Potom podmíněná entropie rozkladu $\mathcal{A}$ za předpokladu, že nastal jev B_j je

$$H(\mathcal{A} \mid B_j) = -\sum_{i=1}^{N_A} P(A_i \mid B_j) \ln P(A_i \mid B_j)$$

Podmíněná entropie $H(\mathcal{A} \mid \mathcal{B})$ je pak střední hodnota přes $\mathcal{B}$

$$H(\mathcal{A} \mid \mathcal{B}) = \sum_{j=1}^{N_B} P(B_j)\, H(\mathcal{A} \mid B_j) = -\sum_{j=1}^{N_B} P(B_j) \sum_{i=1}^{N_A} P(A_i \mid B_j)\, \ln P(A_i \mid B_j) =$$

$$= -\sum_i \sum_j P(A_i,\, B_j) \ln P(A_i \mid B_j)$$

$P(B_j) \cdot P(A_i \mid B_j)$

Pozn: Střední nejistota o $\mathcal{A}$ je-li pozorováno $\mathcal{B}$ je $H(\mathcal{A} \mid \mathcal{B})$.

Příklad

Parlament, $\mathbf{S} = \{v_1, v_2, v_3\}$, v_1 indukuje rozklad $\mathcal{A}$

	v_1	v_2	v_3	$p(\mathbf{S})$	$p(v_2, v_3 \mid v_1 = 0)$
s_1	0	0	0	0.13	0.2203
s_2	0	0	1	0.18	0.3051
s_3	0	1	0	0.13	0.2203
s_4	0	1	1	0.15	0.2542
					$p(v_2, v_3 \mid v_1 = 1)$
s_5	1	0	0	0.10	0.2439
s_6	1	0	1	0.05	0.1220
s_7	1	1	0	0.13	0.3171
s_8	1	1	1	0.13	0.3171

$$\frac{p(v_2, v_3, v_1 = 0)}{p(v_1 = 0)} = \frac{0.13}{0.59}$$

1.3769 Nat

$$H(v_2, v_3 \mid v_1) = 0.59 \cdot H(v_2, v_3 \mid v_2 = 0) +$$

$$0.41 \cdot H(v_2, v_3 \mid v_1 = 1) = 1.3573$$

1.3291

$$H(v_2, v_3) - H(v_2, v_3 \mid v_1)$$

$$H(v_2, v_3 \mid v_1 = 0) = -0.2203 \cdot \ln 0.2203 - \cdots = 1.3769 \text{ [Nat]}$$

$$H(v_2, v_3 \mid v_1 = 1) = \cdots = 1.3291 \text{ [Nat]}$$

$$H(v_2, v_3 \mid v_1) = 0.59 \cdot H(v_2, v_3 \mid v_1 = 0) + 0.41 \cdot H(v_2, v_3 \mid v_1 = 1) = 1.3573$$

Interpretace: Nemáme-li žádnou informaci, je naše nejistota o stavu systému $H(\mathcal{E}) = 2.0342$ Nat. Pokud víme, jak hlasude v_1, na3e nejistota klesne na $H(v_2, v_3 \mid v_1) = 1.3573$ Nat.

Vlastnosti

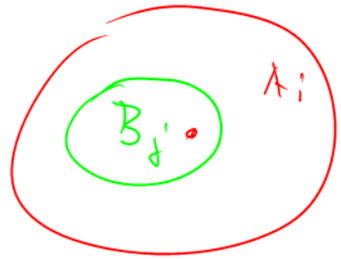
Pro každé rozklady $\mathcal{A}$ a $\mathcal{B}$ platí:

V5. Jestliže $\mathcal{B} \preceq \mathcal{A}$ potom $H(\mathcal{A} \mid \mathcal{B}) = 0$.

Intuice: víme, které jevy v $\mathcal{A}$ nastaly.

$\forall j \; \exists! \, i \qquad B_j \subseteq A_i$

$$P(A_i \mid B_j) = \frac{P(A_i, B_j)}{P(B_j)} = \begin{cases} 1 \\ 0 \end{cases}$$



Def. Rozklady $\mathcal{A}$ a $\mathcal{B}$ nezávislé $\Leftrightarrow \forall A_i \in \mathcal{A}, \, B_j \in \mathcal{B}\colon P(A_i, \, B_j) = P(A_i) \cdot P(B_j)$.

V6. Jsou-li $\mathcal{A}$ a $\mathcal{B}$ nezávislé, potom $H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{A}), \qquad H(\mathcal{B} \mid \mathcal{A}) = H(\mathcal{B})$

Intuice: Jsou-li jevy nezávislé, potom pozorováním $\mathcal{B}$ nezískáme žádnou informaci o $\mathcal{A}$. ⊛1

V7. $H(\mathcal{A} \cdot \mathcal{B}) \leq H(\mathcal{A}) + H(\mathcal{B})$

↖ ne nutně nezávislé

Intuice: entropie složeného jevu není větší než entropie dílčích jevů.

V8. Jsou-li $\mathcal{A}$ a $\mathcal{B}$ nezávislé, potom $H(\mathcal{A} \cdot \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B})$ ⊛1

Vlastnosti

$$H(\mathcal{A}\mid\mathcal{B}) = H(\mathcal{A}\cdot\mathcal{B}) - H(\mathcal{B})$$

V9. $H(\mathcal{A}\cdot\mathcal{B}) = H(\mathcal{B}) + H(\mathcal{A}\mid\mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}\mid\mathcal{A})$

Intuice: Pozorujeme-li $\mathcal{B}$, získáme o $\mathcal{A}\cdot\mathcal{B}$ informaci $H(\mathcal{B})$, takže zbytková nejistota je $H(\mathcal{A}\mid\mathcal{B}) = H(\mathcal{A}\cdot\mathcal{B}) - H(\mathcal{B})$.

$$P(B_j)\,H(\mathcal{A}\mid B_j) = -\sum_i P(A_i,B_j)\ln\frac{P(A_i,B_j)}{P(B_j)}$$

$$H(\mathcal{A}\mid\mathcal{B}) = -\sum_j\sum_i P(A_i,B_j)\ln P(A_i,B_j) + \sum_j\sum_i P(A_i,B_j)\ln P(B_j)$$

$H(\mathcal{A}\cdot\mathcal{B})$ $\qquad$ $-H(\mathcal{B})$

2 TRIKY

$$\sum_i P(A_i,B_j) = P(B_j)$$

$$\sum_j P(B_j)\ln P(B_j)$$

$$\sum_j \ln P(B_j)\left(\sum_i P(A_i,B_j)\right)$$

$P(B_j)$

Vlastnosti

$I(\mathcal{A},\mathcal{B}) = I(\mathcal{B},\mathcal{A})$

V10. $H(\mathcal{A}) - H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{B}) - H(\mathcal{B} \mid \mathcal{A})$ (plyne z V9)

V11. $H(\mathcal{B}) \leq H(\mathcal{A} \cdot \mathcal{B}) \leq H(\mathcal{A}) + H(\mathcal{B})$ (plyne z V9+V7)

V12. Pro každé $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ platí: (důsledek vlastnosti V9)

$$H(\mathcal{B} \cdot \mathcal{C} \mid \mathcal{A}) = H(\mathcal{B} \mid \mathcal{A}) + H(\mathcal{C} \mid \mathcal{A} \cdot \mathcal{B}) = H(\mathcal{C} \mid \mathcal{A}) + H(\mathcal{B} \mid \mathcal{C} \cdot \mathcal{A})$$

V13. $0 \leq H(\mathcal{A} \mid \mathcal{B}) \leq H(\mathcal{A})$ (plyne z V9+V11)

Intuice:

1. Pozorováním $\mathcal{B}$ se nejistota o $\mathcal{A}$ nemůže zvětšit.
2. Entropie podmnožiny rozkladu je menší.

V14. Jestliže $\mathcal{B} \preceq \mathcal{C}$ potom $H(\mathcal{A} \mid \mathcal{B}) \leq H(\mathcal{A} \mid \mathcal{C})$. (plyne z V7 a vlastnosti 5 rozkladu MEJ.)

Intuice: Jemnějším rozkladem se dozvíme více o $\mathcal{A}$. ⊛2

Vzájemná informace

Pozorování rozkladu $\mathcal{B}$ redukuje nejistotu o $\mathcal{A}$ z $H(\mathcal{A})$ na $H(\mathcal{A} \mid \mathcal{B})$ a získá tedy o $\mathcal{A}$ informaci $I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) - H(\mathcal{A} \mid \mathcal{B})$. srv. V9, kde jsme získávali inf. o $\mathcal{A} \cdot \mathcal{B}$

Def.

$$I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A} \cdot \mathcal{B})$$ plyne z V9

Vlastnosti

neplatí $H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{B} \mid \mathcal{A})$!

$$I(\mathcal{A}, \mathcal{B}) = I(\mathcal{B}, \mathcal{A})$$ symetrie

$$I(\mathcal{A}, \mathcal{B}) \geq 0$$ nonnegativita, plyne z V7

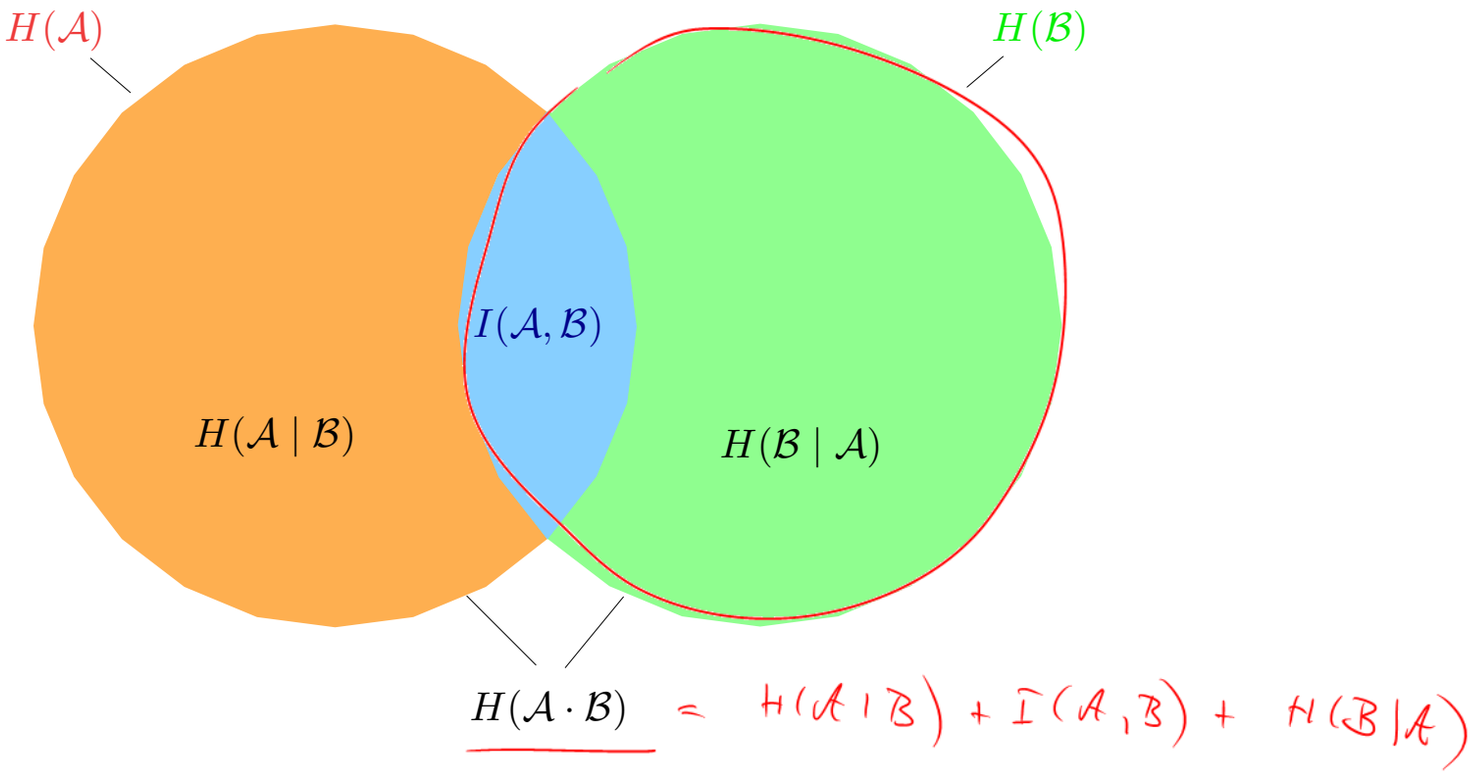
$I(\mathcal{A}, \mathcal{B})$

1. Informace o $\mathcal{A}$ obsažená v $\mathcal{B}$
2. Informace o $\mathcal{B}$ obsažená v $\mathcal{A}$

Pro více rozkladů:

$$I(\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_k) = \sum_{i=1}^{k} H(\mathcal{A}_i) - H(\mathcal{A}_1 \cdot \mathcal{A}_2 \cdots \mathcal{A}_k)$$

Mnemotechnická pomůcka: Vztah mezi podmíněnými druhy entropie



Čteme takto: platí aditivita plochy, příčemž levý kruh představuje $H(\mathcal{A})$, pravý kruh $H(\mathcal{B})$ a jejich sjednocení $H(\mathcal{A}\cdot\mathcal{B})$. Potom

$$H(\mathcal{A}) + H(\mathcal{B}\mid\mathcal{A}) = H(\mathcal{A}\cdot\mathcal{B})$$
$$H(\mathcal{A}\cdot\mathcal{B}) - I(\mathcal{A}\cdot\mathcal{B}) = H(\mathcal{A}\mid\mathcal{B}) + H(\mathcal{B}\mid\mathcal{A})$$
$$\dots$$

Entropie náhodné proměnné

Nechť X je diskrétní náhodná proměnná nabývající hodnot $x_i \in R(X)$ s pravděpodobnostmi

$$P(X = x_i) \stackrel{\text{def}}{=} p_i$$

range

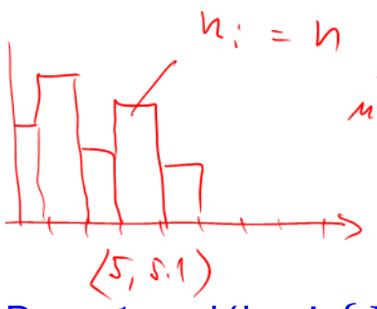
Sjednocení událostí $\{X = x_i\}$ tvoří rozklad $\mathcal{A}_X$ (tj. pokrytí oboru hodnot $R(X)$).

Def. 1 Entropie $H(X)$ diskrétní náhodné proměnné X je rovna

$$H(X) = H(\mathcal{A}_X) = -\sum_i p_i \ln p_i = -\sum_{x_i \in R(X)} P(X = x_i) \ln P(X = x_i)$$

$H(\mathcal{A}_X) \leq H(\mathcal{E}_X)$

Def. 2 Diferenciální entropie $H(X)$ spojité náhodné proměnné X je



$$n_i = n \int_{u \in \Delta} f(u)\,du \approx n \cdot p(\tilde{u}_i) \cdot \Delta$$

$$H(X) \stackrel{\text{def}}{=} -\int_{-\infty}^{\infty} f(x) \ln f(x)\,dx$$

$$\lim_{\Delta \to 0} \sum_i p(\tilde{u}_i) \cdot \Delta \ln p(\tilde{u}_i) \cdot \Delta = \lim_{\Delta \to 0} \sum_i p(\tilde{u}_i) \cdot \Delta \ln p(\tilde{u}_i) + \sum_i p(\tilde{u}_i) \cdot \Delta \cdot \ln \Delta$$

$\Delta \to dx$

$$\ln \Delta \int f(x)\,dx = \ln \Delta$$

Pozn 1: události $\{X = x_i\}$ spojité X netvoří rozklad, jsou nespočetné

Pozn 2: pro spojitou X je $H(X) \in (-\infty, \infty)$

Pozn 3: rovnoměrné rozdělení na intervalu $\langle 0, a \rangle$: $H(x) = \ln a$, normální rozdělení: $H(x) = \ln \sigma\sqrt{2\pi e}$

⊛2 entropie multidimenzionálního normálního rozdělení

Sdružená a podmíněná entropie diskrétní náhodné proměnné

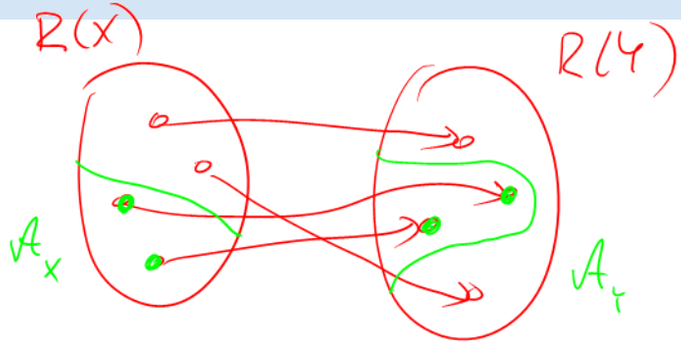
Sdružená entropie

$$H(X,Y) = H(\mathcal{A}_X \cdot \mathcal{A}_Y) = -\sum_i \sum_j p_{ij} \ln p_{ij}$$

Podmíněná entropie

$$H(X \mid Y) = H(\mathcal{A}_X \mid \mathcal{A}_Y)$$

Věta o bijekci



Jestliže $y = f(x)$ je zobrazení prosté a na (bijekce), potom

V15. $H(Y \mid X) = H(X \mid Y) = 0$

D: Rozklady $\mathcal{A}_X$ a $\mathcal{A}_Y$ jsou ekvivalentní, $\mathcal{A}_X = \mathcal{A}_Y$. Takže $\mathcal{A}_X \preceq \mathcal{A}_Y$ a platí V5.

V16. $I(X,Y) = H(Y), \quad I(X,Y) = H(X).$ Také $H(X,Y) = H(X) = H(Y).$ (z V9)

Aplikace 1: Který poslanec má největší vliv na výsledek hlasování?

Systém: $\mathbf{S} = \{v_1, v_2, v_3, v_4, v_5\}$

Pozorování:

$$H_1 = H(v_2, v_3, v_4, v_5 \mid v_1) \text{ tím menší, čím větší je vliv } v_1$$
$$= H(v_1, v_2, v_3, v_4, v_5) - H(v_1) \qquad \text{(z V9)}$$

Výsledky:

$\{v_1, v_2, v_3, v_4, v_5\}$

$H_i = H(S) - H(v_i)$

v_i	H_i
v_1	2.6423
v_2	2.6292
v_3	2.6262
v_4	2.6310
v_5	2.6310

Pozn: $H(v_1, v_2, v_3, v_4, v_5) = 3.3191$ [Nat]