Mathematics for Data Science

Khoat Than

School of Information and Communication Technology

Hanoi University of Science and Technology

2018

VIASM

Vietnam Institute for Advanced Study in Mathematics

PHẦN 1
ĐẠI SỐ TUYẾN TÍNH

Ma trận và chuyển vị

- *__Ma trận__* là một bảng hình chữ nhật gồm các phần tử được sắp xếp thành các hàng và các cột. Nếu ma trận A có m hàng và n cột và các phần tử của A là các số thực thì ký hiệu $A \in R^{m\times n}$.
- Cho $A \in R^{m\times n}$, ta nói $B \in R^{n\times m}$ là *__chuyển vị__* của A nếu:

$$b_{ij} = a_{ji} \; \forall 1 \le i \le n, 1 \le j \le m$$

Ký hiệu: $B = A^T$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \Rightarrow \mathbf{x}^T = \begin{bmatrix} x_1 \; x_2 \; \ldots \; x_m \end{bmatrix}; \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ddots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \Rightarrow \mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \ldots & a_{m1} \\ a_{12} & a_{22} & \ldots & a_{m2} \\ \ldots & \ldots & \ddots & \ldots \\ a_{1n} & a_{2n} & \ldots & a_{mn} \end{bmatrix}$$

Nếu $A = A^T$ thì ta gọi A là *__ma trận đối xứng__*

Phép nhân hai ma trận

- Cho hai ma trận $A \in R^{m\times n}, B \in R^{n\times p}$, tích của hai ma trận được ký hiệu là $C \in R^{m\times p}$ với:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}\,, 1 \le i \le m, 1 \le j \le p$$

Tính chất:

- Phép nhân hai ma trận không có tính giao hoán: $AB \ne BA$
- Tính kết hợp: $ABC = (AB)C = A(BC)$
- Tính phân phối đối với phép cộng: $A(B + C) = AB + AC$
- $(AB)^T = A^T B^T$
- ***Inner product***: bằng cách coi vector là một trường hợp đặc biệt của ma trận $n \times 1$, tích vô hướng của 2 vector $x, y \in R^n$ được định nghĩa là:

$$x^T y = y^T x = \sum_{i=1}^{n} x_i y_i$$

- ***Element-wise product*** (*Hadamard product*): Tích Hadamard của hai ma trận cùng kích thước $A, B \in R^{m\times n}$ được ký hiệu là:

$$C = A \odot B \in R^{m\times n} \text{ trong đó } c_{ij} = a_{ij} \times b_{ij}$$

Ma trận đơn vị, Ma trận nghịch đảo

- ***Đường chéo chính*** của một ma trận (có thể không vuông) là tập hợp các phần tử có chỉ số hàng và chỉ số cột như nhau

- Một ma trận chỉ có các phần tử trên đường chéo chính là khác 0 được gọi là ***ma trận đường chéo*** *(diagonal matrix).* Giả sử D là một ma trận đường chéo có các phần tử trên đường chéo chính là $d_1, d_2, \dots, d_n$, ký hiệu: $D = diag(d_1, d_2, \dots, d_n)$

- Một ma trận vuông với các phần tử trên đường chéo chính bằng 1, còn lại bằng 0 được gọi là ***ma trận đơn vị***, và ký hiệu là I_n.

- Cho một ma trận vuông $A \in R^{n \times n}$, nếu tồn tại ma trận vuông B $\in R^{n \times n}$ sao cho: $AB = I_n$ thì ta nói A là ***khả nghịch*** và B được gọi là ***ma trận nghịch đảo*** của A.

Ký hiệu: $B = A^{-1}$.

Tính chất:

- $A.A^{-1} = I_n$
- $(AB)^{-1} = B^{-1}A^{-1}$
- Giả sử $D = diag(d_1, d_2, \dots, d_n)$ là một ma trận đường chéo vuông thì D khả nghịch iff $d_i \neq 0\ \forall i = 1, \dots, n$ và nếu D khả nghịch thì:

$$D^{-1} = diag(d_1^{-1}, d_2^{-1}, \dots, d_n^{-1})$$

Định thức

- **Định nghĩa:** ***Định thức*** của một ma trận vuông A được ký hiệu là $detA$. Có nhiều cách định nghĩa định thức, cách phổ biến nhất là sử dụng quy nạp theo bậc của ma trận.
  - Với $n = 1$, $detA$ chính là phần tử duy nhất của ma trận đó
  - Với một ma trận vuông bậc $n > 1$:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \ddots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \Rightarrow \det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{A}_{ij})$$

Với A_{ij} là ma trận thu được bằng cách xoá hàng thứ i và cột thứ j của ma trận A, hay còn gọi là ***phần bù đại số*** của A ứng với phần tử ở hàng i, cột j.

Định thức

- **Tính chất:**

- $detA = detA^T$
- $detI_n = 1$
- $\det(AB) = detA.detB$
- $detA^{-1} = \frac{1}{detA}$
- Nếu một ma trận có một hàng hoặc một cột là một vector 0 thì định thức của nó bằng 0
- Một ma trận là khả nghịch khi và chỉ khi định thức của nó khác 0
- Định thức của một ma trận tam giác (vuông) bằng tích các phần tử trên đường chéo chính

Tổ hợp tuyến tính-Không gian sinh

❑ Tổ hợp tuyến tính (*linear combination*)

Cho các vector khác không $a_1, \dots, a_n \in R^m$ và các số thực $x_1, x_2, \dots, x_n$. Khi đó vector:

$$b = x_1 a_1 + x_2 a_2 + \cdots + x_n a_n$$

được gọi là một ***tổ hợp tuyến tính*** của $a_1, \dots, a_n \in R^m$.

Xét ma trận $A = [a_1, a_2, \dots, a_n] \in R^{m \times n}$ và $x = [x_1, x_2, \dots, x_n]^T$ là một vector cột $n \times 1$, ta có thể viết lại:

$$b = Ax$$

và b là một tổ hợp tuyến tính các cột của A

Tổ hợp tuyến tính-Không gian sinh

- Tập hợp tất cả các vector có thể biểu diễn được như là một tổ hợp tuyến tính của các vector khác không $a_1, \dots, a_n \in R^m$ được gọi là ***không gian sinh* (*span space*)** của hệ các vector đó, và được ký hiệu là $span(a_1, \dots, a_n)$

- Nếu phương trình:

$$x_1 a_1 + x_2 a_2 + \cdots + x_n a_n = 0$$

Có nghiệm duy nhất $x_1 = x_2 = \cdots = x_n = 0$ thì ta nói hệ $\{a_1, a_2, \dots, a_n\}$ là ***độc lập tuyến tính***. Ngược lại ta nói hệ đó là ***phụ thuộc tuyến tính.***

Tổ hợp tuyến tính-Không gian sinh

❑ Tính chất:

- Một hệ là phụ thuộc tuyến tính iff tồn tại một vector trong hệ đó là tổ hợp tuyến tính của các vector còn lại
- Tập con khác rỗng của một hệ độc lập tuyến tính là một hệ độc lập tuyến tính
- Tập hợp các cột (các hàng) của một ma trận khả nghịch tạo thành một hệ độc lập tuyến tính

Cơ sở của một không gian

- Một hệ các vector $\{a_1, a_2, \dots, a_n\}$ trong không gian vector m chiều $V = R^m$ được gọi là một cơ sở nếu hai điều kiện sau được thoả mãn:
  - $V \equiv span(a_1, a_2, \dots, a_n)$
  - $\{a_1, a_2, \dots, a_n\}$ là một hệ độc lập tuyến tính

→ Nhận thấy: $n = m$

Khi đó, mọi vector $b \in V$ đều có thể biểu diễn duy nhất dưới dạng một tổ hợp tuyến tính của các a_i

Hạng của ma trận

- Hạng (rank) của một ma trận $A \in R^{m \times n}$, ký hiệu là $rank(A)$, được định nghĩa là số lượng lớn nhất các cột (các hàng) của nó tạo thành một hệ độc lập tuyến tính

- **Tính chất:**

- $rank(A) = rank(A^T)$
- Nếu $A \in R^{m \times n}$ thì $rank(A) \leq \min(m, n)$
- $rank(AB) \leq \min(rank(A), rank(B))$
- $rank(A + B) \leq rank(A) + rank(B)$
- Nếu $A \in R^{m \times n}, B \in R^{n \times k}$ thì: $rank(A) + rank(B) - n \leq rank(AB)$
- Nếu A là một ma trận vuông khả nghịch thì $rank(A) = n$

Hệ trực chuẩn, ma trận trực giao

- Một hệ cơ sở $\{u_1, u_2, \dots, u_m\} \in R^m$ được gọi là ***trực giao*** nếu:

$$u_i \neq 0 \text{ và } u_i^T u_j = 0 \ \forall 1 \leq i \neq j \leq m$$

- Một hệ cơ sở $\{u_1, u_2, \dots, u_m\} \in R^m$ được gọi là ***trực chuẩn*** nếu:

$$\|u_i\|_2^2 = u_i^T u_i = 1 \text{ và } u_i^T u_j = 0 \ \forall 1 \leq i \neq j \leq m$$

- Gọi $U = [u_1, u_2, \dots, u_m]$ với $\{u_1, u_2, \dots, u_m\} \in R^m$ là một hệ trực chuẩn thì $UU^T = U^T U = I_{\mathrm{m}}$.

Ngược lại nếu một ma trận U thoả mãn: $UU^T = U^T U = I_{\mathrm{m}}$ thì U được gọi là ***ma trận trực giao***.

Vietnam Institute for Advanced Study in Mathematics

Hệ trực chuẩn, ma trận trực giao

❑ Tính chất của ma trận trực giao

Cho U là một ma trận trực giao $n \times n$, khi đó:

- U^T cũng là một ma trận trực giao
- $U^{-1} = U^T$
- $\det(U) \in \{-1, 1\}$
- Ma trận trực giao thể hiện ***phép xoay*** một vector. Giả sử có 2 vector $x, y \in R^n$, dùng ma trận U để xoay 2 vector này ta được Ux, Uy và:
$$(Ux)^T(Uy) = x^T(U^TU)y = x^Ty$$

Như vậy phép xoay không làm thay đổi thích vô hướng giữa hai vector

Trị riêng và vector riêng

- Cho một ma trận vuông $A \in R^{n \times n}$, một vector khác không $x \in R^n$ và một số vô hướng (có thể thực hoặc phức) λ.

- Nếu $Ax = \lambda x$ thì ta nói λ và x là một cặp ***trị riêng, vector riêng*** của ma trận A
- $Ax = \lambda x \leftrightarrow (A - \lambda I)x = 0$ nên λ là nghiệm của phương trình $\det(A - tI) = 0$ → $p_A(t) = \det(A - tI)$ được gọi là đa thức đặc trưng của ma trận A.

- **Tính chất:**

- Nếu x là một vecto riêng của A ứng với λ thì kx với $k \neq 0$ cũng là vecto riêng ứng với λ.
- Tích tất cả các giá trị riêng của một ma trận bằng định thức của ma trận đó.
- Tổng tất cả các giá trị riêng của một ma trận bằng tổng các phần tử trên đường chéo của ma trận đó
- Mọi ma trận vuông bậc n đều có n trị riêng (thực hoặc phức, kể cả lặp)
- Nếu A, B là 2 ma trận vuông cùng bậc thì AB và BA có cùng tập các giá trị riêng

Trị riêng và vector riêng

❑ Thuật toán PageRank của Google

- Khi truy cập một trang web, có thể tìm thấy một số liên kết chỉ đến các trang web khác → Nếu có nhiều trang web chỉ đến một trang web cụ thể nào đó thì trang web đó phải khá quan trọng
- Một cách liên hệ đơn giản, khi bắt đầu "surfing" trên internet thông qua các liên kết từ trang web này đến trang web khác, thì thứ hạng của một trang web phải tỷ lệ với ***xác suất*** mà chúng ta duyệt đến trang web đó sau một thời gian "surfing" rất dài.
- Về mặt toán học, biểu diễn các trang web dưới dạng một đồ thị có hướng với các đỉnh là $P_1, P_2, \dots, P_N$ và (P_i, P_j) là một cạnh nếu trang web P_i có link liên kết đến trang web P_j
- Thuật toán PageRank sinh ra một vector ranking π thoả mãn:

$$\pi(P) = \frac{\pi(P_1)}{\deg^-(P_1)} + \frac{\pi(P_2)}{\deg^-(P_2)} + \dots + \frac{\pi(P_m)}{\deg^-(P_m)}$$

Trong đó $P_1, \dots, P_m$ là các trang web có link liên kết đến P và $\deg^-(P_i)$ là số link đi ra từ trang web P_i (giả sử các giá trị này đều khác 0 để quá trình duyết web được liên tục)

- Đặt A là ma trận kề của đồ thị có hướng này, và D là ma trận đường chéo:

$$D = diag(\deg^-(P_1), \deg^-(P_1), \dots, \deg^-(P_N))$$

Đặt $W = A^T D^{-1}$ thì $\pi = W\pi$ hay π là một ***giá trị riêng*** của ma trận W

Chéo hoá và phân tích ma trận

❑ Chéo hoá ma trận

Giả sử $x_1, \dots, x_n \neq 0$ là các vector riêng của một ma trận vuông A ứng với các giá trị riêng $\lambda_1, \dots, \lambda_n$. Các x_i thường sẽ được chọn sao cho $x_i^T x_i = 1$.

Đặt $D = diag(\lambda_1, \dots, \lambda_n)$ và $X = [x_1, \dots, x_n]$ ta sẽ có: $AX = XD$

Hơn nữa nếu các giá trị riêng x_i là độc lập tuyến tính thì X là một ma trận khả nghịch, do đó:

$$A = XDX^{-1}$$

Do D là một ma trận đường chéo nên biểu diễn trên được gọi là ***chéo hoá ma trận***

Chéo hoá và phân tích ma trận

❑ Chéo hoá ma trận

Tính chất:

- Cách biểu diễn ma trận bằng cách chéo hoá trên còn được gọi là ***eigen decomposition***
- Chéo hoá ma trận chỉ áp dụng với ma trận vuông
- Một ma trận vuông bậc n là chéo hoá được iff nó có đủ n trị riêng độc lập tuyến tính
- Chéo hoá ma trận giúp tính toán dễ dàng các A^k

$$A^2 = (XDX^{-1})(XDX^{-1}) = XD^2X^{-1}$$
$$A^k = XD^kX^{-1}$$

Nếu A khả nghịch: $A^{-1} = (XDX^{-1})^{-1} = XD^{-1}X^{-1}$

- Nếu A mà một ma trận đối xứng thì tồn tại ma trận trực giao U sao cho U^TAU là ma trận chéo.

Chéo hoá và phân tích ma trận

❑ Phân tích ma trận

- Việc phân tích một ma trận ra thành tích của nhiều ma trận đặc biệt khác (Matrix Factorization hoặc Matrix Decomposition) mang lại nhiều ích lợi quan trọng như:
  - giảm số chiều dữ liệu, nén dữ liệu
  - tìm hiểu các đặc tính của dữ liệu, clustering
  - giải các hệ phương trình tuyến tính, và nhiều ứng dụng khác…
- ***Singular Value Decomposition (SVD)*** là một phương pháp phân tích ma trận rất đẹp và có nhiều ứng dụng trong thực tế

Chéo hoá và phân tích ma trận

❑ Phân tích ma trận

Định lý: một ma trận $A_{m\times n}$ ($A \in R^{m\times n}$) đều có thể phân tích dưới dạng:

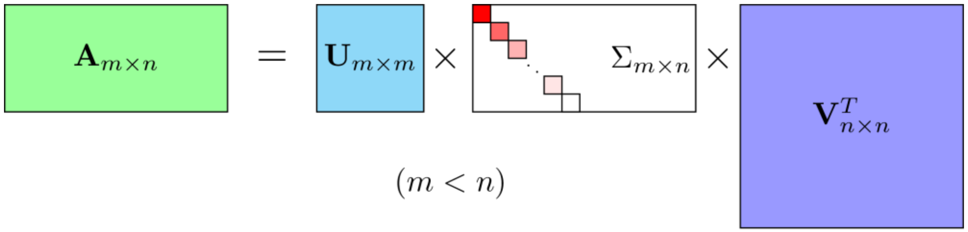
$$A_{m\times n} = U_{m\times m} \times \Sigma_{m\times n} \times V_{n\times n}^T$$

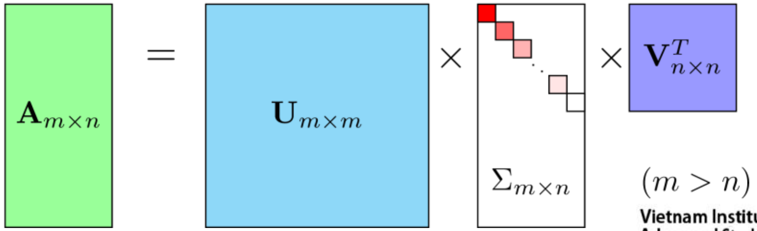
trong đó U, V là các ma trận trực giao và Σ là ma trận đương chéo không vuông với các phần tử trên đường chéo là $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0 = \cdots = 0$ với r là hạng của ma trận A

→ Phép phân tích này được gọi là

Singular Value Decomposition

(SVD)





Vietnam Institute for Advanced Study in Mathematics

Chéo hoá và phân tích ma trận

❑ Phân tích ma trận

Một ứng dụng minh hoạ của SVD trong ***image compression.***

- Để đơn giản, xét một bức ảnh đa mức xám, và xem bức ảnh này như một ma trận kích thước $n \times n$, mỗi phần tử của ma trận này tương ứng với 1 pixel, giá trị của mỗi pixel này là một số nguyên nằm trong khoảng $[0, 255]$ → để biểu diễn một pixel này cần 8 bit, nên để lưu cả bức ảnh cần $8 \times n^2$ bit

- Vấn đề với một bức ảnh với độ phân giải cao thì sẽ rất tốn dung lượng lưu trữ, nhưng trong nhiều hoàn cảnh, không cần ảnh với độ phân giải gốc, có thể chấp nhận mất mát thông tin qua một số phép xấp xỉ → chẳng hạn nén ảnh bằng ***truncated SVD***

- Trong ma trận Σ, các giá trị trên đường chéo là không âm và giảm dần $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0 = \cdots = 0$, thông thường chỉ có một số σ_i là có giá trị lớn, còn lại nhỏ và gần 0. khi đó có thể xấp xỉ mà trận Σ bởi tổng $k < r$ ma trận có rank bằng 1.

$$\mathbf{A} \approx \mathbf{A}_k = \mathbf{U}_k \Sigma_k (\mathbf{V}_k)^T = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^2 + \cdots + \sigma_k \mathbf{u}_k \mathbf{v} k^T \qquad \frac{||\mathbf{A} - \mathbf{A}_k||_F^2}{||\mathbf{A}||_F^2} = \frac{\sum_{i=k+1}^{r} \sigma_i^2}{\sum_{j=1}^{r} \sigma_j^2}$$

- Sai số do xấp xỉ càng nhỏ nếu phần singular values bị *truncated* có giá trị càng nhỏ so với phần singular values được giữ lại → chọn k phù hợp với lượng thông tin cần giữ lại → cần $8 \times k \times n$ bit để lưu trữ

Ma trận xác định dương

❑ Chỉ xét trên họ các ma trận đối xứng

- Một ma trận đối xứng A bậc n được gọi là ***xác định dương*** nếu:

$$x^T A x > 0 \; \forall x \neq 0$$

- Một ma trận đối xứng A bậc n được gọi là ***bán xác định dương*** nếu:

$$x^T A x \geq 0 \; \forall x \neq 0$$

❑ Tính chất:

- Một ma trận đối xứng xác định dương đều:
  - Có mọi giá trị riêng đều là một số thực dương
  - khả nghịch, hơn nữa định thức của nó là một số dương
- $A = B^T B$ là ma trận bán xác định dương với mọi ma trận B bất kỳ
- Với A là một ma trận bán xác định dương thì:
  - tồn tại duy nhất ma trận B bán xác định dương sao cho: $B^2 = A$ và B được gọi là căn bậc hai của A
  - A có thể biểu diễn duy nhất dưới dạng LL^T trong đó L là một ma trận tam giác dưới với các phần tử trên đường chéo là dương
  - SVD của A có thể biểu diễn duy nhất dưới dạng: $V^T D V$ trong đó V là một ma trận trực giao
- Ma trận xác định dương được ứng dụng rất nhiều trong các bài tối ưu lồi

Chuẩn của vector và ma trận

- ***Chuẩn*** là một đại lượng dùng để đo khoảng cách trong không gian nhiều chiều

- **Định nghĩa: *chuẩn*** là một hàm số f thoả mãn đồng thời các tính chất sau:

  - $f(x) \geq 0 \; \forall x$. Dấu bằng xảy ra khi và chỉ khi $x = 0$
  - $f(ax) = |a| f(x) \; \forall a \in R$
  - $f(x + y) \leq f(x) + f(y) \;\; \forall x, y$

Chuẩn của vector và ma trận

❑ Chuẩn của vector

Cho $x \in R^n$, ta có họ $p - norms$ hoặc $l_p - norms$ được định nghĩa như sau:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$$

- Với $p = 1$: $l_1 - norm$

$$|x| = |x_1| + |x_2| + \cdots + |x_n|$$

- Với $p = 2$: $l_2 - norm$ hoặc Euclidean norm:

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{x^T x}$$

- Với $p = \infty$: $l_\infty - norm$

$$\|x\|_\infty = \max_{1 \le i \le n} |x_i|$$

Chuẩn của vector và ma trận

❑ Chuẩn của ma trận

Xét ma trận $A \in R^{m \times n}$, chuẩn của A được định nghĩa thông qua chuẩn của vector như sau:

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p = 1} \|Ax\|_p$$

- Với $l_1 - norm$: $\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|$
- Với $l_2 - norm$ hoặc Euclidean norm: $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{1 \le i \le n} \sqrt{\lambda_i(A^T A)}$
- Với $l_\infty - norm$: $\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|$
- Thường dùng nhất là ***chuẩn Frobenius***, ký hiệu là $\|A\|_F$ là căn bậc hai của tổng bình phương tất cả các phần tử của ma trận đó:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$$

Vietnam Institute for Advanced Study in Mathematics

Vết của ma trận

- **Định nghĩa:** Vết của một ma trận vuông là tổng tất cả các phần tử trên đường chéo chính của nó, được ký hiệu là trace(A).
- **Tính chất**:

- $trace(A) = trace(A^T)$
- $trace(\sum_{i=1}^{k} A_i) = \sum_{i=1}^{k} trace(A_i)$
- $trace(A) = \sum_{i=1}^{n} \lambda_i$ với λ_i là các giá trị riếng của A
- $trace(AB) = trace(BA)$
- Nếu X là một ma trận khả nghịch cùng chiều với ma trận vuông A thì: $trace(XAX^{-1}) = trace(X^{-1}XA) = trace(A)$
- $trace(A^TA) = trace(AA^T) = ||A||_F^2 \geq 0$ với A là ma trận bất kỳ

PHẦN 2
GIẢI TÍCH

Tổng quan

- Xem xét các vấn đề liên quan đến hàm số

- Các đặc tính quan trọng của một hàm số: tính liên tục, tính khả vi, tính đồng biến / nghịch biến, tính lồi/lõm (convex/concave)…

- Cụ thể trong lĩnh vực Học máy, rất nhiều bài toán đưa quy về bài toán tối ưu của một hàm mục tiêu nào đó, để giải được bài toán này cần dựa vào các thuật toán tối ưu dựa trên đạo hàm, hoặc xấp xỉ

Tổng quan

❑ **Một số khái niệm cơ bản:**

- Với hàm một biến, xét hàm số $f: R \to R$, khi đó:

f được gọi là liên tục tại $a \in R$ iff:

$$\lim_{x \to a} f(x) = f(a)$$

f được gọi là khả vi tại $a \in R$ iff giới hạn:

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

là tồn tại hữu hạn, và giá trị hữu hạn đó được ký hiệu là $f'(a)$

- Với hàm nhiều biến, các định nghĩa trên được áp dụng cho từng biến khi khác biến còn lại được cố định

Đạo hàm của hàm nhiều biến

❑ Hàm cho giá trị là một số vô hướng

Đạo hàm (gradient) của một hàm số: $f(x): R^n \rightarrow R$ theo vector x được định nghĩa như sau:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) \triangleq \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n$$

Trong đó $\frac{\partial f(x)}{\partial x_i}$ là đạo hàm của hàm số theo thành phần thứ i của vector x. Đạo hàm này được lấy khi giả sử tất cả các biến còn lại là hằng số

Đạo hàm của hàm nhiều biến

❑ **Hàm cho giá trị là một số vô hướng**

Đạo hàm bậc hai (second-order gradient) của hàm số trên còn được gọi là Hessian và được định nghĩa như sau:

$$\nabla^2 f(\mathbf{x}) \triangleq \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix} \in \mathbb{S}^n$$

Với $S^n \in R^{n \times n}$ là tập các ma trận vuông đối xứng $n \times n$

Đạo hàm của hàm nhiều biến

❑ Hàm cho giá trị là một số vô hướng

Đạo hàm của một hàm số $f(X)$: $R^{n\times m} \to R$ theo ma trận X được định nghĩa là:

$$\nabla^2 f(\mathbf{x}) \triangleq \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix} \in \mathbb{S}^n$$

Đạo hàm của hàm nhiều biến

❑ Hàm cho giá trị là một vector

Giả sử một hàm số với đầu vào là một số thực $v(x): R \to R^n$:

$$v(x) = \begin{bmatrix} v_1(x) \\ v_2(x) \\ \vdots \\ v_n(x) \end{bmatrix}$$

Đạo hàm bậc nhất và bậc hai của nó là một vector hàng như sau:

$$\nabla v(x) \triangleq \begin{bmatrix} \frac{\partial v_1(x)}{\partial x} & \frac{\partial v_2(x)}{\partial x} & \dots & \frac{\partial v_n(x)}{\partial x} \end{bmatrix}$$

$$\nabla^2 v(x) \triangleq \begin{bmatrix} \frac{\partial^2 v_1(x)}{\partial x^2} & \frac{\partial^2 v_2(x)}{\partial x^2} & \dots & \frac{\partial^2 v_n(x)}{\partial x^2} \end{bmatrix}$$

Đạo hàm của hàm nhiều biến

❑ Hàm cho giá trị là một vector

Nếu đầu vào cũng là một vector, tức có hàm số $h(x): R^k \rightarrow R^n$ thì đạo hàm của nó là một ma trận kxn

$$\nabla h(\mathbf{x}) \triangleq \begin{bmatrix} \frac{\partial h_1(\mathbf{x})}{\partial x_1} & \frac{\partial h_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial h_n(\mathbf{x})}{\partial x_1} \\ \frac{\partial h_1(\mathbf{x})}{\partial x_2} & \frac{\partial h_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial h_n(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_1(\mathbf{x})}{\partial x_k} & \frac{\partial h_2(\mathbf{x})}{\partial x_k} & \cdots & \frac{\partial h_n(\mathbf{x})}{\partial x_k} \end{bmatrix}$$

$$= \left[\nabla h_1(\mathbf{x})\ \nabla h_2(\mathbf{x}) \ldots \nabla h_n(\mathbf{x})\right] \in \mathbf{R}^{k \times n}$$

Đạo hàm của hàm nhiều biến

❑ Tính chất quan trọng

Để cho tổng quát, ta giả sử biến đầu vào là một ma trận và các hàm số có chiều phù hợp để các phép nhân thực hiện được

- ***Product rules:***

Với hàm biến một chiều:

$$\big(f(x)g(x)\big)' = f'(x)g(x) + f(x)g'(x)$$

Với hàm biến nhiều chiều:

$$\nabla\big(f(X)^T g(X)\big) = \big(\nabla f(X)\big)g(X) + \big(\nabla g(X)\big)f(X)$$

- ***Chain rules:***

Với hàm biến một chiều:

$$\Big(f\big(g(x)\big)\Big)' = g'(x)f'_g(g(x))$$

Với hàm biến nhiều chiều:

$$\nabla_X f\big(g(X)\big) = \nabla_X g^T \nabla_f f$$

Đạo hàm của hàm nhiều biến

❑ Bảng các đạo hàm thường gặp:

$f(\mathbf{x})$	$\nabla f(\mathbf{x})$
$\mathbf{a}^T\mathbf{x}$	$\mathbf{a}$
$\mathbf{x}^T\mathbf{A}\mathbf{x}$	$\mathbf{A} + \mathbf{A}^T)\mathbf{x}$
$\mathbf{x}^T\mathbf{x} = \\|\mathbf{x}\\|_2^2$	$2\mathbf{x}$
$\\|\mathbf{A}\mathbf{x} - \mathbf{b}\\|_2^2$	$2\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$
$\mathbf{a}^T\mathbf{x}^T\mathbf{x}\mathbf{b}$	$2\mathbf{a}^T\mathbf{b}\mathbf{x}$
$\mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{b}$	$(\mathbf{a}\mathbf{b}^T + \mathbf{b}\mathbf{a}^T)\mathbf{x}$

Đạo hàm theo vector

$f(\mathbf{X})$	$\nabla f(\mathbf{X})$
$\\|\mathbf{X}\\|_F^2$	$2\mathbf{X}$
$\mathbf{A}\mathbf{X}$	$\mathbf{A}^T$
$\\|\mathbf{A}\mathbf{X} - \mathbf{B}\\|_F^2$	$2\mathbf{A}^T(\mathbf{A}\mathbf{X} - \mathbf{B})$
$\\|\mathbf{X}\mathbf{A} - \mathbf{B}\\|_F^2$	$2(\mathbf{X}\mathbf{A} - \mathbf{B})\mathbf{A}^T$
$\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{b}$	$\mathbf{X}(\mathbf{a}\mathbf{b}^T + \mathbf{b}\mathbf{a}^T)$
$\mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{b}$	$(\mathbf{a}\mathbf{b}^T + \mathbf{b}\mathbf{a}^T)\mathbf{X},$
$\mathbf{a}^T\mathbf{Y}\mathbf{X}^T\mathbf{b}$	$\mathbf{b}\mathbf{a}^T\mathbf{Y}$
$\mathbf{a}^T\mathbf{Y}^T\mathbf{X}\mathbf{b}$	$\mathbf{Y}\mathbf{a}\mathbf{b}^T$
$\mathbf{a}^T\mathbf{X}\mathbf{Y}^T\mathbf{b}$	$\mathbf{a}\mathbf{b}^T\mathbf{Y}$
$\mathbf{a}^T\mathbf{X}^T\mathbf{Y}\mathbf{b}$	$\mathbf{Y}\mathbf{b}\mathbf{a}^T$

Đạo hàm theo ma trận

Khai triển Taylor

- Khai triển Tayor là một công cụ rất hữu dụng trong nhiều bài toán liên quan đến hàm số mà trong đó phải sử dụng đến đạo hàm của hàm số. Về bản chất, khai triển Taylor cho phép xấp xỉ một hàm số với một đa thức có bậc tuỳ ý.
- **Khai triển Taylor cho hàm một biến:**

Taylor's theorem : Gọi $f : R \to R$ là một hàm liên tục có đạo hàm tới bậc k tại x với k là một số nguyên và ϵ là một số thực đủ nhỏ hơn 1. Ta có:

$$f(x+\epsilon) = f(x) + f'(x)\epsilon + \frac{f''(x)}{2!}\epsilon^2 + \ldots + \frac{f^{(k)}(x)}{k!}\epsilon^k + o(\epsilon^k) \qquad (8)$$

Ở đây, $f^{(k)}(x)$ là đạo hàm cấp k của f tại x.

Trong đó o có nghĩa là: $\lim_{\epsilon \to 0} \frac{o(\epsilon^k)}{\epsilon^k} = 0$

Khai triển Taylor

❑ **Khai triển Taylor cho hàm nhiều biến:**

Taylor's theorem : Gọi $f : \mathbb{R}^d \to R$ là một hàm liên tục có đạo hàm tới cấp 2 tại $\mathbf{x}$ và ϵ là một số thực đủ nhỏ hơn 1. Với mọi vector cố định cho trước $\mathbf{y}$, ta có:

$$\begin{aligned} f(\mathbf{x}+\epsilon\mathbf{y}) &= f(\mathbf{x}) + \epsilon\langle\nabla f(\mathbf{x}), \mathbf{y}\rangle + o(\epsilon) \\ f(\mathbf{x}+\epsilon\mathbf{y}) &= f(\mathbf{x}) + \epsilon\langle\nabla f(\mathbf{x}), \mathbf{y}\rangle + \epsilon^2\frac{1}{2}\mathbf{y}^T\nabla^2 f(\mathbf{x})\mathbf{y} + o(\epsilon^2) \end{aligned} \qquad (14)$$

→ Khai triển Taylor là cơ sở lý thuyết cho rất nhiều thuật toán tối ưu bằng cách xấp xỉ, trong đó điển hình là ***Gradient descent*** và ***Newton step***

Khai triển Taylor

❑ Sử dụng xấp xỉ kiểm tra đạo hàm

- Việc tính toán đạo hàm của hàm nhiều biến thường khá phức tạp, trong trường hợp tính được ***dạng đóng (closed form)*** của đạo hàm, để kiểm tra tính toán có đúng không, có thể dùng một phép xấp xỉ
- Theo định nghĩa:

$$f'(x) = \lim_{e \to 0} \frac{f(x+e) - f(x)}{e} = \lim_{e \to 0} \frac{f(x+e) - f(x-e)}{2e}$$

Bằng việc chọn $e = 10^{-6}$ là một số đủ nhỏ, có thể xấp xỉ:

$f'(x) \approx \frac{f(x+e)-f(x)}{e}$ (1) và $f'(x) \approx \frac{f(x+e)-f(x-e)}{2e}$ (2)

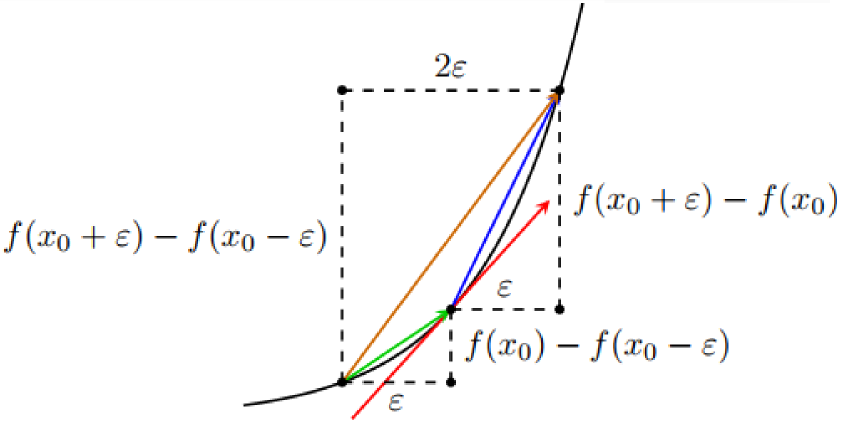
Nếu closed form tính được là đúng gì $f'(x)$ phải gần với các xấp xỉ trên

Vấn đề là xấp xỉ này sai số càng nhỏ thì càng tốt !?

Khai triển Taylor

❑ Sử dụng xấp xỉ kiểm tra đạo hàm

- Theo khai triển Taylor:



$$f(x+e) = f(x) + f'(x)e + \frac{f''(x)}{2!}e^2 + o(e^2)$$
$$f(x-e) = f(x) - f'(x)e + \frac{f''(x)}{2!}e^2 + o(e^2)$$

Do đó:

$$\frac{f(x+e)-f(x)}{e} = f'(x) + o(e) \text{ và } \frac{f(x+e)-f(x-e)}{2e} = f'(x) + o(e^2)$$

Khi e đủ nhỏ thì $o(e^2) \ll o(e)$ tức là sai số của phép xấp xỉ thứ (2) là nhỏ hơn và là tốt hơn

Cách tính bằng xấp xỉ này được gọi là ***numerical gradient***

PHẦN 3
XÁC SUẤT CƠ BẢN

Sự kiện và xác suất

- **Định nghĩa 1:** Một ***không gian xác suất*** bao gồm 3 thành phần:
  - Một không gian mẫu Q: là một tập các kết quả có thể của một quá trình ngẫu nhiên được mô hình hoá bởi không gian xác suất đó.
  - Sự kiện: mỗi sự kiện có thể được coi là một tập con của Q. Tập các sự kiện được kí hiệu là F.
  - Một hàm xác suất: Pr: F → R thoả mãn những điều kiện sau:
    - Với mỗi sự kiện E: $0 \leq \Pr[E] \leq 1$
    - $\Pr[Q] = 1$
    - Với một tập hữu hạn hoặc đếm được các sự kiện $E_1, E_2, \ldots,$ đôi một không giao nhau: $\Pr[\cup_{i \geq 1} E_i] = \sum_{i \geq 1} \Pr[E_i]$

Sự kiện và xác suất

- **Bổ đề 1:** Cho hai sự kiện E_1, E_2 bất kỳ:

$$\Pr[E_1 \cup E_2] = \Pr[E_1] + \Pr[E_2] - \Pr[E_1 \cap E_2]$$

- **Bổ đề 2:** Cho một tập hữu hạn hoặc đếm được các sự kiện $E_1, E_2, \ldots$ bất kỳ:

$$\Pr[\cup_{i \geq 1} E_i] \leq \sum_{i \geq 1} \Pr[E_i]$$

- **Bổ đề 3:** Nguyễn lý bù trừ

Cho một tập n sự kiện $E_1, E_2, \ldots, E_n$ bất kỳ:

$$\begin{aligned}
\Pr[\cup_{i \geq 1} E_i] &= \sum_{i \geq 1} \Pr[E_i] - \sum_{i<j} \Pr[E_i \cap E_j] \\
&+ \sum_{i<j<k} \Pr[E_1 \cap E_j \cap E_k] \\
&- \ldots + (-1)^{\ell+1} \sum_{i_1 < i_2 < \ldots < i_\ell} \Pr[\cap_{r=1}^{\ell} E_{i_r}]
\end{aligned}$$

Sự kiện và xác suất

- **Định nghĩa 2:** Hai sự kiện E_1, E_2 được gọi là ***độc lập*** nếu:
$$\Pr[E_1 \cap E_2] = \Pr[E_1] . \Pr[E_2]$$

Tương tự như vậy, các sự kiện $E_1, E_2, \ldots, E_n$ được gọi là độc lập nếu: $\Pr[\bigcap_{i=1}^{n} E_i] = \prod_{i=1}^{n} \Pr[E_i]$

- **Định nghĩa 3:** ***Xác suất có điều kiện*** của một sự kiện E khi biết sự kiện F xảy ra là:
$$\Pr[E|F] = \frac{\Pr[E \cap F]}{\Pr[F]}$$

Sự kiện và xác suất

Một luật rất quan trọng để tính xác suất là ***luật tổng xác suất và luật Bayes***:

- **Định lý 1 (Law of total probability)**: Gọi $E_1, E_2, \ldots, E_n$ là các sự kiện đôi một không giao nhau trong một không gian mẫu Q thoả mãn $\bigcup_{i=1}^{n} E_i = Q$, ta có:

$$\Pr[B] = \sum_{i=1}^{n} \Pr[B \cap E_i] = \sum_{i=1}^{n} \Pr[B|E_i] \Pr[E_i]$$

- **Định lý 2 (Bayes' Law)**: Gọi $E_1, E_2, \ldots, E_n$ là các sự kiện đôi một không giao nhau trong một không gian mẫu Q thoả mãn $\bigcup_{i=1}^{n} E_i = Q$, ta có:

$$\Pr[E_j|B] = \frac{\Pr[E_j \cap B]}{\Pr[B]} = \frac{\Pr[B|E_j] \Pr[E_j]}{\sum_{i=1}^{n} \Pr[B|E_i] \Pr[E_i]}$$

Biến ngẫu nhiên

- **Định nghĩa 4:** ***Biến ngẫu nhiên*** (đại lượng ngẫu nhiên) là một đại lượng mà giá trị của nó là ngẫu nhiên, phụ thuộc vào kết quả phép thử.
  - Biến ngẫu nhiên được gọi là ***rời rạc***, nếu tập giá trị của nó là một tập hữu hạn hoặc vô hạn đếm được các phần tử
  - Biến ngẫu nhiên được gọi là ***liên tục***, nếu tập giá trị của nó lấp kín một khoảng hoặc một số khoảng của trục số hoặc cũng có thể là cả trục số.

Hàm xác suất

❑ Định nghĩa 5:

- Đối với biến ngẫu nhiên rời rạc X, mỗi giá trị x_i của nó gắn với một xác suất $p_i = P(X = x_i)$.
- Hàm số $p(x) = P(X = x)$, với mọi x thuộc tập giá trị của X, được gọi là ***hàm xác suất*** của X. Hàm số này có thể được mô tả bởi bảng sau:

$X = x$	x_1	x_2	$\ldots$	x_n	$\ldots$
$P(X = x)$	p_1	p_2	$\ldots$	p_n	$\ldots$

Tính chất:

- $p(x) \geq 0 \ \forall x$
- $\sum_{\forall x} p(x) = 1$

Hàm phân phối xác suất

- **Định nghĩa 6:** ***Hàm phân phối xác suất*** của biến ngẫu nhiên X, kí hiệu là F(x) và được xác định như sau:

$$F(x) = P(X < x)$$

Hàm phần phối xác suất phản ánh mức độ tập trung xác suất ở bên trái điểm x

Tính chất:

- $0 \leq F(x) \leq 1$
- $\lim_{x \to -\infty} F(x) = 0,\ \lim_{x \to +\infty} F(x) = 1$
- $F(x)$ là hàm không giảm: $\forall a < b, F(a) \leq F(b)$
- $P(a \leq X < b) = F(b) - F(a)$
- Nếu X là biến ngẫu nhiên rời rạc thì $F(x)$ còn được gọi là hàm phân phối tích luỹ:

$$F(x) = P(X < x) = \sum_{i: x_i < x} P(X = x_i) = \sum_{i: x_i < x} p_i$$

Hàm mật độ xác suất

Với biến ngẫu nhiên liên tục, hàm phân phối xác suất có một hạn chế là không cho biết rõ phân phối xác suất ở lân cận một điểm nào đó trên trục số.

- **Định nghĩa 7:** Hàm mật độ xác suất $f(x)$ của biến ngẫu nhiên liên tục X, có hàm phân phối xác suất $F(x)$ khả vi, được định nghĩa thông qua đẳng thức:

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)dt$$

Dễ thấy $f(x) = F'(x)$

Tính chất:

- $f(x) \geq 0\ \forall x$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$
- $P(a \leq X \leq b) = \int_a^b f(x)dx$ (có thể bỏ các dấu "=")

Các tham số đặc trưng

❑ Kỳ vọng:

- Kỳ vọng của biến ngẫu nhiên X, ký hiệu là $E(X)$ được định nghĩa như sau:
  - Nếu X là biến ngẫu nhiên rời rạc có hàm xác suất $p(x_i) = p_i$ thì:

$$E(X) = \sum_{\forall i} x_i p_i$$

  - Nếu X là biến ngẫu nhiên liên tục có hàm mật độ xác suất $f(x)$ thì:

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

- ***Tính chất:***
  - Kỳ vọng là tổng có trọng số các giá trị của X, khác với trung bình cộng các giá trị
  - E(c)=c với c là hằng số
  - E(aX)=aEX với a là hằng số
  - E(X+Y)=EX+EY với X, Y là hai biến ngẫu nhiên bất kỳ
  - E(XY)=EX.EY nếu X, Y là hai biến ngẫu nhiên độc lập
  - Nếu $Y = \varphi(X)$ thì tương ứng với X là biến ngẫu nhiên rời rạc hay liên tục ta có:

$E(Y) = \sum_{\forall i} \varphi(x_i) p_i$ và $E(Y) = \int_{-\infty}^{+\infty} \varphi(x) f(x) dx$

Các tham số đặc trưng

❑ Phương sai:

- Phản ánh mức độ phân tán của các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình của nó là kỳ vọng.
- Phương sai của biến ngẫu nhiên X, ký hiệu là $V(X)$ được định nghĩa như sau:

$$VX = E(X - EX)^2 = E(X^2) - (EX)^2$$

  - Nếu X là biến ngẫu nhiên rời rạc có hàm xác suất $p(x_i) = p_i$ thì:

$$V(X) = \sum_{\forall i} x_i^2 p_i - \left(\sum_{\forall i} x_i p_i\right)^2$$

  - Nếu X là biến ngẫu nhiên liên tục có hàm mật độ xác suất $f(x)$ thì:

$$V(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - \left(\int_{-\infty}^{+\infty} xf(x)dx\right)^2$$

- ***Tính chất:***
  - Vc=0 với c là hằng số
  - V(aX)=a^2VX với a là hằng số
  - V(X+b)=VX
  - V(X+Y)=VX+VY nếu X, Y là hai biến ngẫu nhiên độc lập

Các tham số đặc trưng

❑ Độ lệch chuẩn:

- Độ lệch chuẩn của biến ngẫu nhiên X, ký hiệu là $\sigma(X)$ được định nghĩa như sau:

$$\sigma(X) = \sqrt{V(X)}$$

- Độ lệch chuẩn có cùng đơn vị đo với biến ngẫu nhiên

❑ Mode

- Mode của biến ngẫu nhiên X, kí hiệu là $mod(X)$, là giá trị của biến ngẫu nhiên X có khả năng xuất hiện lớn nhất trong một lân cận nào đó của nó.
- Đối với biến ngẫu nhiên rời rạc, $mod(X)$ là giá trị của X ứng với xác suất lớn nhất.
- Như vậy một biến ngẫu nhiên có thể có một mode hoặc nhiều mode.

Các tham số đặc trưng

❑ Hiệp phương sai (*Covariance*)

Giả sử X, Y là các biến ngẫu nhiên, hiệp phương sai của X và Y được ký hiệu là μ_{XY}, và được xác định bởi:

$$\mu_{XY} = E[(X - EX)(Y - EY)] = E(XY) - EX.EY$$

Trong đó $E(XY)$ được xác định theo công thức:

$$E(XY) = \begin{cases} \sum_i \sum_j x_i y_j p_{ij}, & \text{đối với biến ngẫu nhiên rời rạc} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy.f(x,y), & \text{đối với biến ngẫu nhiên liên tục} \end{cases}$$

- Nếu X, Y độc lập thì $E(XY) = E(X)E(Y) \leftrightarrow \mu_{XY} = 0$, điều ngược lại không đúng

Các tham số đặc trưng

❑ Hệ số tương quan (*Correlation*)

Giả sử X, Y là các biến ngẫu nhiên, hệ số tương quan của X và Y được ký hiệu là ρ_{XY}, xác định bởi:

$$\rho_{XY} = \frac{\mu_{XY}}{\sigma_X \sigma_Y} = \frac{\mu_{XY}}{\sqrt{VX.VY}}$$

- Có thể chứng minh được $|\rho_{XY}| \leq 1$
- Nếu $\rho_{XY} = \pm 1$ ta nói X và Y có ***tương quan tuyến tính***, tức là tồn tại a, b sao cho: $Y = aX + b$
- Nếu $\rho_{XY} = 0 \leftrightarrow \mu_{XY} = 0$, ta nói X và Y là ***không tương quan***
- Nếu $\rho_{XY} > 0$, thì X và Y có ***tương quan dương*** (tương quan đồng biến)
- Nếu $\rho_{XY} < 0$, thì X và Y có ***tương quan âm*** (tương quan nghịch biến)
- *Hai biến độc lập thì không tương quan, chiều ngược lại không đúng*
- *Hai biến tương quan thì phụ thuộc (không độc lập)*

Các tham số đặc trưng

❑ Ma trận hiệp phương sai

- Giả sử X, Y là các biến ngẫu nhiên. Khi đó ***ma trận hiệp phương sai*** của X và Y được xác định bởi:

$$T = \begin{bmatrix} \mu_{XX} & \mu_{XY} \\ \mu_{YX} & \mu_{YY} \end{bmatrix} = \begin{bmatrix} VX & \mu_{XY} \\ \mu_{YX} & VY \end{bmatrix}$$

Trên đường chéo chính là các phương sai của từng thành phần và do $\mu_{XY} = \mu_{YX}$ nên ma trận này là ma trận đối xứng

- Giả sử $X \in R^m$, $Y \in R^n$ là các vetor ngẫu nhiên. Khi đó ma trận hiệp phương sai của X và Y là một ma trận kích thước $m \times n$, trong đó phần tử (i, j) của ma trận này chính là $\mu_{X_i Y_j}$

Khi $X = Y$, ta còn gọi ma trận này là ***variance-covariance*** của X, vì khi đó các giá trị trên đường chéo chính là các phương sai

Ước lượng các tham số đặc trưng

Trong thực tế, chúng ta thường có một bộ dữ liệu gồm N điểm vector $x_1, x_2, \dots, x_N$. Để biết được đặc tính phân bố của các điểm dữ liệu này, có thể ***ước lượng các tham số đặc trưng*** như sau:

- Vector kỳ vọng: $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$
- Ma trận hiệp phương sai: $S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T = \frac{1}{N}\hat{X}\hat{X}^T$

Trong đó $\hat{X}$ được tạo bằng cách trừ mỗi cột của $X = [x_1, x_2, \dots, x_N]$ đi $\bar{x}$

- Mọi phần tử trên đường chéo của ma trận S là phương sai của từng chiều dữ liệu
- Các phần tử nằm ngoài đường chéo thể hiện sự tương quan giữa các thành phần của dữ liệu, chính là hiệp phương sai

Một số phân phối xác suất thường gặp

❑ Phân phối Bernoulli:

Phân phối Bernoulli là một ***phân phối rời rạc*** mô tả biến ngẫu nhiên mà đầu ra chỉ nhận một trong hai giá trị 0, 1.

Phân phối Bernoulli được mô tả bằng một tham số $p \in [0,1]$ và là xác suất để biến ngẫu nhiên $X = 1$: $p(X = 1) = p$, $p(X = 0) = 1 - p$

- Hàm xác suất: $p(x) = p^x(1-p)^{1-x}$
- Kỳ vọng: $E(X) = p$
- Phương sai: $V(X) = p(1-p)$

❑ Phân phối nhị thức (*Binomial Distribution*)

Phân phối nhị thức mô tả biến ngẫu nhiên trong n phép thử Bernoulli, cụ thể một biến ngẫu nhiên X nhận giá trị trong tập $\{1,2,\dots,n\}$ với hàm xác suất theo công thức Bernoulli:

$$p(X = k) = C_n^k p^k (1-p)^{n-k} \text{ với } 1 \le k \le n, p \in (0,1)$$

Khi đó X được gọi là tuân theo phân phối nhị thức với tham số n, p. Ký hiệu $X \sim B(n,p)$

- Kỳ vọng: $E(X) = np$
- Phương sai: $V(X) = np(1-p)$

Một số phân phối xác suất thường gặp

❑ **Phân phối Categorical:**

Trong nhiều trường hợp, đầu ra của biến ngẫu nhiên rời rạc có thể là K đầu ra, phân phối Categorical sẽ được mô tả bởi K tham số, viết dưới dạng vector: $p = [p_1, p_2, \dots, p_K]$ với p_k là các số không âm và có tổng bằng 1

$$p(x = k) = p_k$$

Một số phân phối xác suất thường gặp

❑ Phân phối Chuẩn:

- Tổng quát với biến ngẫu nhiên D chiều. Có hai tham số mô tả phân phối này là: vecto kỳ vọng $\mu \in R^D$ và ma trận hiệp phương sai $\Sigma \in S^D$ là một ma trận đối xứng xác định dương:

- Hàm mật độ xác suất có dạng:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Một số phân phối xác suất thường gặp

❑ Phân phối Beta:

- Phân phối Beta là một phần phối liên tục được định nghĩa trên một biến ngẫu nhiên $\lambda \in [0,1]$, được dung để mô tả sự biến động của tham số λ trong phân phối Bernoulli.
- Phân phối Beta được mô tả bởi hai tham số dương: α, β.
- Hàm mật độ xác suất là:

$$p(\lambda) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \lambda^{\alpha-1}(1-\lambda)^{\beta-1}$$

Với hàm số Gama: $$\Gamma(z) = \int_0^\infty t^{z-1}\exp(-t)dt$$

Một số phân phối xác suất thường gặp

❑ Phân phối Dirichlet:

- Phân phối Dirichlet là trường hợp tổng quát của phân phối Beta khi được dung để mô tả tham số của phần phối Categorical.
- Phân phối Dirichlet được định nghĩa trên K biến liên tục $\lambda_1, \lambda_2, \dots, \lambda_k$ với λ_k là các số không âm và có tổng bằng 1.
- Có K tham số dương để mô tả phân phối Dirichlet là: $\alpha_1, \alpha_2, \dots, \alpha_k$
- Hàm mật độ xác suất có dạng:

$$p(\lambda_1, \dots, \lambda_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1}$$

5 công cụ của xác suất

❑ Tính tuyến tính của kỳ vọng (1)

Gọi $X_1, X_2, \dots, X_n$ là n biến ngẫu nhiên trong cùng một không gian xác suất. Gọi $X = \sum_{i=1}^{n} X_i$, ta có:

$$E(X) = \sum_{i=1}^{n} EX_i$$

❑ Union bound (2) (bỏ)

Union Bound: Gọi $E_1, E_2, \dots E_n$ là tập n sự kiện, ta có:

$$\Pr[\text{ ít nhất một sự kiện } E_i \text{ xảy ra}] \leq \sum_{i=1}^{n} \Pr[E_i] \qquad (16)$$

5 công cụ của xác suất

❑ Bất đẳng thức Markov (3)

Cho một biến ngẫu nhiên X không âm, với mọi hằng số $c > 1$, ta có:

$$\Pr[X > cE(X)] \leq \frac{1}{c}$$

Ví dụ: trong bài giải thuật Quicksort ngẫu nhiên, ta tính được kỳ vọng thời gian của Quiksort khi chọn pivot ngẫu nhiên là:

$$E[T(n)] = O(nlogn).$$

Theo bất đẳng thức Markov, xác suất để giải thuật này chạy lâu hơn 10 lần $E[T(n)]$ là không quá 10%

5 công cụ của xác suất

❑ Bất đẳng thức Chebyshev (4)

Cho một biến ngẫu nhiên X, với mọi hằng số $c > 1$, ta có:

$$\Pr[|X - EX| \geq c.\sigma(X)] \leq \frac{1}{c^2}$$

❑ Bất đẳng thức Chernoff (5)

Chernoff Bounds: Giả sử $X = \sum_{i=1}^{n} X_i$ trong đó $X_1, X_2, \ldots, X_n$ là n biến ngẫu nhiên độc lập có giá trị trong đoạn $[0,1]$. Gọi $\mu = E[X]$. Ta có:

- Với mọi $\delta > 0$,

$$\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu} \qquad (10)$$

- Với mọi $\delta \in (0,1)$,

$$\Pr[X \geq (1-\delta)\mu] \leq e^{-\delta^2\mu/2} \qquad (11)$$

Tài liệu tham khảo

- *Linear algebra refresher course*, Stanford University
- *Xác suất thống kê* – Tống Đình Quỳ
- *Đại số tuyến tính* – Nguyễn Hữu Việt Hưng
- *Machine Learning cơ bản* - Vũ Hữu Tiệp
- *Giải thuật và lập trình* - Lê Việt Hùng
- *Convex Optimization* – Boyd and Vandenberghe
- *Matrix calculus,* Stanford University
- *Matrix Cookbook,* Kaare Brandt Petersen
- *Bài giảng Application of random matrices* – Vũ Hà Văn