Sự độc lập của các sự kiện

Trần Vĩnh Đức

HUST

Ngày 24 tháng 8 năm 2017

Tài liệu tham khảo

- Eric Lehman, F Thomson Leighton & Albert R Meyer, *Mathematics for Computer Science*, 2013 (Miễn phí)
- Michael Mitzenmacher và Eli Upfal, *Probability and Computing*, 2005
- Nguyễn Tiến Dũng và Đỗ Đức Thái, *Nhập Môn Hiện Đại Xác Suất & Thống Kê*.

Định nghĩa

Sự kiện A là độc lập với sự kiện B nếu

$$\Pr[A \mid B] = \Pr[A]$$

hoặc nếu $\Pr[B] = 0$.

Biết B xảy ra không làm thay đổi xác suất A xảy ra.

Ví dụ

Tung hai đồng xu độc lập

$$A = \text{ sự kiện đồng xu } 1 \text{ ngửa}$$
$$B = \text{ sự kiện đồng xu } 2 \text{ ngửa}$$

Ta có

$$\Pr[A \mid B] = \Pr[A] = 1/2$$

Câu hỏi

Hai sự kiện rời nhau có luôn độc lập?

Định lý

Sự kiện A là độc lập với sự kiện B nếu và chỉ nếu

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$$

Chứng minh.

- Nếu $\Pr[B] = 0$, vậy thì

$$\Pr[A \cap B] = 0 = \Pr[A] \cdot \Pr[B].$$

- Nếu $\Pr[B] > 0$, vậy thì

$$\begin{aligned} & \Pr[A \mid B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \Pr[B] \\ \iff \quad & \Pr[A \cap B] = \Pr[A] \cdot \Pr[B] \end{aligned}$$

$\square$

Hệ quả (Tính đối xứng)

Sự kiện A *độc lập với sự kiện* B *nếu và chỉ nếu sự kiện* B *độc lập với sự kiện* B*.*

Câu hỏi

Tung hai đồng xu độc lập

$$A = \text{ sự kiện chúng cùng mặt}$$
$$B = \text{ sự kiện đồng xu thứ nhất ngửa}$$

Hỏi sự kiện A có độc lập với sự kiện B?

Câu hỏi

Xét hai đồng xu với tính chất

$$\Pr[H] = p, \qquad \Pr[T] = 1 - p.$$

Xét hai sự kiện

$$\begin{aligned} A &= \text{ sự kiện chúng cùng mặt} \\ B &= \text{ sự kiện đồng xu thứ nhất ngửa} \end{aligned}$$

Với những giá trị nào của p thì hai sự kiện này độc lập?

Định nghĩa

Các sự kiện $E_1, E_2, \ldots, E_n$ là độc lập nếu, với mọi tập con $S \subseteq \{1, \ldots, n\}$,

$$\Pr\left[\bigcap_{j \in S} E_j\right] = \prod_{j \in S} \Pr[E_j].$$

Ví dụ

Khi $n = 3$: các sự kiên E_1, E_2, E_3 là độc lập nếu

$$\begin{aligned}
\Pr[E_1 \cap E_2] &= \Pr[E_1] \cdot \Pr[E_2] \\
\Pr[E_1 \cap E_3] &= \Pr[E_1] \cdot \Pr[E_3] \\
\Pr[E_2 \cap E_3] &= \Pr[E_2] \cdot \Pr[E_3] \\
\Pr[E_1 \cap E_2 \cap E_3] &= \Pr[E_1] \cdot \Pr[E_2] \cdot \Pr[E_3]
\end{aligned}$$

Ví dụ

Tung ba đồng xu độc lập

$$
\begin{aligned}
A_1 &= \text{ sự kiện kết quả đồng 1 giống kết quả đồng 2} \\
A_2 &= \text{ sự kiện kết quả đồng 2 giống kết quả đồng 3} \\
A_3 &= \text{ sự kiện kết quả đồng 3 giống kết quả đồng 1}
\end{aligned}
$$

Có phải các sự kiện này độc lập?

Định nghĩa

Các sự kiện $E_1, E_2, \ldots, E_n$ là độc lập từng đôi một nếu, với mọi $i, j \in \{1, 2, \ldots, n\}$, $i \neq j$,

$$\Pr[E_i \cap E_j] = \Pr[E_i] \cdot \Pr[E_j].$$

Bài toán ngày sinh nhật

Một lớp học bất kỳ có 30 sinh viên. Xác suất để ít nhất hai người trong lớp trùng ngày sinh nhật là bao nhiêu?

Không ai trùng ngày sinh

- Nếu đã có một người trong phòng, người thứ 2 vào phòng, xác suất chị ta có ngày sinh khác với người đầu tiên là $(1 - 1/365)$.
- Người thứ 3 vào phòng, xác suất anh ta có khác với hai người trong phòng là $(1 - 2/365)$
- Giả sử trong phòng đã có $k - 1$ người khác ngày sinh nhật. Khi người thứ k vào phòng, xác suất chị ta có ngày sinh khác với $k - 1$ người trong phòng là $(1 - (k - 1)/365)$.
- Vậy xác suất 30 người trong phòng có ngày sinh nhật đôi một khác nhau là

$$\left(1 - \frac{1}{365}\right) \cdot \left(1 - \frac{2}{365}\right) \cdot \left(1 - \frac{3}{365}\right) \cdots \left(1 - \frac{29}{365}\right) \approx 0.2937$$

Khai triển Taylor của hàm e^{-x} là

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \cdots .$$

Vậy khi $x < 1$, ta có

$$1 - x \le e^{-x}.$$

Tổng quát hoá

Nếu có m người và n ngày sinh có thể, vậy xác suất để mọi người có ngày sinh khác nhau là

$$\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\cdot\left(1-\frac{3}{n}\right)\cdots\left(1-\frac{m-1}{n}\right)=\prod_{j=1}^{m-1}\left(1-\frac{j}{n}\right)$$

Dùng công thức $1-k/n \approx e^{-k/n}$ khi k nhỏ so với n, ta được

$$\begin{aligned}\prod_{j=1}^{m-1}\left(1-\frac{j}{n}\right) &\approx \prod_{j=1}^{m-1} e^{-j/n}\\ &= \exp\left(-\sum_{j=1}^{m-1}\frac{j}{n}\right)\\ &= \exp\left(-m(m-1)/2n\right)\\ &\approx \exp(-m^2/2n)\end{aligned}$$

Khi xác suất trùng ngày sinh gần $1/2$

- Giá trị của m từ phương trình

$$\frac{m^2}{2n} = \ln 2 \quad \Longleftrightarrow \quad m = \sqrt{2n \ln 2} \approx 1.1774\sqrt{n}.$$

  đảm bảo rằng "xác suất m người có ngày sinh nhật khác nhau là $1/2$".
- Trong trường hợp $n = 365$, ta được $m = 22.49$. Có nghĩa rằng:

Trong một nhóm gồm 23 người, xác suất để có hai người trùng ngày sinh ít nhất bằng $1/2$.

Nguyên lý ngày sinh nhật

Nếu một năm có d ngày và có $\sqrt{2d}$ người trong phòng, vậy thì xác suất có hai người cùng ngày sinh nhật khoảng $1 - 1/e \approx 0.632$.

Đánh giá thô

- Đặt $E_k =$ sự kiện ngày sinh người thứ k không trùng với $k-1$ người trước đó.
- Xác suất k người đầu tiên có trùng ngày sinh là

$$\begin{aligned}
\Pr[\overline{E_1} \cup \overline{E_1} \cup \cdots \cup \overline{E_k}] &\le \sum_{i=1}^{k} \Pr[E_i] \\
&\le \sum_{i=1}^{k} \frac{i-1}{n} \\
&= \frac{k(k-1)}{2n}
\end{aligned}$$

Với $\lfloor \sqrt{n} \rfloor$ người, xác suất ít nhất $1/2$ là mọi ngày sinh nhật của họ đều khác nhau.

Đánh giá thô (tiếp)

- Giả sử $\lceil\sqrt{n}\rceil$ người đầu tiên có ngày sinh nhật khác nhau.
- Mỗi người tiếp theo có xác suất $\sqrt{n}/n$ trùng ngày sinh nhật với $\lceil\sqrt{n}\rceil$ người đầu tiên.
- Vậy xác suất $\lceil\sqrt{n}\rceil$ người tiếp theo có ngày sinh nhất khác với $\lceil\sqrt{n}\rceil$ người đầu tiên là

$$\left(1 - \frac{1}{\lceil\sqrt{n}\rceil}\right)^{\lceil\sqrt{n}\rceil} < \frac{1}{e} < \frac{1}{2}.$$

Khi có $2\lceil\sqrt{n}\rceil$ người, xác suất nhiều nhất $1/e$ là mọi ngày sinh của họ đều khác nhau.

Hàm băm mật mã

Định nghĩa

Hàm băm là hàm tính được một cách "hiệu quả"

$$H : \{0, 1\}^* \to \{0, 1\}^d.$$

Nó ánh xạ một xâu nhị phân độ dài bất kỳ trong không gian thông điệp thành một xâu nhị phân độ dài cố định, gọi là *mã băm*.

Thông thường độ dài của mã băm là $d = 128, 160, 256$ hoặc 512 bit.

Ví dụ

Một số hàm băm trong thực tế.

Hàm băm	d
MD4, MD5	128
SHA-1	160
SHA-256	256
SHA-512	512
SHA-3 (Keccak)	$224, 256, 384, 512$

Hàm băm và tính nén

- Giả sử hàm băm được thiết kế một cách lý tưởng (như ngẫu nhiên), khi đó cho trước mã băm x, xác suất tìm được một dữ liệu m thỏa mãn $H(m) = x$ chỉ là 2^{-d}.
- Con số này rất gần 0 khi d đủ lớn. Như vậy hàm băm cho ta một biểu diễn *nén* hợp lý của dữ liệu.

15.04

(Vivid Vervet): April 2015 (Supported until January 2016)

md5 Hash	***Version***
53c869eba8686007239a650d903847fd	ubuntu-15.04-desktop-amd64.iso
6ea04093b767ad6778aa245d53625612	ubuntu-15.04-desktop-i386.iso
487f4a81f22f8597503db3d51a1b502e	ubuntu-15.04-server-amd64.iso
49de7a0ed202d11456126589e2d1db22	ubuntu-15.04-server-i386.iso
fcfba8de8848944705cd200ff76c53cf	ubuntu-15.04-snappy-amd64-generic.img.xz
ef2a4951a2e889908a55c980d2bea475	ubuntu-15.04-snappy-armhf-bbb.img.xz

Định nghĩa

Một *xung đột* cho hàm H là một cặp $m_0, m_1 \in \{0,1\}^*$ thỏa mãn

$$H(m_0) = H(m_1) \quad \text{và} \quad m_0 \neq m_1.$$

- Vì kích thước đầu vào của hàm băm lớn hơn so với kích thước đầu ra, nên theo nguyên lý "chuồng bồ câu", luôn tồn tại xung đột.
- Tuy vậy, để hàm băm là an toàn thì việc tìm thấy xung đột phải rất "khó". Có nghĩa rằng, xác suất tìm thấy xung đột phải "nhỏ".

Nguyên lý ngày sinh nhật

Xét tập thông điệp M với $|M| > \sqrt{2 \cdot 2^d}$ và nếu các giá trị trên M được chọn ngẫu nhiên (đều) và độc lập. Vậy thì

$$\exists x, y \in M \quad \text{thoả mãn} \quad H(x) = H(y)$$

với xác suất $> 1/2$.