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.

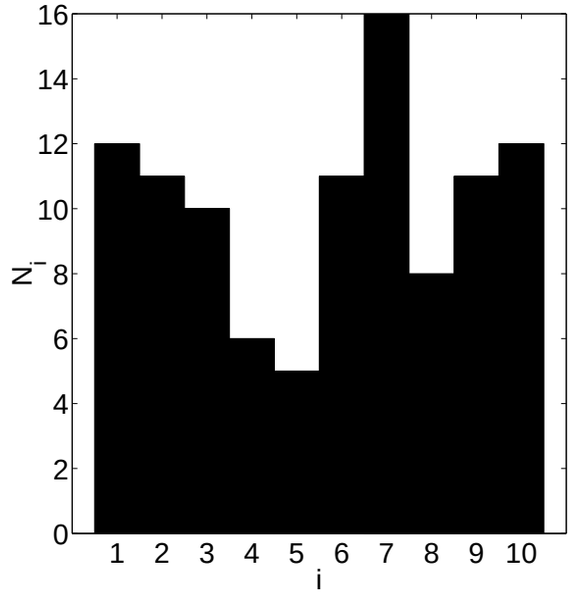
Anděl, J. *Statistické metody*. Praha: Matfyzpress, 1998; str. 157-167.

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

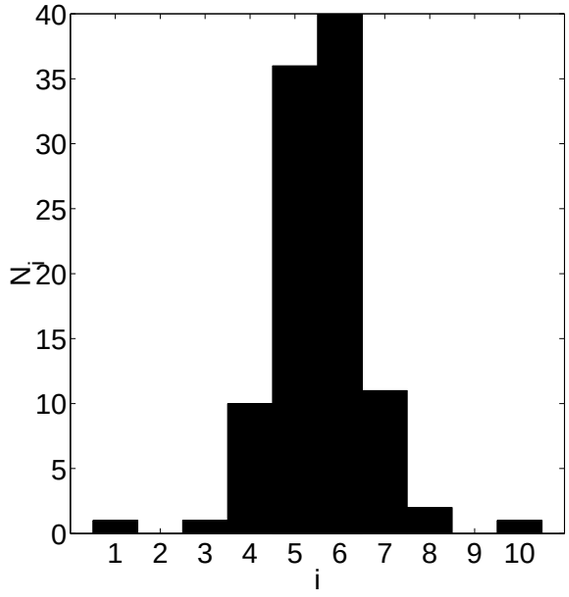
Entropie jako míra neuspořádanosti

Jev x_i: kulička padla do přihrádky (‚třídy‘) 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, $i = 1, 2, \ldots, m$ (zde $m = 10$)

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

pokračování

$p_i \stackrel{\text{def}}{=} P(\text{kulička padla do přihrádky } i)$

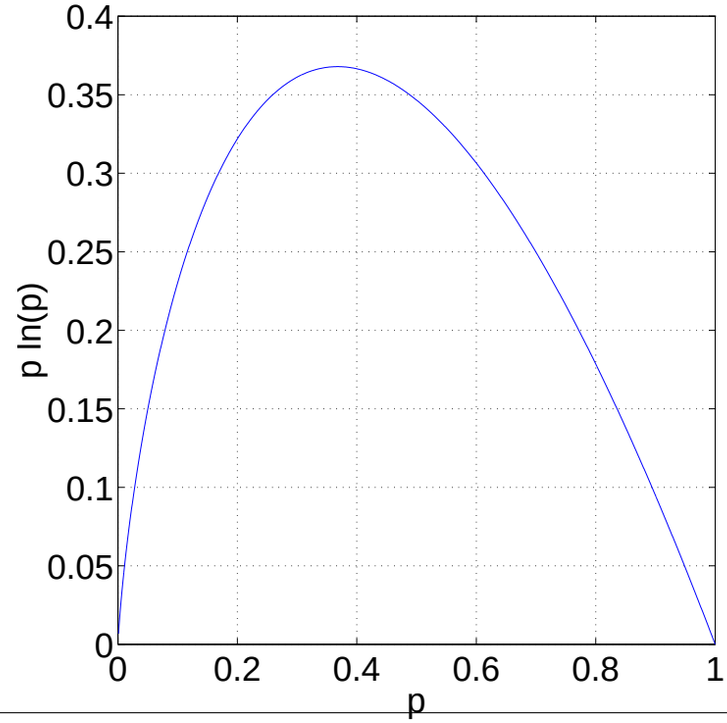
Entropie

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

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

$$\varphi(p) = -p \ln p$$



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 $\mathbf{x} = \{x_k\}_{k=1}^{K}$

$$H(\mathbf{x}) = -\sum_{k=1}^{K} p(x_k) \ln p(x_k)$$

[Shannon 1948]

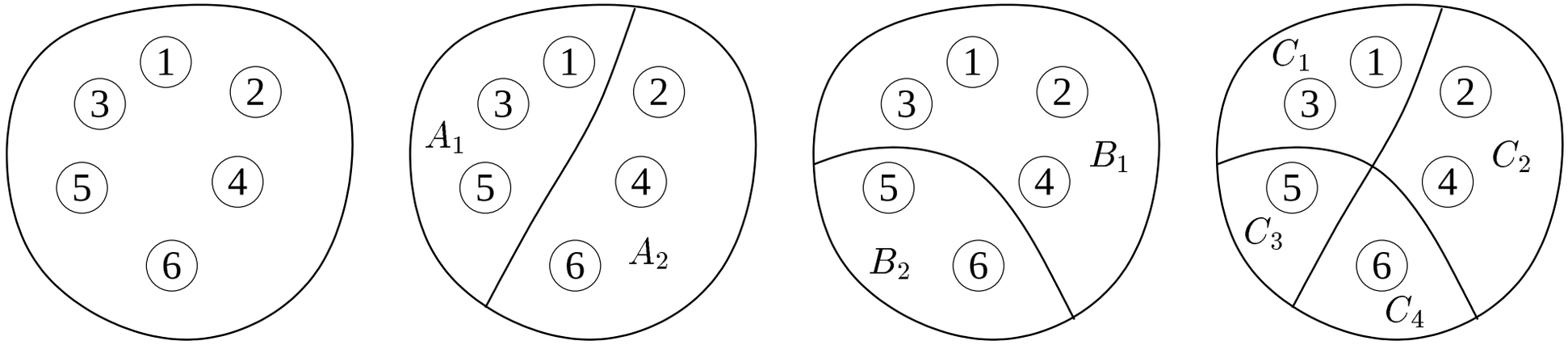
Vlastnosti rozkladu množiny elementárních jevů na třídy ekvivalence

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

Definice

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

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



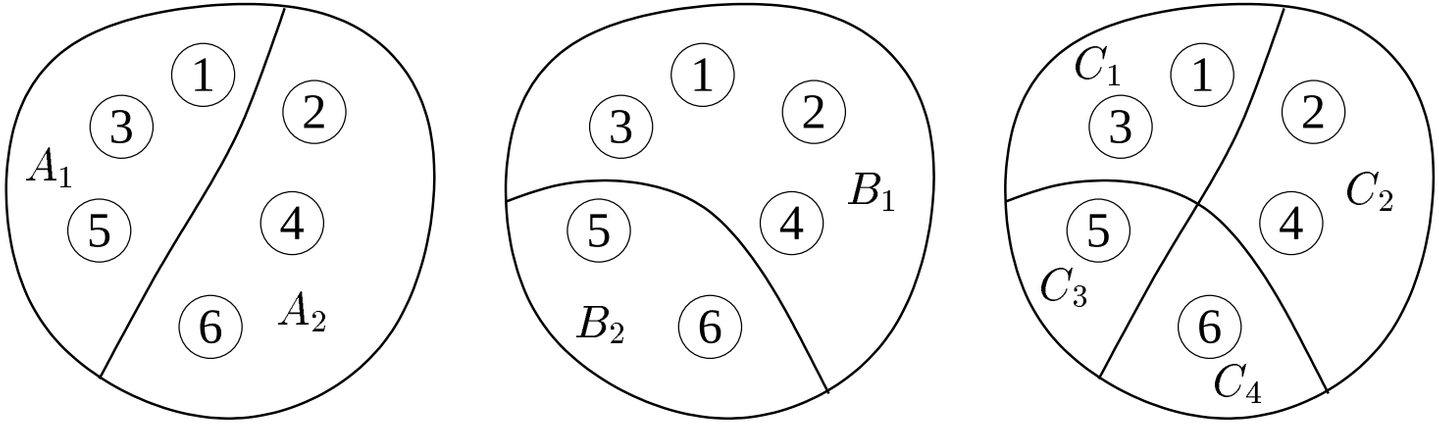
$\mathcal{E}$ $\quad \mathcal{A} = \{A_1, A_2\}$ $\quad \mathcal{B} = \{B_1, B_2\}$ $\quad \mathcal{C} \preceq \mathcal{B}, \mathcal{C} \preceq \mathcal{A}$

$\mathcal{C} \preceq \mathcal{B}$ protože $C_1 \subseteq B_1$, $C_2 \subseteq B_1$, $C_3 \subseteq B_2$, $C_4 \subseteq B_2$.

Součin rozkladů

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

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



$\mathcal{A}$ $\mathcal{B}$ $\mathcal{C} = \mathcal{A} \cdot \mathcal{B}$

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 $\mathcal{A} \cdot \mathcal{B} = \mathcal{B}$ (z toho plyne idempotence $\mathcal{A} \cdot \mathcal{A} = \mathcal{A}$)

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)

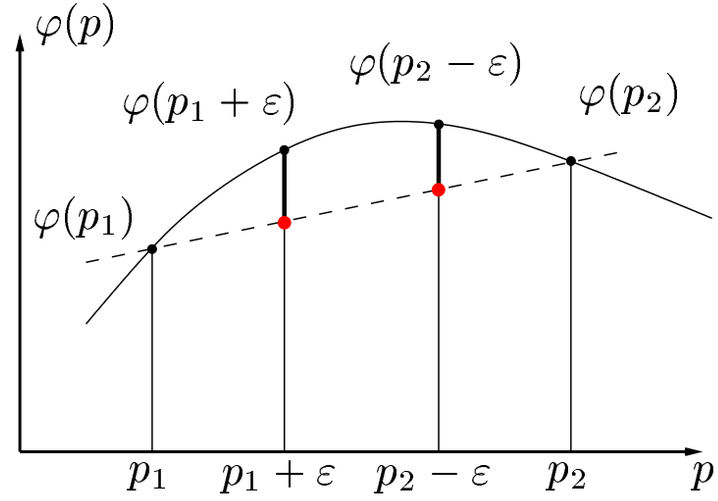
Entropie

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

Pomocná věta **V0**:

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

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



D: Plyne z monotonicity $\ln p$ a konkavity $\varphi(p)$:

$$\begin{aligned}\varphi(p_1 + p_2) &= -p_1 \ln(p_1 + p_2) - p_2 \ln(p_1 + p_2) = -p_1(\ln p_1 + \delta_1) - p_2(\ln p_2 + \delta_2) \quad \delta_1, \delta_2 \ge 0 \\ &= \varphi(p_1) + \varphi(p_2) - (p_1\delta_1 + p_2\delta_2) \le \varphi(p_1) + \varphi(p_2)\end{aligned}$$

$$\begin{array}{r} \varphi(p_1 + \varepsilon) \ge \varphi(p_1) + \frac{\varepsilon}{p_2 - p_1}\big(\varphi(p_2) - \varphi(p_1)\big) \\ \varphi(p_2 - \varepsilon) \ge \varphi(p_2) - \frac{\varepsilon}{p_2 - p_1}\big(\varphi(p_2) - \varphi(p_1)\big) \\ \hline \varphi(p_1 + \varepsilon) + \varphi(p_2 - \varepsilon) \ge \varphi(p_1) + \varphi(p_2) \end{array}$$

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.

D:

1. Nechť $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$, zkonstruujeme $\mathcal{B} = \{B_1, B_2, A_2, \ldots, A_n\}$ rozkladem $A_1 = B_1 \cup B_2$, $p(B_1) = p_1$, $p(B_2) = p_2$. Pak

$$H(\mathcal{A}) - \varphi(p_1 + p_2) = H(\mathcal{B}) - \varphi(p_1) - \varphi(p_2)$$
$$H(\mathcal{A}) - H(\mathcal{B}) = \varphi(p_1 + p_2) - \varphi(p_1) - \varphi(p_2) \leq 0 \qquad \text{(z V0)}$$

2. Pro libovolné $\mathcal{B} \preceq \mathcal{A}$ platí, že existuje posloupnost zjemnění

$$\mathcal{A} = \mathcal{A}_1 \succeq A_2 \cdots \succeq \mathcal{A}_m = \mathcal{B},$$

takových, že $\mathcal{A}_{k+1}$ je zkonstruována z $\mathcal{A}_k$ rozkladem jedné podmnožiny jako v předchozím kroku. Takže

$$H(\mathcal{A}) = H(\mathcal{A}_1) \leq H(\mathcal{A}_2) \leq \cdots \leq H(\mathcal{A}_m) = H(\mathcal{B})$$

pokračování

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

pokračování

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

D:

1. Všechny prvky mají stejnou pst $\Rightarrow$ $H(\mathcal{A})$ je maximální — přímou implikací
   - Nechť $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ a $p(A_i) = p_0 = 1/n$
   - Zkonstruujeme $\mathcal{B} = \{B_1, B_2, A_3, \ldots A_n\}$ rozložením $B_1 \cup B_2 = A_1 \cup A_2$ tak, aby $p(B_1) = p_0 + \varepsilon$ a $p(B_2) = p_0 - \varepsilon$. — ε přidané v B_1 musí být odebráno z B_2, $\sum_i p(B_i) = 1$
   - Potom, zavedeme-li $p_1 = p_0 - \varepsilon$ a $p_2 = p_0 + \varepsilon$,

$$H(\mathcal{B}) - \varphi(p_0 + \varepsilon) - \varphi(p_0 - \varepsilon) = H(\mathcal{A}) - \varphi(p_0) - \varphi(p_0)$$

$$H(\mathcal{B}) - \varphi(p_1) - \varphi(p_2) = H(\mathcal{A}) - \varphi(p_1 + \varepsilon) - \varphi(p_2 - \varepsilon)$$

$$0 \leq \varphi(p_1 + \varepsilon) + \varphi(p_2 - \varepsilon) - \varphi(p_1) - \varphi(p_2) = H(\mathcal{A}) - H(\mathcal{B}) \qquad \Leftarrow \text{V0}$$

   - takže každým ‚rozvážením‘ pravděpodobností entropie klesne

2. $H(\mathcal{A})$ je maximální $\Rightarrow$ všechny prvky mají stejnou pst — sporem
   - nechť $H(\mathcal{B})$ je maximální a existují $B_1, B_2 \in \mathcal{B}$ takové, že $p(B_1) \neq p(B_2)$
   - potom stejnou konstrukcí zjistíme, že $H(\mathcal{B}) \leq H(\mathcal{A})$, což je spor.