Statistika pro informatiku

Souhrn látky

únor 2014

Obsah

Rejstřík

1 Základy statistiky a pravděpodobnosti

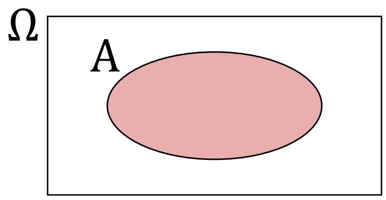
1.1 Rozlišení pojmů

Statistika TODO

Pravděpodobnost TODO

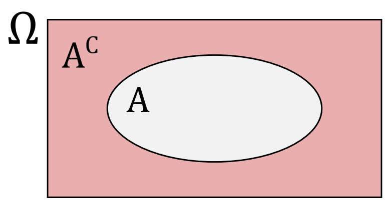
1.2 Pravděpodobnost jevu a jeho doplňku

$$\mathbb{P}(A) = \frac{size(A)}{size(\Omega)}$$



Obrázek 1: Vennův diagram základní pravděpodobnosti jevu

$$\mathbb{P}\left(A^C\right) = 1 - \mathbb{P}(A)$$

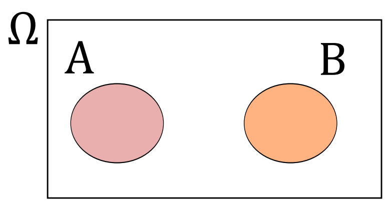


Obrázek 2: Vennův diagram doplňku jevu

1.3 Sjednocení jevů

Pro disjunktní jevy platí

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

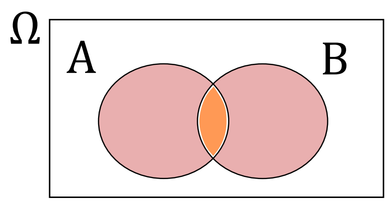


Obrázek 3: Disjunktní jevy

Pro jevy platí

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Oblast průniku by byla započítána dvakrát, proto je potřeba ji odečíst.

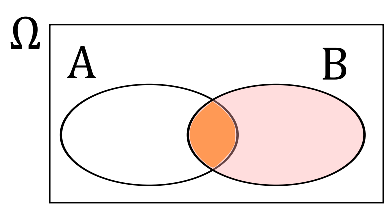


Obrázek 4: Jevy

1.4 Podmíněná pravděpodobnost

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \mathbb{P}(B) \neq 0$$

„Pravděpodobnost jevu A za podmínky, že jsme v B a že jev B nastal."



Obrázek 5: Podmíněná pravděpodobnost

1.5 Nezávislost jevů (průnik)

Pro **nezávislé** jevy platí

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) * \mathbb{P}(B).$$

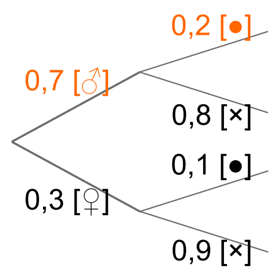
Jinak platí

$$\begin{aligned} \mathbb{P}(A \cap B) &= \mathbb{P}(A|B)\,\mathbb{P}(B) \\ \mathbb{P}(A \cap B) &= \mathbb{P}(B|A)\,\mathbb{P}(A) \\ \mathbb{P}(A \cap B \cap C \ldots) &= \mathbb{P}(A)\,\mathbb{P}(B|A)\,\mathbb{P}(C|A \cap B) \ldots \end{aligned}$$

1.6 Bayessova věta

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$$

$$\mathbb{P}\left(\male|\bullet\right)=\mathbb{P}\left(\bullet|\male\right)*\mathbb{P}\left(\male\right)=\mathbb{P}\left(\male\right)*\mathbb{P}\left(\bullet|\male\right)=0,7*0,2=\underline{\underline{0,14}}$$



Obrázek 6: Bayessova věta pomocí stromu

2 Vlastnosti

2.1 Střední hodnota

Pro diskrétní veličiny

$$\mathbb{E}X = \sum_i p_i x_i = \sum_i x_i * \mathbb{P}\left(X = x_i\right)$$

Pro spojité veličiny

$$\mathbb{E}X = \int_{-\infty}^{+\infty} x * f_x\left(x\right)\mathrm{d}x$$

(P a f jsou funkce hustoty.)

Pro libovolné náhodné veličiny platí:

$$\begin{aligned}
\mathbb{E}\left(aX+Y\right) &= a\mathbb{E}\left(X\right)+\mathbb{E}\left(Y\right)\ \text{(linearita)}\\
\mathbb{E}\left(X+Y\right) &= \mathbb{E}\left(X\right)+\mathbb{E}\left(Y\right)\\
\mathbb{E}\left(X+Y\right) &= \mathbb{E}\left(max\left\{X,\,Y\right\}\right)+\mathbb{E}\left(min\left\{X,\,Y\right\}\right)\\
\mathbb{E}X^2 &= \sum_i p_i x_i^2\ \text{(pro diskrétní jevy)}\\
\mathbb{E}\left(\max\left\{X,\,Y\right\}\right) &= \mathbb{E}\left(X\right)+E\left(Y\right)-\mathbb{E}\left(\min\left\{X,\,Y\right\}\right)\\
\mathbb{E}\left(XY\right) &= \mathbb{E}X*\mathbb{E}Y\ \text{(platí jen pro nezávislé jevy)}
\end{aligned}$$

2.2 Rozptyl

$$\sigma_x = varX = \mathbb{E}\left(X-\mathbb{E}X\right)^2 = \mathbb{E}X^2-\left(\mathbb{E}X\right)^2$$

2.3 Distribuční funkce

- Funkce je zprava spojitá.

Distribuční funkce pro diskrétní veličiny

$$F = \mathbb{P}(X \leq x_i) = \sum_{x_i \leq x} p_x\left(x_i\right)$$

Distribuční funkce pro spojité veličiny

$$F = \mathbb{P}(X \leq x_i) = \int\limits_{-\infty}^{x} f_x\left(u\right) \mathrm{d}u \ \forall x \in \mathbb{R}$$

(X je náhodná veličina, x_i je číslo)

2.4 Hustota

Funkce hustoty pro diskrétní veličiny

$$p\left(X\right) = \mathbb{P}\left(X = x\right)$$

Funkce hustoty pro spojité veličiny

$$f\left(x\right) = F_x'(x)$$

3 Rozdělení pravděpodobnosti

3.1 Diskrétní (nespojité) rozdělení

Diskrétní veličiny mohou nabývat pouze spočetného počtu hodnot (i nekonečného).

3.1.1 Bernoulliho rozdělení

$$X \sim Be\left(p\right)$$

Hustota

$$\mathbb{P}\left(1\right) = p,\ \mathbb{P}\left(0\right) = 1 - p$$

Střední hodnota

$$\mathbb{E}X = p$$

Rozptyl

$$varX = p\left(1 - p\right)$$

3.1.2 Binomické rozdělení

$$X \sim Bi\left(n,\ p\right)$$

Hustota

$$\mathbb{P}_X\left(k\right) = \binom{n}{k} p^k \left(1 - p\right)^{n-k}$$

Střední hodnota

$$\mathbb{E}X = n * p$$

Rozptyl

$$varX = np\left(1 - p\right)$$

3.1.3 Geometrické rozdělení

$$X \sim geom\left(p\right)$$

- „Počet hodů mincí než padne první panna. Tedy čekání na úspěch.“

Hustota

$$\mathbb{P}_x\left(k\right) = \left(1-p\right)^{k-1} * p$$

Distribuční funkce

$$\mathbb{P}\left(T \leq n\right) = 1 - \left(1-p\right)^n$$

Funkce přežití

$$\mathbb{P}\left(T > n\right) = \left(1-p\right)^n$$

Střední hodnota

$$\mathbb{E}X = \frac{1}{p}$$

Rozptyl

$$varX = \frac{1}{p}\left(\frac{1}{p} - 1\right)$$

3.1.4 Poissonovo rozdělení

TODO

3.2 Spojité rozdělení

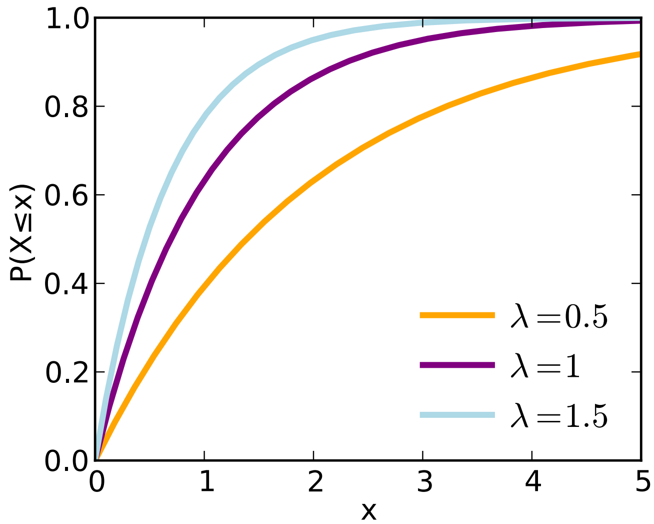
Spojité náhodné veličiny nabývají na rozdíl od diskrétních veličin nějakého intervalu.

3.2.1 Rovnoměrné rozdělení

TODO

3.2.2 Exponenciální rozdělení

$$\begin{aligned}
X &\sim exp\left(\lambda\right)\\
f\left(x\right) &= \lambda e^{-\lambda x},\ x \geq 0\\
P\left(X \leq x\right) &= 1 - e^{-\lambda x}\\
\mathbb{E}X &= \frac{1}{\lambda}\\
\mathrm{var}X &= \frac{1}{\lambda^2}
\end{aligned}$$



Obrázek 7: Graf distribuční funkce exponenciálního rozdělení[1]

3.2.3 Normální (gaussovo) rozdělení

TODO

4 Entropie

Entropie diskrétní veličiny

$$\boxed{H_b\left(X\right) = -\sum p_i \log_b p_i}$$

Entropie spojité veličiny

$$\boxed{H_b\left(X\right) = -\int\limits_{-\infty}^{+\infty} f\left(x\right) \log_b f\left(x\right) \mathrm{d}x}$$

(b je základ abecedy pro kódová slova, nejčastěji používáme binární abecedu, tedy $b = 2$)

Aditivita entropie

$$\begin{aligned} H\left(X,Y\right) &= H\left(X\right) + H\left(Y|X\right) \\ H\left(X,Y\right) &= H\left(X\right) + H\left(Y\right) \text{ (speciálně jen pro nezávislé náhodné veličiny)} \end{aligned}$$

Sdružená entropie

$$H\left(X,Y\right) = -\sum_{i,j} p_{i,j} \log p_{i,j}$$

(sdružená hustota $p_{i,j} = P\left(X = x_i,\, Y = Y_j\right)$)

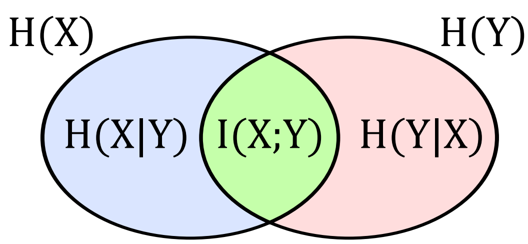
Podmíněná entropie

$$\begin{aligned} H(X|Y) &= -\sum_{i,j} p(x_i, y_i) \log p(y_j|x_i) \\ \mathbb{P}(x_i|y_j) &= \frac{\mathbb{P}(x_i,\, y_j)}{\mathbb{P}(y_i)} \begin{array}{c}\text{sdružená hustota}\\ \text{marginální hustota}\end{array} \\ \mathbb{P}(X,\, Y) &= \mathbb{P}(X|Y) * \mathbb{P}(Y) \end{aligned}$$

4.1 Vzájemná informace

$$\begin{aligned} I(X;\, Y) &= H(X) - H(X|Y) \\ I(X;\, Y) &= H(Y) - H(Y|X) \\ I(X;\, Y) &= H(X) + H(Y) - H(X,\, Y) \\ I(X;\, Y) &= I(Y;\, X) \\ I(X;\, X) &= H(X) \end{aligned}$$

$$I(X,\, Y) = \sum_{i,\, j} p_{i,j} \log \frac{p_{i,j}}{p_i * p_j} = \ldots = H(X) - H(X|Y)$$



Obrázek 8: Vzájemná informace a entropie (tedy $H(X,\, Y)$)

4.2 Kódování

Střední délka kódového slova

$$L(C) = \mathbb{E}\ell(X) = \sum_i \ell(x_i) * \mathbb{P}(X = x_i)$$

Kódování je optimální, pokdu je entropie menší než střední délka kódového slova.

4.3 Suma

Výpočet sumy

$$\begin{aligned} \sum_{n=0}^{\infty} r^n &= \frac{1}{1-r} \\ \sum_{n=0}^{\infty} n * r^n &= \frac{r}{(1-r)^2} \end{aligned}$$

4.4 Logaritmus

$$\begin{aligned}
\log_a (x_1 * x_2) &= \log_a x_1 + \log_a x_2 \\
\log_a \left(\frac{x_1}{x_2}\right) &= \log_a x_1 - \log_a x_2 \\
\log_a x^r &= r * \log_a x \\
\log_a a &= 1 \\
\log_a 1 &= 0
\end{aligned}$$

5 Náhodné procesy

Značení procesu

$$X(t,\ \omega) = X_t = X(t)$$

Střední hodnota

$$\begin{aligned}
\eta_x(t) &= \mathbb{E}X(t) = \int x(t) * f_{X_t}(x)\,\mathrm{d}x \\
& \mathbb{E}X(t) = \sum x_i(t)\,\mathbb{P}(X_t = x_i(t))
\end{aligned}$$

Pokud je střední hodnota **nezávislá** na t, **je proces stacionární** v průměru. Tedy

$$\eta(t) = \eta_x\ \forall t$$

Střední hodnota integrálu

$$\mathbb{E}\int_a^b X(t)\,\mathrm{d}t = \int_a^b \mathbb{E}X(t)\,\mathrm{d}t$$

Autokorelační funkce

$$\begin{aligned}
R_{xx}(t_1,\ t_2) &= \mathbb{E}\left[X(t_1) * \overline{X(t_2)}\right] \text{ v } \mathbb{C} \\
&= \mathbb{E}[X(t_1) * X(t_2)] = \sum_i x_i(t) * P(X(t_1) * X(t_2) = x_i(t)) \text{ v } \mathbb{R}
\end{aligned}$$

V ukázkovém příkladu se autokorelační funkce spočítala pomocí následujícího „triku“ $(0 \le t_1 \le t_2 \le 2)$

$$\begin{aligned}
R_x(t_1, t_2) &= \mathbb{E}[X(t_1) * X(t_2)] = 0 * \mathbb{P}(X(t_1) * X(t_2) = 0) + 1 * \mathbb{P}(X(t_1) * X(t_2) = 1) \\
&= \mathbb{P}(X(t_1) * X(t_2) = 1) = \mathbb{P}(X(t_1) = 1,\ X(t_2) = 1) = \\
&= \boxed{\mathbb{P}(t_1 \le A,\ t_2 \le A) = \mathbb{P}(A \ge t_1,\ t_2 \ge A) = P(A \ge t_2)}
\end{aligned}$$

6 Markovovy řetězce

Matice přechodů

$$\mathbb{P}_{i,j} = p(i,\ j) = P(X_{n+1} = s_j | X_n = s_i) = p_{i,j}$$

Markovova podmínka
TODO
Řetězec zapisujeme

- diagramem,
- maticí přechodů.

6.1 Stacionární distribuce

- Po čase se některé Markovovy řetězce ustálí v nějakém stavu.
- Rovnovážná distribuce
$$\pi : \pi = \pi * \mathbb{P}$$
- Detailní rovnováha
$$\pi_i p_{i,j} = \pi_j p_{j,i} \, \forall i, j$$

Reference

[1] The Free Encyclopedia Wikipedia. Exponential distribution. 2014.