Lineární korelační koeficient

Úvod

Mějme množinu uspořádaných dvojic reálných čísel

$$\{\,(x_1, y_1), \ldots, (x_n, y_n)\,\}.$$

Lineární korelační koeficient $r(x, y)$ udává, do jaké míry jsou hodnoty x_i a y_i svázány lineární funkcí. Pokud jsou přesně svázány lineární funkcí s kladnou derivací (tj. čím více x, tím více y a naopak), je $r(x, y) = 1$; pokud se zápornou (tj. čím více x, tím méně y a naopak), bude $r(x, y) = -1$. Pokud je to něco mezi, bude $-1 \leq r(x, y) \leq +1$.

Lineární korelační koeficient se také (v počítačovém vidění často) nazývá normalizovaná vzájemná korelace (*normalized cross-correlation, NCC).*

Vodorovnou čarou nad čímkoliv označíme střední hodnotu. Tedy střední hodnota množiny čísel $\{x_1, \ldots, x_n\}$ je $\overline{x}$ a jejich variance $\overline{(x - \overline{x})^2}$. Variance je tedy střední hodnota čtverců, jenže nikoliv původních čísel, ale čísel posunutých tak, že jejich střední hodnota je 0. Je

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \overline{(x - \overline{x})^2} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Výpočet

Jak se r počítá? Uděláme to v následujících třech krocích:

1. Posuneme čísla $\{x_1, \ldots, x_n\}$ tak, aby měla nulovou střední hodnotu. Nová čísla označíme jako

   $$x'_i = x_i - \overline{x}, \qquad i = 1, \ldots, n.$$

   Totéž uděláme s čísly $\{y_1, \ldots, y_n\}$.

2. Čísla $\{x'_1, \ldots, x'_n\}$ vydělíme všechna stejnou konstantou tak, aby nová čísla měla jednotkovou varianci:

   $$x''_i = x'_i / \sqrt{\overline{x'^2}}, \qquad i = 1, \ldots, n.$$

   Totéž uděláme s $\{y'_1, \ldots, y'_n\}$.

3. Pomocí normalizovaných čísel x''_i a y''_i je lineární korelační koeficient dán jednoduchým vztahem

   $$r(x, y) = \overline{x''y''} = \frac{1}{n} \sum_{i=1}^{n} x''_i y''_i. \tag{1}$$

Vlastnosti

Z algoritmu jsou zjevné následující důležité vlastnosti $r(x, y)$:

- *Symetrie*, tj. $r(x, y) = r(y, x)$.
- *Invariance vůči lineární transformaci x a y*. Tj. pro jakákoliv b_x a b_y a jakákoliv kladná a_x a a_y je

$$r(\, a_x x + b_x,\, a_y y + b_y \,) = r(x, y).$$

- $-1 \leq r(x, y) \leq +1$. To lze snadno uvidět z toho, ze vztah (1) lze interpretovat jako skalární součin dvou vektorů z n-rozměrného lineárního prostoru,

$$\mathbf{x}'' = \frac{(x''_1, \ldots, x''_n)}{\sqrt{n}}, \qquad \mathbf{y}'' = \frac{(y''_1, \ldots, y''_n)}{\sqrt{n}}.$$

  Díky jednotkové varianci a nulové střední hodnotě x'' a y'' mají vektory $\mathbf{x}''$ a $\mathbf{y}''$ jednotkovou délku. Vlastnost vyplývá ze známé skutečnosti, že skalární součin dvou libovolných jednotkových vektorů leží v intervalu $\langle -1, +1 \rangle$.

  Položíme-li $y_i = x_i$, bude $\mathbf{x}'' = \mathbf{y}''$ a tedy $r(x, y) = +1$. Položíme-li $y_i = -x_i$, bude $\mathbf{x}'' = -\mathbf{y}''$ a $r(x, y) = -1$.