Korelacja vs kauzacja

Krzysztof Szymański Mateusz Rajs Filip Ficek

Wstęp

Motywacja

Często doświadczamy korelacji - sytuacji, gdy jedna wielkość jest wiązana z drugą. Choć intuicyjnie wydaje się, że występuje wtedy powiązanie, wcale tak nie musi być. Celem projektu jest pokazanie, że przy odpowiednio dużej liczbie danych korelacje są nieuniknione.

Problem badawczy

Ile co najmniej potrzeba zestawów danych o długości n, aby wśród nich istniały takie dwa, że ich współczynnik korelacji jest większy niż zadana wartość, np. 95%?

Postawienie problemu

Współczynnik korelacji zestawów (list) danych X oraz Y definiuje się jako:

$$r_{XY} := \frac{E(X \cdot Y) - E(X) \cdot E(Y)}{\sigma_X \cdot \sigma_Y}$$

Ta wielkość mierzy, jak podobne są dwa zestawy danych. Przyjmuje wartość z zakresu:

$$-1 \leqslant r_{XY} \leqslant 1.$$

Normalizacja danych

Współczynnik korelacji jest niezmienniczy ze względu na liniowe transformacje danych, zatem można zastąpić dane zestawy poprzez:

$$X \rightsquigarrow \frac{X - E(X)}{\sigma_X} .$$

Po takiej transformacji średnia wynosi 0, a odchylenie standardowe 1. Wzór na korelację upraszcza się do

$$r_{XY} = E(X \cdot Y) .$$

Geometria

Znormalizowane dane spełniają:

$$\sum x_k = 0$$
$$\sum x_k^2 = n$$

Wektor wodzący takiego zestawu danych leży na przekroju S^{n-1} oraz płaszczyzny przez jej środek, zatem przestrzenią możliwości jest S^{n-2}. Ponieważ wektory danych mają długość $\sqrt{n}$, współczynnik korelacji między nimi to cosinus kąta między danymi wektorami.

Problem badawczy II

Ile punktów można rozmieścić na S^{n-2} aby odległość kątowa każdych dwóch była większa od danej wartości.

Wyniki

Przypadek ogólny

Jeżeli punkty są od siebie odległe na tyle, by nie być skorelowane, każdy zajmuje pewien *zakazany obszar*. Policzywszy stosunek pola całej sfery do pola tych obszarów otrzymujemy oszacowanie ogólne:

$$\boxed{\frac{2}{I\left(\frac{1-r}{2}, \frac{n-2}{2}, 0.5\right)}}$$

Można to oszacowanie poprawić, mnożąc powyższy wzór przez gęstość upakowania kul w przestrzeni $(n-2)$-wymiarowej, którą oznaczamy przez ρ_n (patrz Tabela 1).

Asymptotyka

Wspomniany wyżej wzór dla ustalonego r rośnie asymptotycznie jak

$$\sqrt{n} \cdot \left(\sqrt{\frac{2}{1-r}}\right)^n$$

Zestaw danych dł. 4

Przy $n = 4$ możemy skorzystać z geometrii sferycznej, aby uzyskać lepsze o ok. 10% szacowanie szczególne: [1]

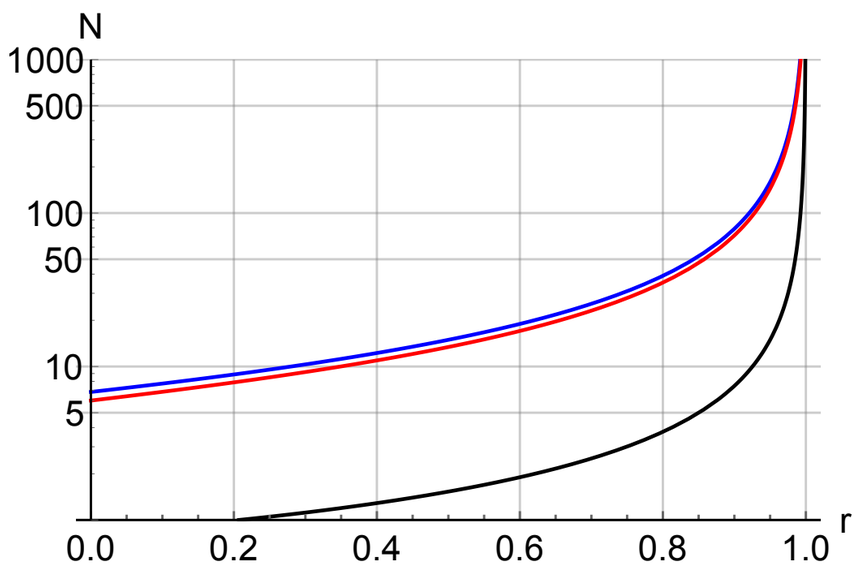
$$\boxed{2 + \frac{2\pi}{6 \operatorname{arcctg} \sqrt{2r+1} - \pi}}$$

Podsumowanie

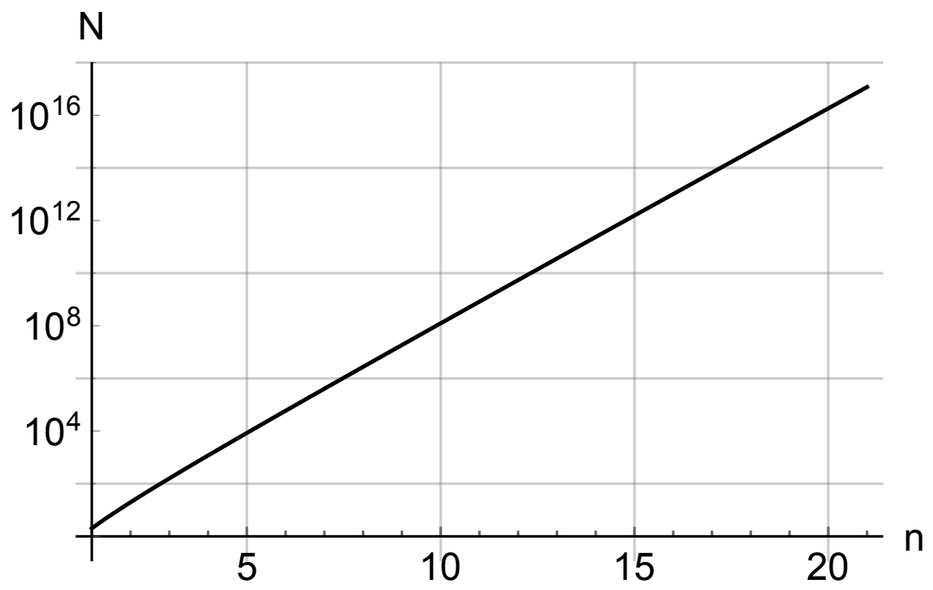
Udało nam się oszacować liczbę zestawów danych potrzebnych, aby pojawiła się między nimi korelacja. Oszacowanie jest wykładnicze, a więc bardzo duże. Problem nie został bynajmniej wyczerpany. Dużym czynnikiem w obliczeniach są gęstości upakowania ρ_n, o których podejrzewamy, że wykładniczo maleją. Jest to bardzo słabo zbadany obszar matematyki.

Oznaczenia:

1. $E(X)$ - średnia zestawu danych X.
2. σ_X - odchylenie standardowe danych X.
3. $I(x, y, z)$ - funkcja specjalna.
4. S^2 - sfera w trzech wymiarach.
5. S^{n-1} - sfera w n wymiarach.



Rysunek 1: Oszacowanie szczególne i ogólne dla $n = 4$.



Rysunek 2: Oszacowanie ogólne przy $r = 0.95$. Skala logarytmiczna.

Wymiary	1	2	3	8	24	Asymptotyka
ρ_n	100%	90%	74%	25%	0.2%	$\frac{2n}{2^n} \leqslant \rho_n \leqslant 2^{-0.599n}$

Tabela 1: [2, 3] Znane obecnie gęstości upakowania przestrzeni kulami.

Literatura

[1] Edith Mooers, *Tammes's Problem*

[2] Thomas Hartman et al., *Sphere packing and quantum gravity*

[3] S. Torquato, F. H. Stillinger, *New Conjectural Lower Bounds on the Optimal Density of Sphere Packings*, arXiv: 0508381