Johns Hopkins University
EN. 601.475 Machine Learning

Probability & Linear Algebra Review

Xuan Zhang

Sep. 2. 2022

I. Probability

- Random Variables
- Bayes' Rule
- Bernoulli distributions & Sigmoid function
- Categorical distributions & Softmax function
- Gaussian distribution

II. Bayes Optimal Rule

III. Linear Algera

- Vector norms
- matrix multiplication
- vector derivatives

I. Probability

Probability

two interpretations:

- frequentist

  probabilities represent long run frequencies of events

- Bayesian (this course)

  probability is used to quantify our uncertainty about sth.

  ♥ can be used to model uncertainty about one-off events

Probability

$\Pr(\underset{\text{event}}{A})$: probability that A is true

$\in [0,1]$

joint probability: $\Pr(A, B)$

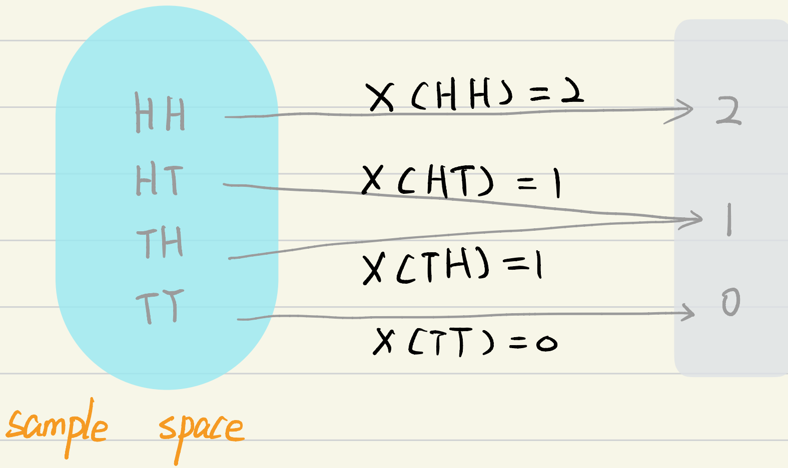
if A & B are independent: $\Pr(A,B) = \Pr(A)\Pr(B)$

conditional probability: $\Pr(B \mid A) \triangleq \dfrac{\Pr(A,B)}{\Pr(A)}$

conditionally independent $A \perp B \mid C$

$\Pr(A, B \mid C) = \Pr(A \mid C)\Pr(B \mid C)$

Random Variable (RV)



X: number of heads

RVs are functions. RV is a numeric function of the outcome.

Random Variable (RV)

- discrete RV :

  sample space is finite or countably infinite

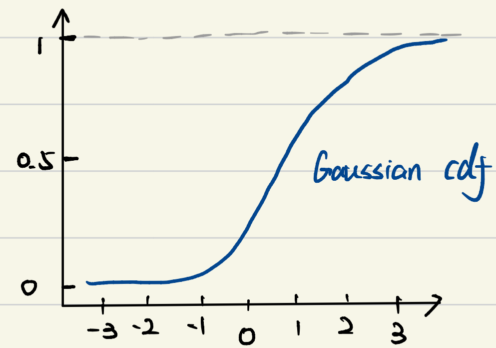
- continuous RV :

  infinite number of values between two values

Cumulative distribution function (cdf)

$$P(x) \triangleq \Pr(X \leq x)$$

e.g. $\Pr(a < X \leq b) = P(b) - P(a)$



cdf's are monotonically non-decreasing functions.

Random Variable (RV)

probability mass function (pmf)

(discrete RV)

$$p(x) \triangleq \Pr(X = x)$$

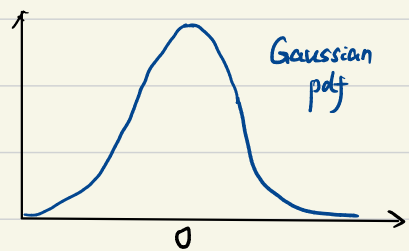
$$\sum_{x \in \mathcal{X}} p(x) = 1$$

probability density function (pdf)

(continuous RV)

$$p(x) \triangleq \frac{d}{dx} p(x)$$

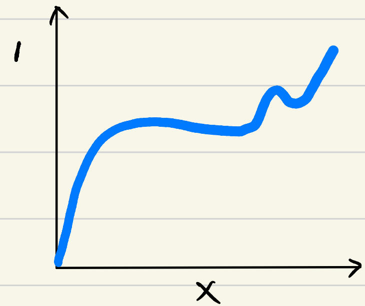
$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$



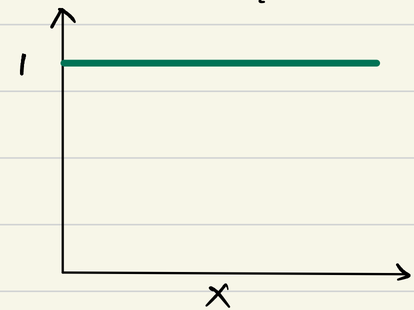
Which of the following are valid cdf?

a.

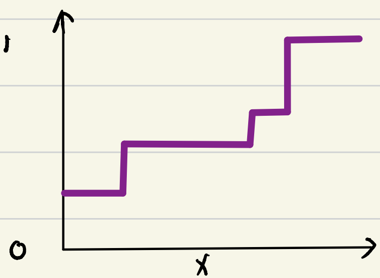
b.



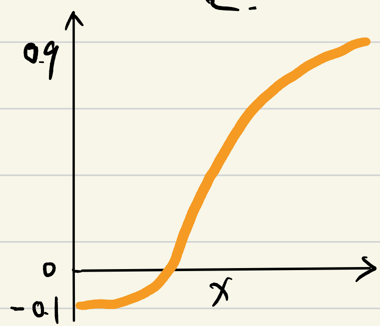
c.



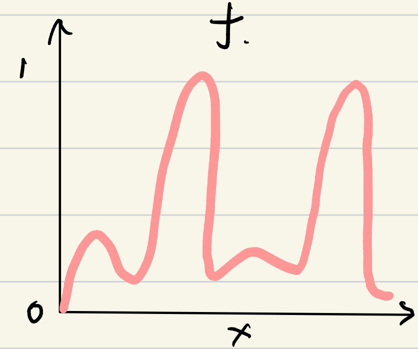
d.



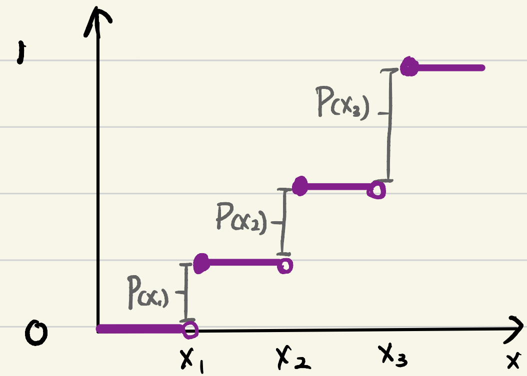
e.



f.



Is it a cdf for a discrete random variable?



Terminology

X	a random variable
$x \sim X$	a sample value of a RV X
$Pr(x) = Pr(X = x)$	probability of event X has value x
$Pr(\bar{x}) = 1 - Pr(x)$	probability of x not happening
$P(x)$	cumulative distribution function (cdf)
$p(x)$	probability mass function (pmf)
	probability density function (pdf)

Related Random Variables

joint distribution:

$$p(x,y) = p(X=x, Y=y)$$

marginal distribution:

$$p(X=x) = \sum_y p(X=x, Y=y)$$ → sum rule

conditional distribution:

$$p(Y=y \mid X=x) = \frac{p(X=x, Y=y)}{p(X=x)}$$

$$\Updownarrow$$

$$p(x,y) = p(x)\,p(y \mid x)$$ → product rule

chain rule of probability.

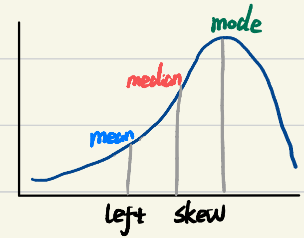
$$p(X_{1:D}) = p(x_1)\,p(x_2 \mid x_1)\,p(x_3 \mid x_1, x_2) \cdots p(x_D \mid x_{1:D-1})$$

Moments of a distribution

- mean (expected value):

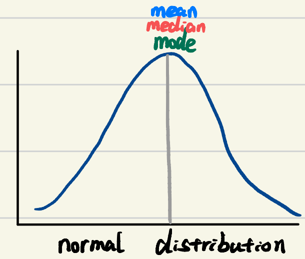
$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x\,p(x) \qquad \mathbb{E}[X] \triangleq \int_{\mathcal{X}} x\,p(x)\,dx$$

(discrete RV) (continuous RV)



linearity of expectation:

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$



- Variance

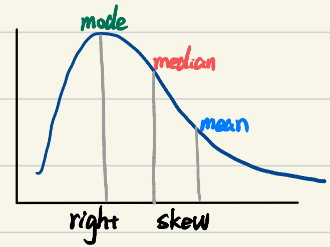
$$\mathbb{V}[X] \triangleq \mathbb{E}[(X-\mu)^2] = \int (x-\mu)^2 p(x)\,dx$$

$$= \mathbb{E}[X^2] - \mu^2$$

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2$$

$$\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i]$$

$$\mathbb{V}[aX+b] = a^2\,\mathbb{V}[X]$$



- mode

$$x^* = \underset{x}{\operatorname{argmax}}\; p(x)$$

Bayes' rule

H : unknown (or hidden) quantity

$Y=y$: observed data

prior : what we know about H before seeing data

likelihood : distribution of data we expect to see if $H=h$

$$\boxed{p(H=h \mid Y=y)} = \frac{\boxed{p(H=h)}\,\boxed{p(Y=y \mid H=h)}}{\boxed{p(Y=y)}}$$

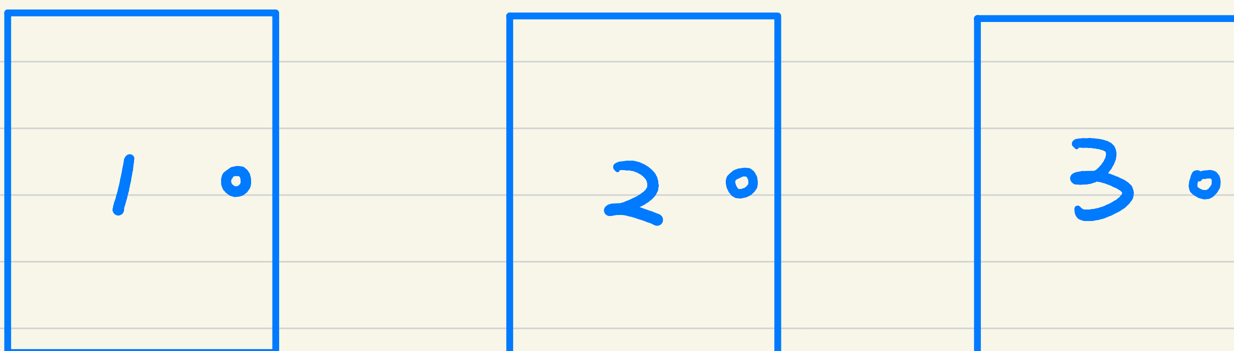
marginal likelihood

posterior : new belief state about H

$$\boxed{\text{posterior} \propto \text{prior} \times \text{likelihood}}$$

Example: The Monty Hall Problem

$1,000,000

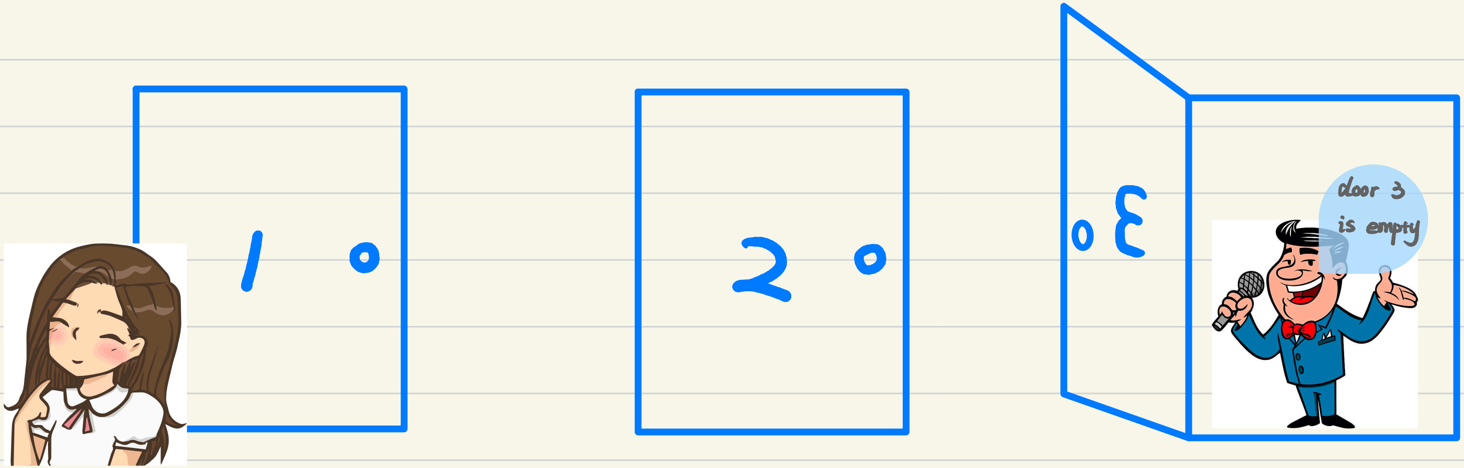


Example: The Monty Hall Problem

What's your choice?

A. stick with door 1

B. switch to door 2



Example: The Monty Hall Problem

H: (hidden quantity) the prize is behind a door $\{1, 2, 3\}$

Y: (observation) a door is opened $\{1, 2, 3\}$

We want to compare $P(H=1 \mid Y=3)$ vs. $P(H=2 \mid Y=3)$

Example: The Monty Hall Problem

H: (hidden quantity) the prize is behind a door $\{1, 2, 3\}$

Y: (observation) a door is opened $\{1, 2, 3\}$

We want to compare $P(H=1 \mid Y=3)$ vs. $P(H=2 \mid Y=3)$

$$P(H=1 \mid Y=3) = \frac{\boxed{P(Y=3 \mid H=1)}^{\frac{1}{2}\text{ why?}}\ \boxed{P(H=1)}^{\frac{1}{3}}}{\underbrace{\boxed{P(Y=3)}}_{\frac{1}{2}}} = \frac{1}{3}$$

$$P(Y=3) = P(Y=3 \mid H=1)P(H=1) + P(Y=3 \mid H=2)P(H=2) + P(Y=3 \mid H=3)\,P(H=3)$$

$$= \frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3} = \frac{1}{2}$$

$$P(H=2 \mid Y=3) = 1 - P(H=1 \mid Y=3) = \frac{2}{3}$$

choose door 2 !

Bernoulli and Binomial Distribution

Bernoulli distribution

toss a coin

θ : probability that it lands head

$p(Y = \underset{\downarrow \text{head}}{1}) = \theta$ $\qquad$ $p(Y = \underset{\downarrow \text{tail}}{0}) = 1 - \theta$

$$Y \sim \mathrm{Ber}(\theta)$$

pmf :
$$\mathrm{Ber}(y|\theta) = \begin{cases} 1-\theta & \text{if } y=0 \\ \theta & \text{if } y=1 \end{cases}$$

$$\Updownarrow$$

$$\boxed{\mathrm{Ber}(y|\theta) \triangleq \theta^{y}(1-\theta)^{1-y}}$$

Binomial distribution

toss a coin N times

number of heads : $S \triangleq \sum_{n=1}^{N} \mathbb{I}(y_n = 1)$

$$\boxed{\mathrm{Bin}(s \mid N, \theta) \triangleq \binom{N}{s} \theta^{s}(1-\theta)^{N-s}}$$

Bernoulli Distribution and Sigmoid function

Bernoulli distribution

$$\boxed{Ber(y|\theta) \triangleq \theta^y (1-\theta)^{1-y}}$$

Sigmoid (logistic) function

We want to predict a binary variable $y \in \{0,1\}$ given inputs $x \in \mathcal{X}$

$p(y|x,\theta) = Ber(y|f(x;\theta))$ $\qquad$ $\underline{0 \le f(x;\theta) \le 1}$

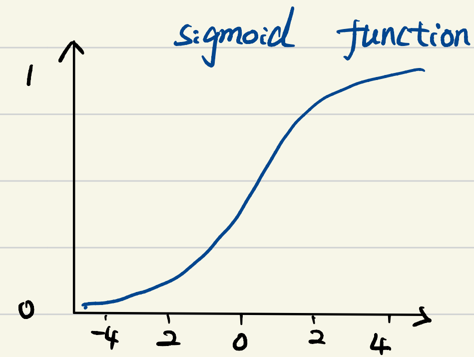
can be relaxed

$p(y|x,\theta) = Ber(y|\sigma(f(x;\theta)))$ $\qquad$ $\sigma(\,)$: sigmoid function $\qquad$ $0 \le \sigma(a) \le 1$

$$\boxed{\sigma(a) \triangleq \frac{1}{1+e^{-a}}, \text{ where } a = f(x;\theta)}$$

Sigmoid function

$$\boxed{\sigma(a) \triangleq \frac{1}{1+e^{-a}}, \text{ where } a = f(x;\theta)}$$



$$p(y=1 \mid x,\theta) = \frac{1}{1+e^{-a}} = \sigma(a)$$

$$p(y=0 \mid x,\theta) = 1 - \frac{1}{1+e^{-a}} = \sigma(-a)$$

log odds: $\boxed{\log\left(\frac{p}{1-p}\right)} = \log\left(\frac{e^a}{1+e^a}\,\frac{1+e^a}{1}\right) = \log(e^a) = \boxed{a}$, where $p = p(y=1 \mid x,\theta)$

logit

(Spoiler) Binary logistic regression

$$p(y \mid x;\theta) = \mathrm{Ber}(y \mid \sigma(w^T x + b)) \Leftrightarrow p(y=1 \mid x;\theta) = \sigma(w^T x + b) = \frac{1}{1+e^{-(w^T x + b)}}$$

decision boundary: $x^* \rightarrow p(y=1 \mid x = x^*, \theta) = 0.5$

Categorical and multinomial distributions

Categorical Distribution

roll a C-sided dice

$C > 2$

$$\boxed{\mathrm{Cat}(y \mid \theta) \triangleq \prod_{c=1}^{C} \theta_c^{\mathbb{I}(y=c)}}$$

$$p(y=c \mid \theta) = \theta_c$$

Multinomial Distribution

roll a C-sided dice N times

$$\boxed{\begin{aligned} &\mathrm{Mu}(s \mid N, \theta) \triangleq \binom{N}{s_1 \cdots s_C} \prod_{c=1}^{C} \theta_c^{s_c}, \\ &\text{where } s_c \triangleq \sum_{n=1}^{N} \mathbb{I}(y_n = c) \end{aligned}}$$

$\binom{N}{s_1 \cdots s_C}$: multinomial coefficient

number of ways to divide a set of size $N = \sum_{c=1}^{C} s_c$ into subsets with sizes s_1 to s_C

Softmax function

In conditional case:

$$p(y|x;\theta) = \text{Cat}(y|f(x;\theta)) \qquad \underline{0 \le f(x;\theta) \le 1}$$

can be relaxed

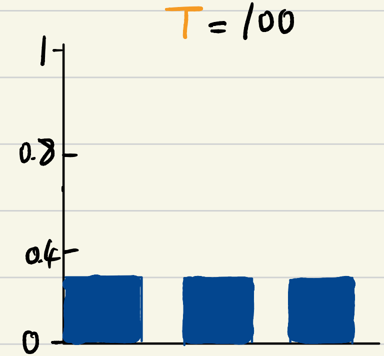
$$p(y|x;\theta) = \text{Cat}(y|S(a)_c) \qquad S(\,): \text{ softmax function} \qquad 0 \le S(a)_c \le 1, \quad \sum_{c=1}^{C} S(a)_c = 1$$

$$\boxed{S(a) \triangleq \left[\frac{e^{a_1}}{\sum_{c'=1}^{C} e^{a_{c'}}}, \dots, \frac{e^{a_C}}{\sum_{c'=1}^{C} e^{a_{c'}}} \right]}$$

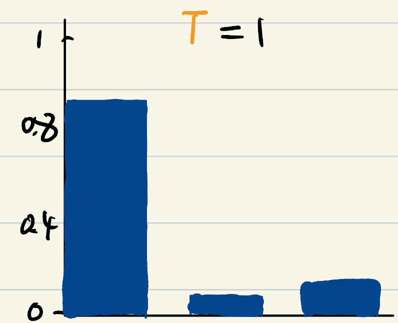
$a = f(x;\theta)$ logits

Softmax function

$$\boxed{S(a) \triangleq \left[\frac{e^{a_1}}{\sum_{c'=1}^{C} e^{a_{c'}}}, \ldots, \frac{e^{a_C}}{\sum_{c'=1}^{C} e^{a_{c'}}} \right]}$$



temperature T : change the output distribution



as $T \to 0$, $S(a/T)_c = \begin{cases} 1.0 & \text{if } c = \operatorname{argmax}_{c'} a_{c'} \\ 0.0 & \text{otherwise} \end{cases}$

Gaussian (normal) distribution

$Y \sim \mathcal{N}(\mu, \sigma^2)$

pdf:

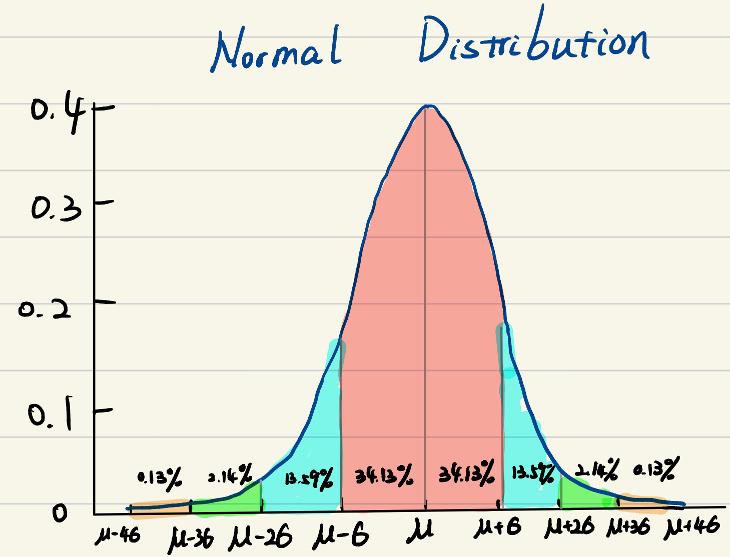
$$\mathcal{N}(y \mid \mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

mean:

$$\mu = \mathbb{E}[\mathcal{N}(\cdot \mid \mu, \sigma^2)]$$

standard deviation:

$$\sigma = \mathrm{std}[\mathcal{N}(\cdot \mid \mu, \sigma^2)]$$



II. Bayes Optimal Rule

Bayes Optimal Rule

we are searching for a "f"
that minimizes the expected loss.

measures distance
between true label and prediction.

$$\boxed{f^*} = \boxed{\underset{f}{\operatorname{argmin}}}\ \boxed{\boxed{\mathbb{E}_{XY}}\left[\boxed{\text{loss}}(\boxed{Y}, \boxed{f(X)})\right]}$$

$f^*: \mathcal{X} \to \mathcal{Y}$

the function that achieves
the minimal possible population risk

The expectation on P_{XY},
i.e. the data distribution.

prediction

true label

population
risk

Can't compute because P_{XY} is unknown.

III. Linear Algebra

Notation

vector $x \in \mathbb{R}^n$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

one-hot vector (unit vector):

$$e_i = (0, \cdots, 0, 1, 0, \cdots, 0)$$

matrix $A \in \mathbb{R}^{m \times n}$

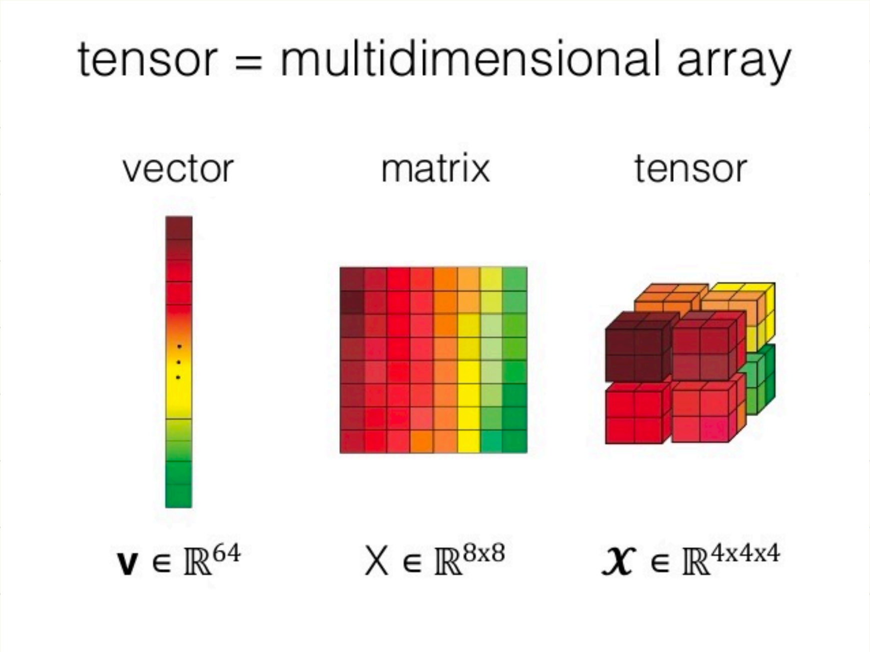
$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

A_{ij} : the entry of A in the i-th row and j-th column

$A_{i,:}$: the i'th row

$A_{:,j}$: the j'th column

Tensor



tensor:

a generalization of a 2d array to more than 2 dimensions.

order or rank of the tensor:

the number of dimensions

Vector Norms

norm of a vector $\|x\|$:

a measure of the "length" of the vector

properties :

1. non-negativity $\|x\| \geqslant 0$
2. definiteness $\|x\| = 0$ if $x = 0$
3. absolute value homogeneity $\|tx\| = |t| \, \|x\|$ for all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$
4. triangle inequality $\|x+y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$

Vector Norms

p-norm

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \text{ for } p \geq 1$$

2-norm (Euclidean norm)

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$ note that $\|x\|_2^2 = x^T x$

1-norm

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

max-norm

$$\|x\|_\infty = \max_i |x_i|$$

Matrix Multiplication

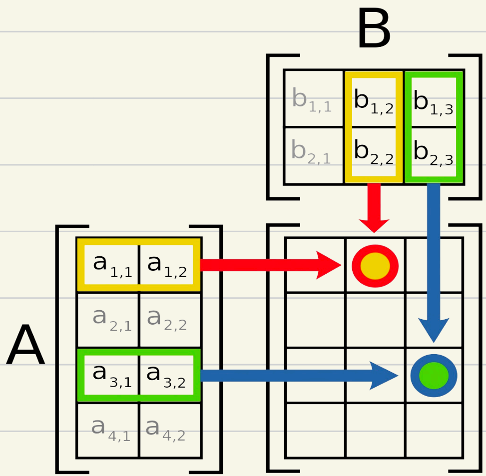
can only occur when $A \in \mathbb{R}^{N \times M}$ $B \in \mathbb{R}^{M \times D}$

$$A \times B = C \in \mathbb{R}^{N \times D}$$

$$C_{i,j} = \sum_{m=1}^{M} a_{i,m} \times b_{m,j}$$

$$\Updownarrow$$

$$C_{i,j} = A[i,:] \cdot B[:,j]$$



Matrix Derivatives and Gradients

consider a function $f: \mathbb{R}^n \to \mathbb{R}$

gradient of f: $\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$

consider a function $f: \mathbb{R}^{m\times n} \to \mathbb{R}$

$$\frac{\partial f}{\partial X} = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial _{1m}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{pmatrix}$$

consider a function $f: \mathbb{R}^n \to \mathbb{R}^m$

Jacobbian $m \times n$

$$J_f(x) = \frac{\partial f}{\partial x^T} \triangleq \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \nabla f_1(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{pmatrix}$$

Consider a function $f: \mathbb{R}^n \to \mathbb{R}$ that is twice differentiable

Hessian $n \times n$

$$H_f = \frac{\partial^2 f}{\partial x^2} = \nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ & \vdots & \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Matrix Derivatives Idioms

$$\frac{\partial (a^T x)}{\partial x} = a \qquad \frac{\partial}{\partial X}(a^T X b) = ab^T$$

$$\frac{\partial (b^T A x)}{\partial x} = A^T b \qquad \frac{\partial}{\partial X}(a^T X^T b) = ba^T$$

$$\frac{\partial (x^T A x)}{\partial x} = (A + A^T) x$$

There are more !

Refer to « The Matrix Cookbook », Petersen et al., Chapter 2.

Questions?