中國石油大學（华东）
CHINA UNIVERSITY OF PETROLEUM

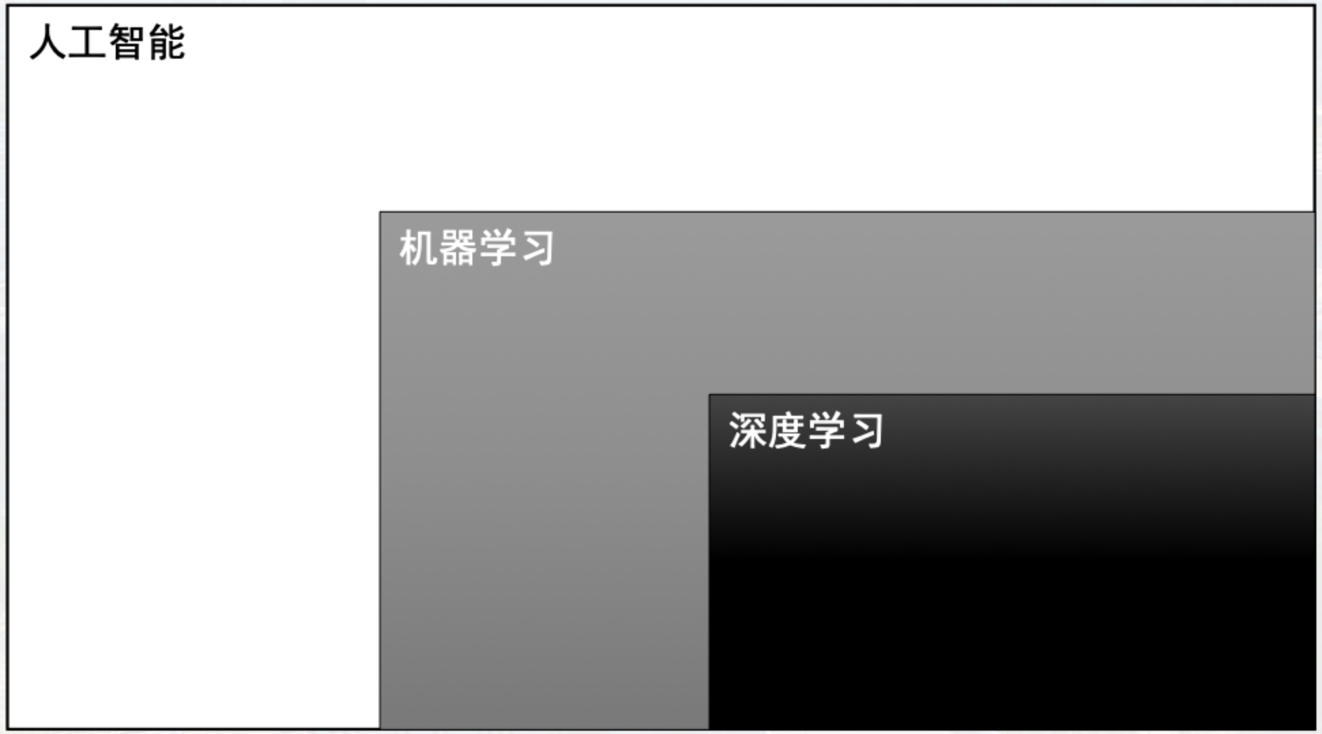
深度学习
Deep Learning

张 琛

2024/2/26

本课程

- 人工智能的一个子领域
  - 深度学习：一类机器学习问题，主要解决贡献度分配问题。

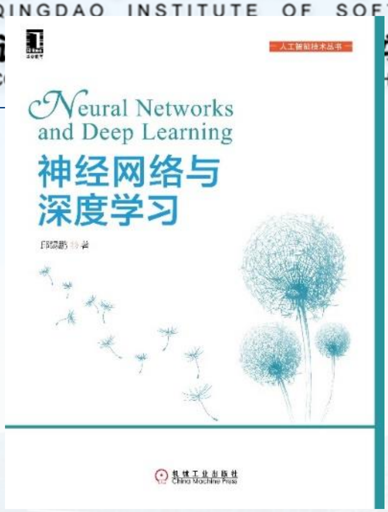


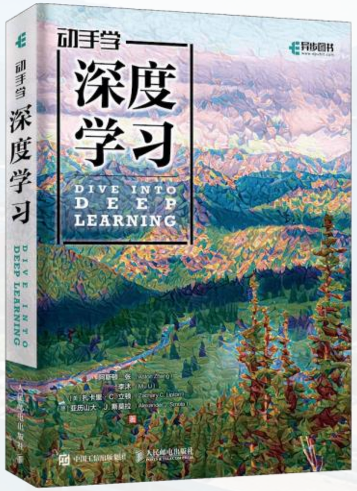
课程大纲

- 概述
  - 机器学习概述
- 基础网络模型
  - 前馈神经网络
  - 卷积神经网络
  - 循环神经网络
  - 网络优化与正则化
  - 记忆与注意力机制
  - 无监督学习
- 进阶模型
  - 深度生成模型
  - 序列生成模型

推荐教材

- 邱锡鹏,神经网络与深度学习,机械工业出版社, 2020, ISBN 9787111649687
  - https://nndl.github.io/
  - 提供配套练习





- 阿斯顿·张等,动手学深度学习, ISBN: 9787115505835
  - https://d2l.ai/
  - 有PyTorch版

- 李航, 机器学习方法. ISBN 9787302597308.
  - 统计学习方法新版升级

推荐网络课程

- 复旦大学 邱锡鹏 《神经网络与深度学习》
  https://nndl.github.io/
- 台湾大学-李宏毅：
  http://speech.ee.ntu.edu.tw/~tlkagk/courses.html
- 李沐老师（跟李沐学AI）《动手学深度学习》 pytorch
- https://space.bilibili.com/1567748478/channel/series

推荐课程

- 斯坦福大学CS224n: Deep Learning for Natural Language Processing
  - https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/
  - Chris Manning 主要讲解自然语言处理领域的各种深度学习模型
- 斯坦福大学CS231n: Convolutional Neural Networks for Visual Recognition
  - http://cs231n.stanford.edu/
  - Fei-Fei Li Andrej Karpathy 主要讲解CNN、RNN在图像领域的应用
- 加州大学伯克利分校 CS 294: Deep Reinforcement Learning
  - http://rail.eecs.berkeley.edu/deeprlcourse/

顶会论文

- NeurIPS、ICLR、ICML、AAAI、IJCAI
- ACL、EMNLP
- CVPR、ICCV
- …

预备知识

- 线性代数
- 微积分
- 数学优化
- 概率论
- 信息论

标量(Scalar)

- 实数，只有大小，没有方向。a, b, c
- 气温，考试成绩

向量(Vector)

- 一组实数组成的有序数组，同时具有大小和方向。

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [x_1, \quad x_2, \quad \cdots, \quad x_n]^T$$

标量

- **简易运算**

$$c = a + b$$

$$c = a \cdot b$$

$$c = \sin a$$

- **长度**

$$|a| = \begin{cases} -a, & a < 0 \\ a, & a \geq 0 \end{cases}$$

$$|a + b| \leq |a| + |b|$$

$$|a \cdot b| = |a| \cdot |b|$$

向量

- 简易运算

$$\boldsymbol{c} = \boldsymbol{a} + \boldsymbol{b} \ \ \textbf{where } c_i = a_i + b_i$$
$$\boldsymbol{c} = \alpha \cdot \boldsymbol{b} \ \ \textbf{where } c_i = \alpha b_i$$
$$\boldsymbol{c} = \sin \boldsymbol{a} \ \ \textbf{where } c_i = \sin a_i$$

- 长度（二范数）

$$\|\boldsymbol{a}\|_2 = \left[\sum_{i=1}^{m} \boldsymbol{a}_i^2\right]^{\frac{1}{2}}$$
$$\|\boldsymbol{a}\| \geq 0 \textbf{ for all } \boldsymbol{a}$$
$$\|\boldsymbol{a} + \boldsymbol{b}\| \leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\|$$
$$\|\boldsymbol{a} \cdot \boldsymbol{b}\| = \|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|$$

向量

- **范数**

■ 满足以下条件的函数$f: \boldsymbol{R}^n \to \boldsymbol{R}, \mathrm{dom} f = \boldsymbol{R}^n$称为范数:

- f是非负的：对所有的$x \in \boldsymbol{R}^n$成立$f(x) \geq 0$,
- f是正定的：仅对$x = 0$成立$f(x) = 0$,
- f是齐次的：对所有的$x \in \boldsymbol{R}^n$和$t \in \boldsymbol{R}$成立$f(tx) = |t| f(x)$,
- f满足三角不等式：对所有的$x, y \in \boldsymbol{R}^n$成立 $f(x+y) \leq f(x) + f(y)$。

向量

- **范数**

  - 向量$x \in \boldsymbol{R}^n$，则$\boldsymbol{R}^n$上的ℓ_1 −范数

$$\|x\|_1 = |x_1| + \cdots + |x_n|$$

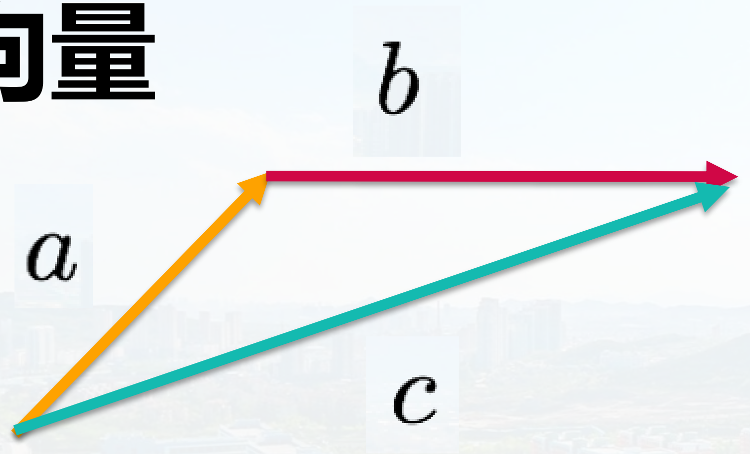
  - ℓ_∞ −范数

$$\|x\|_\infty = \max\{|x_1|, \dots, |x_n|\}$$

  - 更一般地：

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$$

向量



$$\boldsymbol{c} = \boldsymbol{a} + \boldsymbol{b}$$

$$\boldsymbol{c} = \alpha \cdot \boldsymbol{b}$$

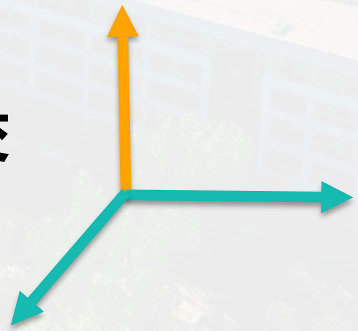
向量

- **点积**

$$a^T b = \sum_i a_i b_i$$

- **正交性**

$$a^T b = \sum_i a_i b_i = \mathbf{0}$$



如果我们有两个向量与第三个正交，它们的线性组合向量也正交

- **矩阵**(Matrix)

- $\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{bmatrix}$

$\boldsymbol{A}$是一个由M行N列个元素排列成的矩形阵列，称为$M \times N$的矩阵

- 矩阵$\boldsymbol{A}$定义了一个从空间R^N到空间R^M的**线性映射**（线性变换）

- **矩阵**(Matrix)
- **线性变换**：线性空间$\mathcal{X}$到线性空间$\mathcal{Y}$的一个映射函数

$f: \mathcal{X} \to \mathcal{Y}$，并满足：对于$\mathcal{X}$中的任何两个向量$\boldsymbol{u}$和$\boldsymbol{v}$以及标量$c$:

$$f(\boldsymbol{u}+\boldsymbol{v}) = f(\boldsymbol{u}) + f(\boldsymbol{v}) \qquad f(c\boldsymbol{v}) = cf(\boldsymbol{v})$$

- 两个有限维欧氏空间的映射函数$f: R^N \to R^M$可以表示为：

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \triangleq \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1N}x_N \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2N}x_N \\ \vdots \\ a_{M1}x_1 + a_{M2}x_2 + \cdots + a_{MN}x_N \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix}$$

矩阵

- 简易运算

$$\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B} \quad \textbf{where } \boldsymbol{C}_{ij} = \boldsymbol{A}_{ij} + \boldsymbol{B}_{ij}$$

$$\boldsymbol{C} = \alpha \cdot \boldsymbol{B} \quad \textbf{where } \boldsymbol{C}_{ij} = \alpha \cdot \boldsymbol{B}_{ij}$$

$$\boldsymbol{C} = \sin \boldsymbol{A} \quad \textbf{where } \boldsymbol{C}_{ij} = \sin \boldsymbol{A}_{ij}$$

矩阵

- Hadamard积　矩阵A和矩阵B的Hadamard积也称为逐点乘积，为A和B中对应的元素相乘。

$$[\boldsymbol{A} \odot \boldsymbol{B}]_{mn} = a_{mn} b_{mn}$$

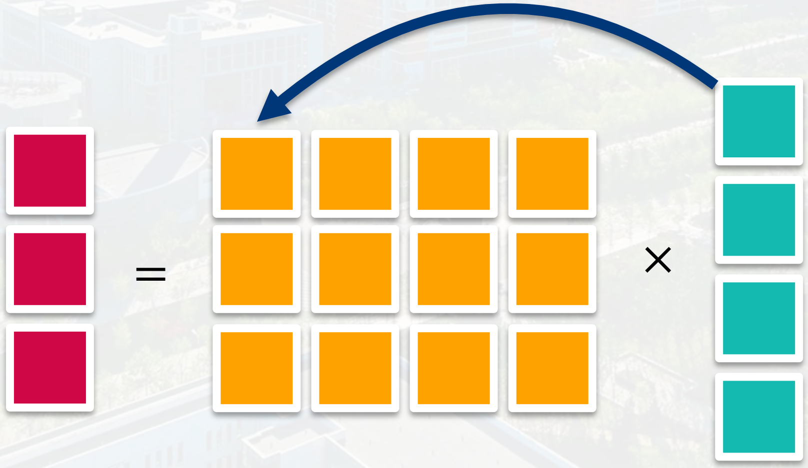
- 一个标量c与矩阵A乘积为A的每个元素是A的相应元素与c的乘积

$$[c\boldsymbol{A}]_{mn} = c a_{mn}$$

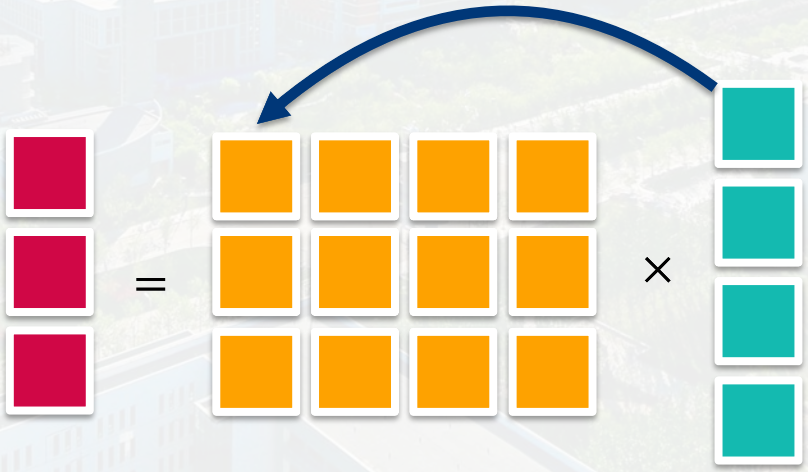
矩阵

- 矩阵相乘 (矩阵 × 向量)

$$\boldsymbol{c} = \boldsymbol{A}\boldsymbol{b} \quad \textbf{where } \boldsymbol{c}_i = \sum_j \boldsymbol{A}_{ij}\, \boldsymbol{b}_j$$



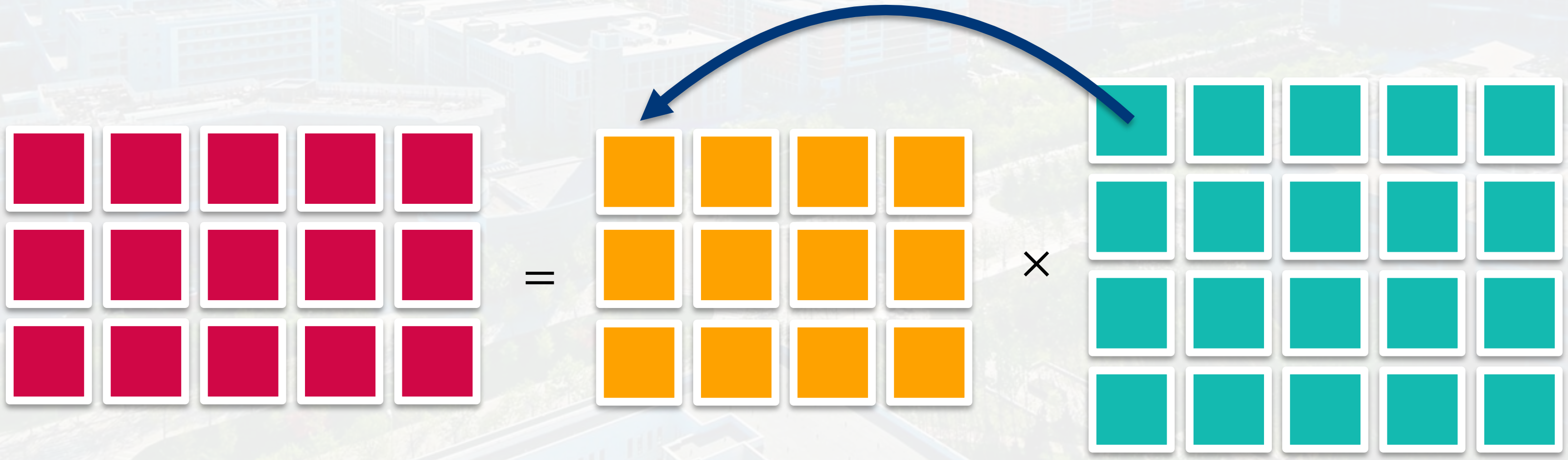
矩阵



矩阵

- 矩阵相乘 (矩阵 × 矩阵)

$$C = AB \quad \text{where } C_{ik} = \sum_j A_{ij}\, b_{jk}$$



矩阵

- **算子（诱导）范数**

$$\boldsymbol{c} = \boldsymbol{A} \cdot \boldsymbol{b} \text{ hence } \|\boldsymbol{c}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{b}\|$$

$$\|\boldsymbol{A}\| = \max\left\{\frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} : \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \boldsymbol{0}\right\}$$

常见算子范数

$$\|\boldsymbol{A}\|_p = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p}$$

矩阵

- 范数

$$\|\boldsymbol{A}\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|$$

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{max}(\boldsymbol{A}^T\boldsymbol{A})}$$

$$\|\boldsymbol{A}\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|$$

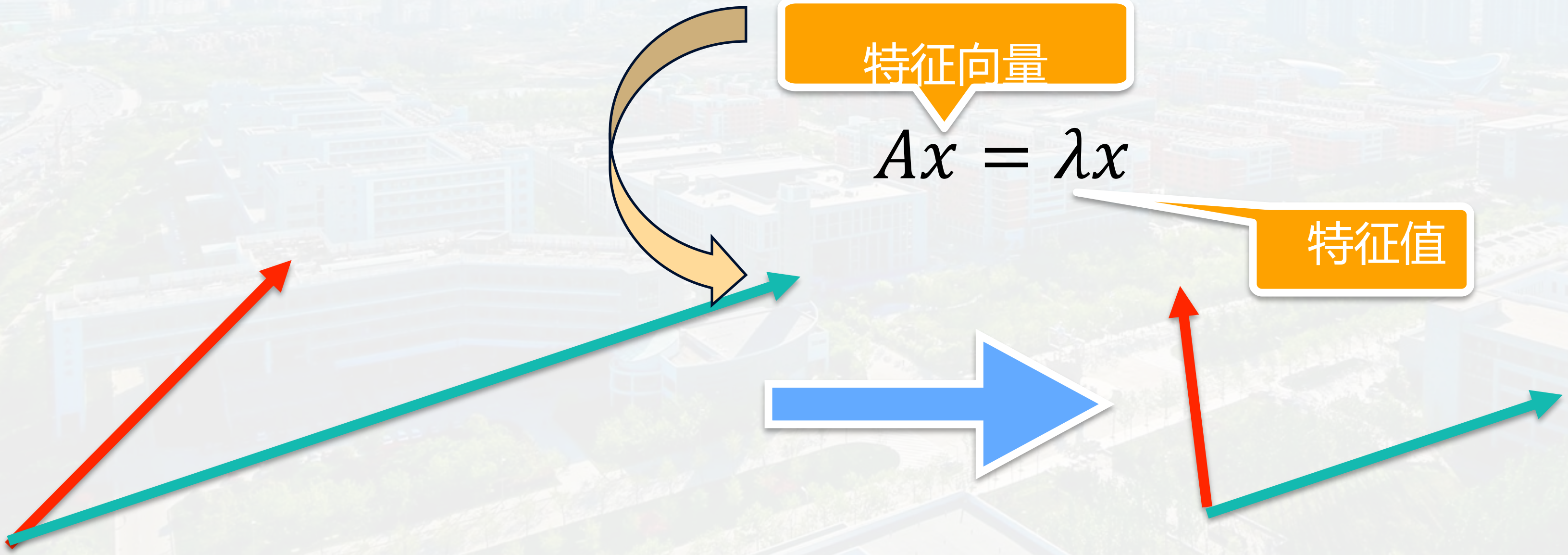
- Frobenius 范数

$$\|A\|_{\text{Frob}} = \left[\sum_{ij} A_{ij}^2\right]^{\frac{1}{2}}$$

矩阵

- **特征值和特征向量**
  - 对于一个给定的线性变换A，它的特征向量x，经过这个线性变换之后，得到的新向量仍然与原来的x保持在同一条直线上，但其长度或方向也许会改变。

$$Ax = \lambda x$$



  - 对称矩阵总会有相应的特征向量和特征值

特殊矩阵

- 对称性 & 反对称性

$$A_{ij} = A_{ji} \qquad\qquad A_{ij} = -A_{ji}$$

- 正定性

$\|x\|^2 = x^T x \geq 0$ **一般化：** $\forall x \neq 0, x^T A x > 0$ **称A为正定矩阵**

- **设A是$n \times n$对称矩阵，当且仅当A的特征值均为非负数，称A为半正定矩阵**

特殊矩阵

- **正交矩阵**
  - ❑所有的列向量都是单位正交向量
  - ❑所有的行向量都是单位正交向量
  - ❑可以写为：

$$UU^T = \boldsymbol{I}$$

- **置换矩阵**
  - ❑ **矩阵的每一行和每一列的元素中只有一个1,其余元素都为0**

$$P \text{ where } P_{ij} = 1 \text{ if and only if } j = \pi(i)$$

  - ❑ **置换矩阵是正交矩阵。**

张量（Tensor）

- 一个数组中的元素分布在若干维坐标的规则网格中。

3-d

[[[1, 2, 3]
[5, 2, 6]
[4, 8, 2]]
[[2, 7, 4]
[4, 2, 3]
[1, 3, 6]]]

一张RGB图像
(长×宽×高)

4-d

[[[[…
…
…]]]]

一组RGB图像
(批量大小×长×宽×高)

5-d

[[[[[…
…
…]]]]]

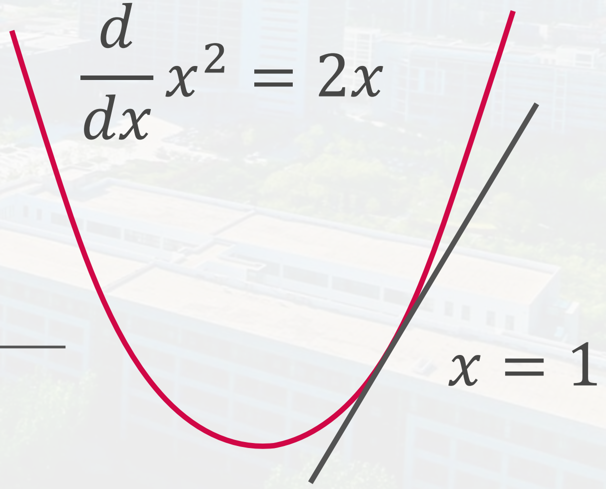
一组RGB视频
(批量大小×时间×长×宽×高)

标量求导回顾

y	a	x^n	$\exp(x)$	$\log(x)$	$\sin(x)$
$\frac{dy}{dx}$	0	nx^{n-1}	$\exp(x)$	$\frac{1}{x}$	$\cos(x)$

y	$u+v$	uv	$y=f(u), u=g(x)$
$\frac{dy}{dx}$	$\frac{du}{dx}+\frac{dv}{dx}$	$\frac{du}{dx}v+\frac{dv}{dx}u$	$\frac{dy}{du}\frac{du}{dx}$

导数是切线的斜率

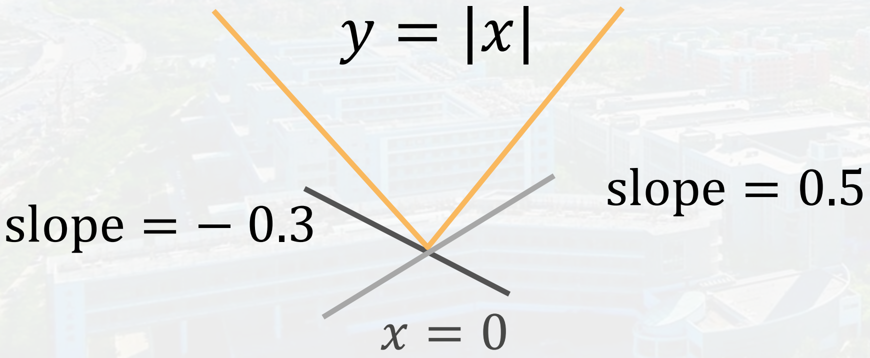


切线的斜率为2

次导数

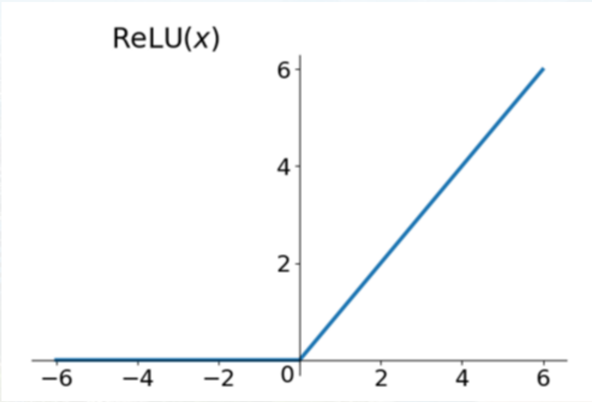
- 不可求导情况下的导数

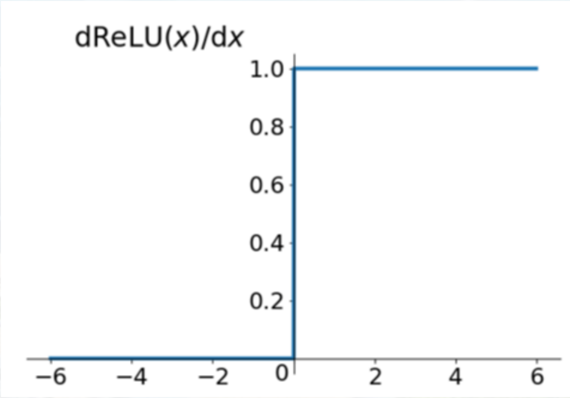
例1:



$$\frac{\partial |x|}{\partial x} = \begin{cases} 1 & \text{if} \quad x > 0 \\ -1 & \text{if} \quad x < 0 \\ a & \text{if} \quad x = 0, a \in [-1,1] \end{cases}$$

例2:





$$\frac{\partial}{\partial x} \max(x, 0) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x < 0 \\ a & \text{if} \quad x = 0, a \in [0,1] \end{cases}$$

梯度

- 矢量求导推广

		标量	矢量
		x	$\mathbf{x}$
标量	y	$\frac{\partial y}{\partial x}$	$\frac{\partial y}{\partial \mathbf{x}}$
矢量	$\mathbf{y}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$

梯度

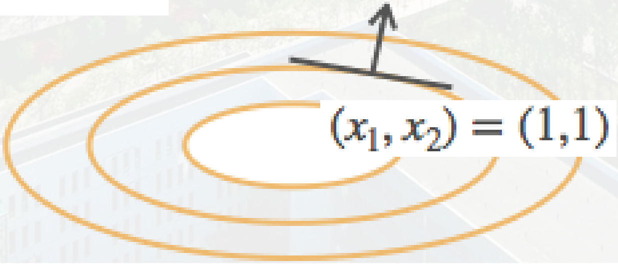
$\partial y/\partial \boldsymbol{x}$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \frac{\partial y}{\partial \boldsymbol{x}} = \left[\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \ldots, \frac{\partial y}{\partial x_n}\right]$$

	x	$\mathbf{x}$
y	$\frac{\partial y}{\partial x}$	$\frac{\partial y}{\partial \mathbf{x}}$
$\mathbf{y}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$

$$\frac{\partial(x_1^2 + 2x_2^2)}{\partial \boldsymbol{x}} = [2x_1, 4x_2]$$

Direction (2, 4), perpendicular to the contour lines



例子

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

y	a	au	$\text{sum}(\mathbf{x})$	$\Vert \mathbf{x} \Vert^2$
$\frac{\partial y}{\partial \mathbf{x}}$	$\mathbf{0}^T$	$a\frac{\partial u}{\partial \mathbf{x}}$	$\mathbf{1}^T$	$2\mathbf{x}^T$

y	$u+v$	uv	$\langle \mathbf{u}, \mathbf{v} \rangle$
$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial u}{\partial \mathbf{x}} + \frac{\partial v}{\partial \mathbf{x}}$	$\frac{\partial u}{\partial \mathbf{x}} v + \frac{\partial v}{\partial \mathbf{x}} u$	$\mathbf{u}^T \frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^T \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$

梯度

$\partial \boldsymbol{y}/\partial x$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \qquad \frac{\partial \boldsymbol{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$

	x	$\mathbf{x}$
y	$\frac{\partial y}{\partial x}$	$\frac{\partial y}{\partial \mathbf{x}}$
$\mathbf{y}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$

$\partial \boldsymbol{y} / \partial \boldsymbol{x}$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \qquad \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} \in \mathbb{R}^m, \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \in \mathbb{R}^{m \times n}$$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \boldsymbol{x}} \\ \frac{\partial y_2}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial y_m}{\partial \boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1}, \frac{\partial y_1}{\partial x_2}, \dots, \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1}, \frac{\partial y_2}{\partial x_2}, \dots, \frac{\partial y_2}{\partial x_n} \\ \vdots \\ \frac{\partial y_m}{\partial x_1}, \frac{\partial y_m}{\partial x_2}, \dots, \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

	x	$\mathbf{x}$
y	$\frac{\partial y}{\partial x}$	$\frac{\partial y}{\partial \mathbf{x}}$
$\mathbf{y}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$

例子

$\mathbf{y}$	a	$\mathbf{x}$	$\mathbf{Ax}$	$\mathbf{x}^T\mathbf{A}$
$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	$\mathbf{0}$	$\mathbf{I}$	$\mathbf{A}$	$\mathbf{A}^T$

$\mathbf{y}$	$a\mathbf{u}$	$\mathbf{Au}$	$\mathbf{u}+\mathbf{v}$
$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\mathbf{A}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}+\frac{\partial \mathbf{v}}{\partial \mathbf{x}}$

$$\mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^m, \frac{\partial \mathbf{y}}{\partial \mathbf{x}}\in\mathbb{R}^{m\times n}$$

$a, \boldsymbol{a}$和 $\boldsymbol{A}$ 不是关于$\mathbf{x}$的函数

$\mathbf{0}$ 和 $\mathbf{I}$ 为矩阵

推广到矩阵

		标量	矢量	矩阵
		x $(1,)$	$\mathbf{x}$ $(n,1)$	$\mathbf{X}$ (n,k)
标量	y $(1,)$	$\frac{\partial y}{\partial x}$ $(1,)$	$\frac{\partial y}{\partial \mathbf{x}}$ $(1,n)$	$\frac{\partial y}{\partial \mathbf{X}}$ (k,n)
矢量	$\mathbf{y}$ $(m,1)$	$\frac{\partial \mathbf{y}}{\partial x}$ $(m,1)$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ (m,n)	$\frac{\partial \mathbf{y}}{\partial \mathbf{X}}$ (m,k,n)
矩阵	$\mathbf{Y}$ (m,l)	$\frac{\partial \mathbf{Y}}{\partial x}$ (m,l)	$\frac{\partial \mathbf{Y}}{\partial \mathbf{x}}$ (m,l,n)	$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$ (m,l,k,n)

链式法则

- 链式法则 – 标量:

$$y = f(u), u = g(x) \qquad \frac{\partial y}{\partial x} = \frac{\partial y}{\partial u}\frac{\partial u}{\partial x}$$

- 链式法则 – 矢量:

$$\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial y}{\partial u}\frac{\partial u}{\partial \mathbf{x}} \qquad \frac{\partial y}{\partial \mathbf{x}} = \frac{\partial y}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}} \qquad \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$

$(1,n) \quad (1,) \ (1,n) \qquad (1,n) \quad (1,k) \ (k,n) \qquad (m,n) \quad (m,k) \ (k,n)$

例1

假设 $\mathbf{x}, \mathbf{w} \in \mathbb{R}^n, \quad y \in \mathbb{R}$ $\qquad z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$

计算 $\dfrac{\partial z}{\partial \mathbf{w}}$

例1

假设 $\mathbf{x}, \mathbf{w} \in \mathbb{R}^n, \quad y \in \mathbb{R}$ $\qquad z = \left(\langle \mathbf{x}, \mathbf{w} \rangle - y\right)^2$

计算 $\dfrac{\partial z}{\partial \mathbf{w}}$

分解

$$a = \langle \mathbf{x}, \mathbf{w} \rangle$$
$$b = a - y$$
$$z = b^2$$

$$\begin{aligned}
\frac{\partial z}{\partial \mathbf{w}} &= \frac{\partial z}{\partial b}\frac{\partial b}{\partial a}\frac{\partial a}{\partial \mathbf{w}} \\
&= \frac{\partial b^2}{\partial b}\frac{\partial a - y}{\partial a}\frac{\partial \langle \mathbf{x}, \mathbf{w} \rangle}{\partial \mathbf{w}} \\
&= 2b \cdot 1 \cdot \mathbf{x}^T \\
&= 2\left(\langle \mathbf{x}, \mathbf{w} \rangle - y\right)\mathbf{x}^T
\end{aligned}$$

$$\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial y}{\partial u}\frac{\partial u}{\partial \mathbf{x}}$$

例2

假设 $\mathbf{X} \in \mathbb{R}^{m\times n}, \quad \mathbf{w} \in \mathbb{R}^{n}, \quad \mathbf{y} \in \mathbb{R}^{m} \qquad z = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$

计算 $\dfrac{\partial z}{\partial \mathbf{w}}$

例2

假设 $\mathbf{X} \in \mathbb{R}^{m\times n}, \quad \mathbf{w} \in \mathbb{R}^{n}, \quad \mathbf{y} \in \mathbb{R}^{m} \qquad z = \|\mathbf{Xw} - \mathbf{y}\|^2$

计算 $\dfrac{\partial z}{\partial \mathbf{w}}$

分解

$$\mathbf{a} = \mathbf{Xw}$$
$$\mathbf{b} = \mathbf{a} - \mathbf{y}$$
$$z = \|\mathbf{b}\|^2$$

$$\begin{aligned}
\frac{\partial z}{\partial \mathbf{w}} &= \frac{\partial z}{\partial \mathbf{b}}\frac{\partial \mathbf{b}}{\partial \mathbf{a}}\frac{\partial \mathbf{a}}{\partial \mathbf{w}} \\
&= \frac{\partial \|\mathbf{b}\|^2}{\partial \mathbf{b}}\frac{\partial \mathbf{a}-\mathbf{y}}{\partial \mathbf{a}}\frac{\partial \mathbf{Xw}}{\partial \mathbf{w}} \\
&= 2\mathbf{b}^T \times \mathbf{I} \times \mathbf{X} \\
&= 2\left(\mathbf{Xw} - \mathbf{y}\right)^T \mathbf{X}
\end{aligned}$$

$$\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial y}{\partial \boldsymbol{u}}\frac{\partial \boldsymbol{u}}{\partial \mathbf{x}}$$

自动微分（AD）

- 自动微分（AD）将符号微分法应用于最基本的算子，然后代入数值，应用于整个函数
- 其它常见微分法
  - 符号微分法

```
In[1]:= D[4 x^3 + x^2 + 3, x]

Out[1]= 2 x + 12 x^2
```

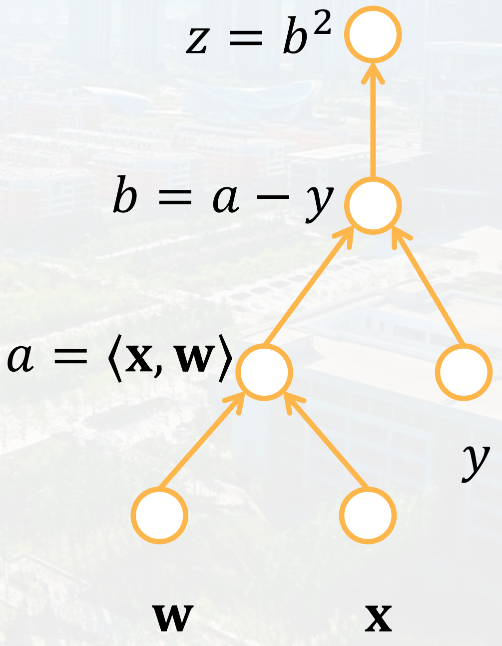
  - 数值微分法

$$\frac{\partial f(x)}{\partial x} = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

计算图

- 将代码分解成最基本的方程（操作子）
- 构造有向无环图来表示运算

假设 $z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$



两种模式

- 通过链式法则

$$\frac{\partial y}{\partial x}=\frac{\partial y}{\partial u_n}\frac{\partial u_n}{\partial u_{n-1}}\cdots\frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial x}$$

- 正向传播

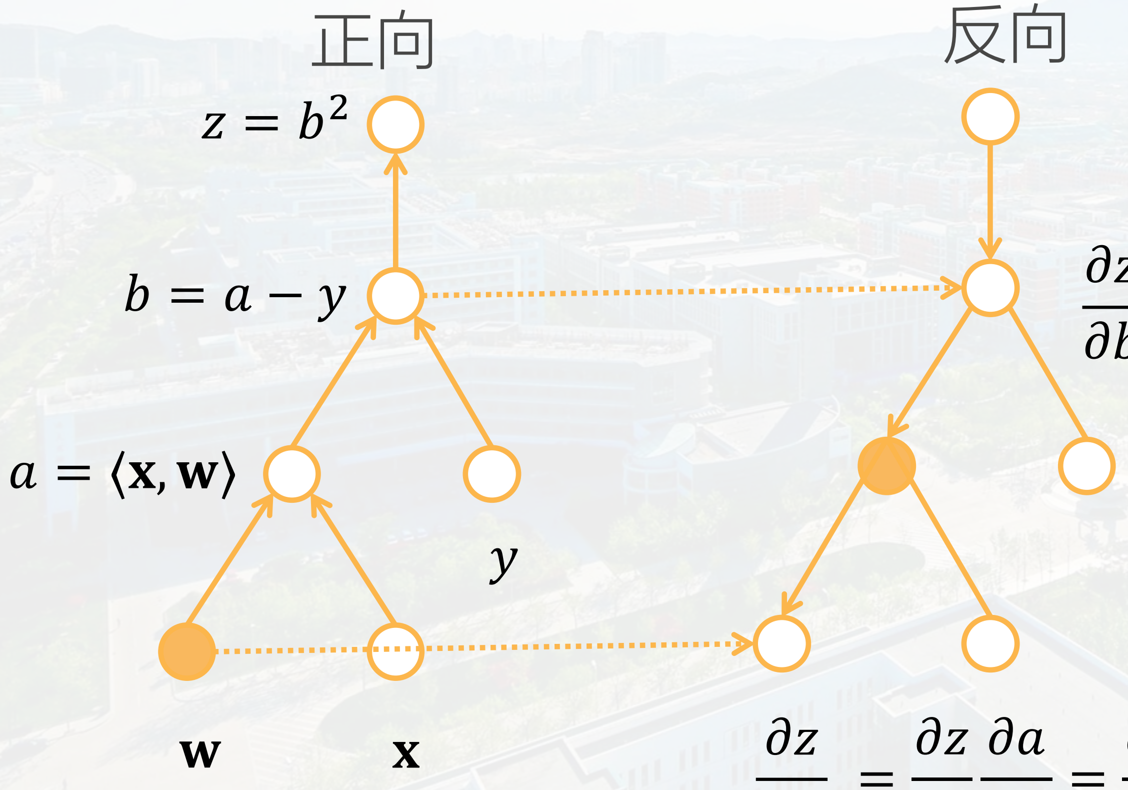
$$\frac{\partial y}{\partial x}=\frac{\partial y}{\partial u_n}\left(\frac{\partial u_n}{\partial u_{n-1}}\left(\cdots\left(\frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial x}\right)\right)\right)$$

- 反向传播

$$\frac{\partial y}{\partial x}=\left(\left(\left(\frac{\partial y}{\partial u_n}\frac{\partial u_n}{\partial u_{n-1}}\right)\cdots\right)\frac{\partial u_2}{\partial u_1}\right)\frac{\partial u_1}{\partial x}$$

反向传播

假设 $z = (\langle \mathbf{x}, \mathbf{w} \rangle - y)^2$



$$\frac{\partial z}{\partial b} = 2b$$

$$\frac{\partial z}{\partial a} = \frac{\partial z}{\partial b}\frac{\partial b}{\partial a} = \frac{\partial z}{\partial b}$$

$$\frac{\partial z}{\partial \mathbf{w}} = \frac{\partial z}{\partial a}\frac{\partial a}{\partial \mathbf{w}} = \frac{\partial z}{\partial a}\boldsymbol{x}^T$$

定义

- 目标函数 $f: \mathcal{A} \to \mathbb{R}$
- 参数 $x^* \in \mathcal{D} \subset \mathcal{A}$
- 目标: $f(x^*) \leq f(x)$ or $f(x^*) \geq f(x)$
- 约束集（可行域）： $\mathcal{D}$

无约束优化

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

目标函数 $f: \mathbb{R}^D \to \mathbb{R}$

输入变量 $\boldsymbol{x} \in \mathbb{R}^D$

可行域 $\mathcal{D} = \mathbb{R}^D$

约束优化

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\text{s.t.} \begin{cases} h_m(\boldsymbol{x}) = 0, \ \ m = 1,2,\dots,M \\ g_n(\boldsymbol{x}) \leq 0, \ \ n = 1,2,\dots,N \end{cases}$$

目标函数 $f: \mathbb{R}^D \rightarrow \mathbb{R}$

输入变量 $\boldsymbol{x} \in \mathbb{R}^D$

可行域 $\mathcal{D} = \mathrm{dom}(f) \cap_{m=1}^{M} \mathrm{dom}(h_m) \cap_{n=1}^{N} \mathrm{dom}(g_n) \subseteq \mathbb{R}^D$

梯度下降法

- 对于函数$f(\boldsymbol{x})$，如果$f(\boldsymbol{x})$在点$\boldsymbol{x}_t$附近是连续可微的，则$f(\boldsymbol{x})$下降最快的方向是$f(\boldsymbol{x})$在$\boldsymbol{x}_t$点的梯度方向的反方向。
- 泰勒一阶展开式：

$$f(\boldsymbol{x}_{t+1}) = f(\boldsymbol{x}_t + \Delta\boldsymbol{x}) \approx f(\boldsymbol{x}_t) + \Delta\boldsymbol{x}^T \nabla f(\boldsymbol{x}_t)$$

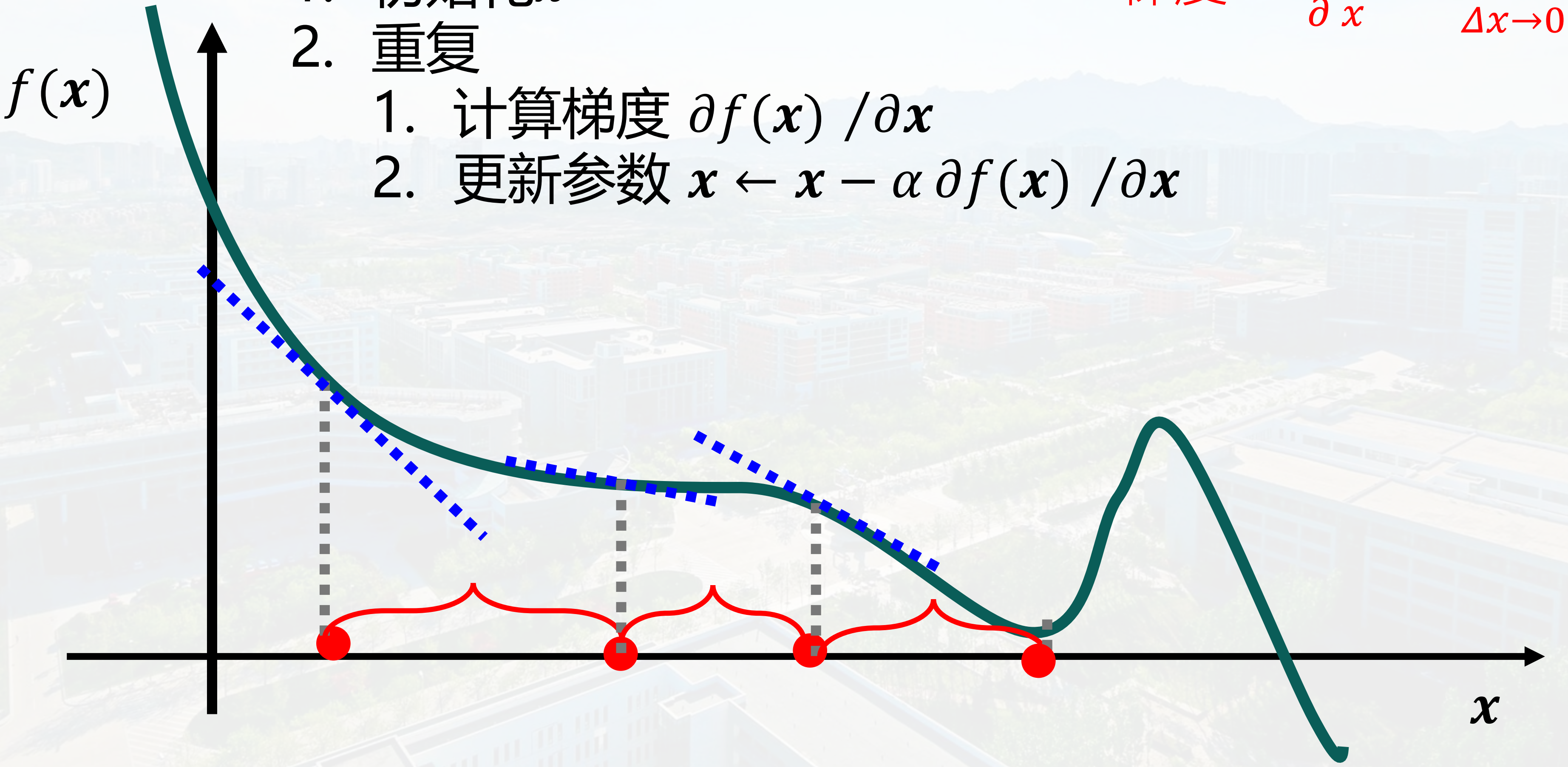
取$\Delta\boldsymbol{x} = -\alpha\nabla f(\boldsymbol{x}_t)$，$\boldsymbol{\alpha} > 0$为一个足够小的数值，则

$$f(\boldsymbol{x}_{t+1}) < f(\boldsymbol{x}_t)$$

数学优化

梯度：$\frac{\partial f(x)}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x)}{\Delta x}$

1. 初始化$\boldsymbol{x}$
2. 重复
   1. 计算梯度 $\partial f(\boldsymbol{x}) / \partial \boldsymbol{x}$
   2. 更新参数 $\boldsymbol{x} \leftarrow \boldsymbol{x} - \alpha \, \partial f(\boldsymbol{x}) / \partial \boldsymbol{x}$



等式约束优化

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$h_m(\boldsymbol{x}) = 0, \qquad m = 1,2,\dots,M$$

构造拉格朗日函数:

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda_m h_m(\boldsymbol{x})$$

令:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}} = 0, \quad \frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = 0$$

求解方程组可得到原始问题的可能解。

不等式约束优化

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\text{s.t.}\begin{cases} h_m(\boldsymbol{x}) = 0,\ m = 1,2,\dots,M \\ g_n(\boldsymbol{x}) \le 0,\ n = 1,2,\dots,N \end{cases}$$

构造拉格朗日函数：

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}) = f(\boldsymbol{x}) + \sum_{m=1}^{M} a_m h_m(\boldsymbol{x}) + \sum_{n=1}^{N} b_n g_n(\boldsymbol{x})$$

当约束条件不满足时，有 $\max_{\boldsymbol{a},\boldsymbol{b}} \mathcal{L}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}) = \infty$

当约束条件满足，且$\boldsymbol{b} \ge \boldsymbol{0}$时，$\max_{\boldsymbol{a},\boldsymbol{b}} \mathcal{L}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}) = f(\boldsymbol{x})$

因此，原问题等价于：

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{a},\boldsymbol{b}} \mathcal{L}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b})$$

$$\text{s.t.}\quad \boldsymbol{b} \ge \boldsymbol{0}$$

min-max优化问题称为**主问题**。

原问题：

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{a},\boldsymbol{b}} \mathcal{L}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b})$$

$$\text{s.t.} \quad \boldsymbol{b} \geq \boldsymbol{0}$$

定义拉格朗日对偶函数：

$$\Gamma(\boldsymbol{a},\boldsymbol{b}) = \inf_{\boldsymbol{x}\in\mathcal{D}} \mathcal{L}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b})$$

当$\boldsymbol{b} \geq \boldsymbol{0}$，对任意$\tilde{\boldsymbol{x}} \in \mathcal{D}$，有：

$$\Gamma(\boldsymbol{a},\boldsymbol{b}) = \inf_{\boldsymbol{x}\in\mathcal{D}} \mathcal{L}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{b}) \leq \mathcal{L}(\tilde{\boldsymbol{x}},\boldsymbol{a},\boldsymbol{b}) \leq f(\tilde{\boldsymbol{x}})$$

令$\boldsymbol{p}^*$为原问题的最优值，有：

$$\Gamma(\boldsymbol{a},\boldsymbol{b}) \leq \boldsymbol{p}^*$$

拉格朗日对偶函数$\Gamma(\boldsymbol{a},\boldsymbol{b})$为原问题最优值的下界。拉格朗日对偶问题：

$$\max_{\boldsymbol{a},\boldsymbol{b}} \Gamma(\boldsymbol{a},\boldsymbol{b}) \qquad \text{s.t.} \quad \boldsymbol{b} \geq \boldsymbol{0}$$

数学优化

令$\boldsymbol{d}^*$表示拉格朗日对偶问题的最优值：

- **弱对偶性：** $\boldsymbol{d}^* \le \boldsymbol{p}^*$
- **强对偶性：** $\boldsymbol{d}^* = \boldsymbol{p}^*$

强对偶性成立时，令$\boldsymbol{x}^*$和$\boldsymbol{a}^*, \boldsymbol{b}^*$分别是原问题和对偶问题的最优解，则它们满足以下条件（KKT条件）：

$$\nabla f(x^*) + \sum_{m=1}^{M} a_m^* \nabla h_m(x^*) + \sum_{n=1}^{N} b_n^* \nabla g_n(x^*) = 0$$

$$h_m(x^*) = 0, \qquad m = 1,2,\dots,M$$

$$g_n(x^*) \le 0, \qquad n = 1,2,\dots,N$$

$$b_n^* g_n(x^*) = 0, \qquad \mathrm{n} = 1,2,\dots,N$$

互补松弛条件

$$b_n^* \ge 0, \qquad n = 1,2,\dots,N$$

- 概率（Probability）
  - 一个随机事件发生的可能性大小，为0到1之间的实数。
- 随机变量（Random Variable）
  - 比如随机掷一个骰子，得到的点数就可以看成一个随机变量X，其取值为$\{1,2,3,4,5,6\}$。
- 概率分布（Probability Distribution）
  - 一个随机变量X取每种可能值的概率

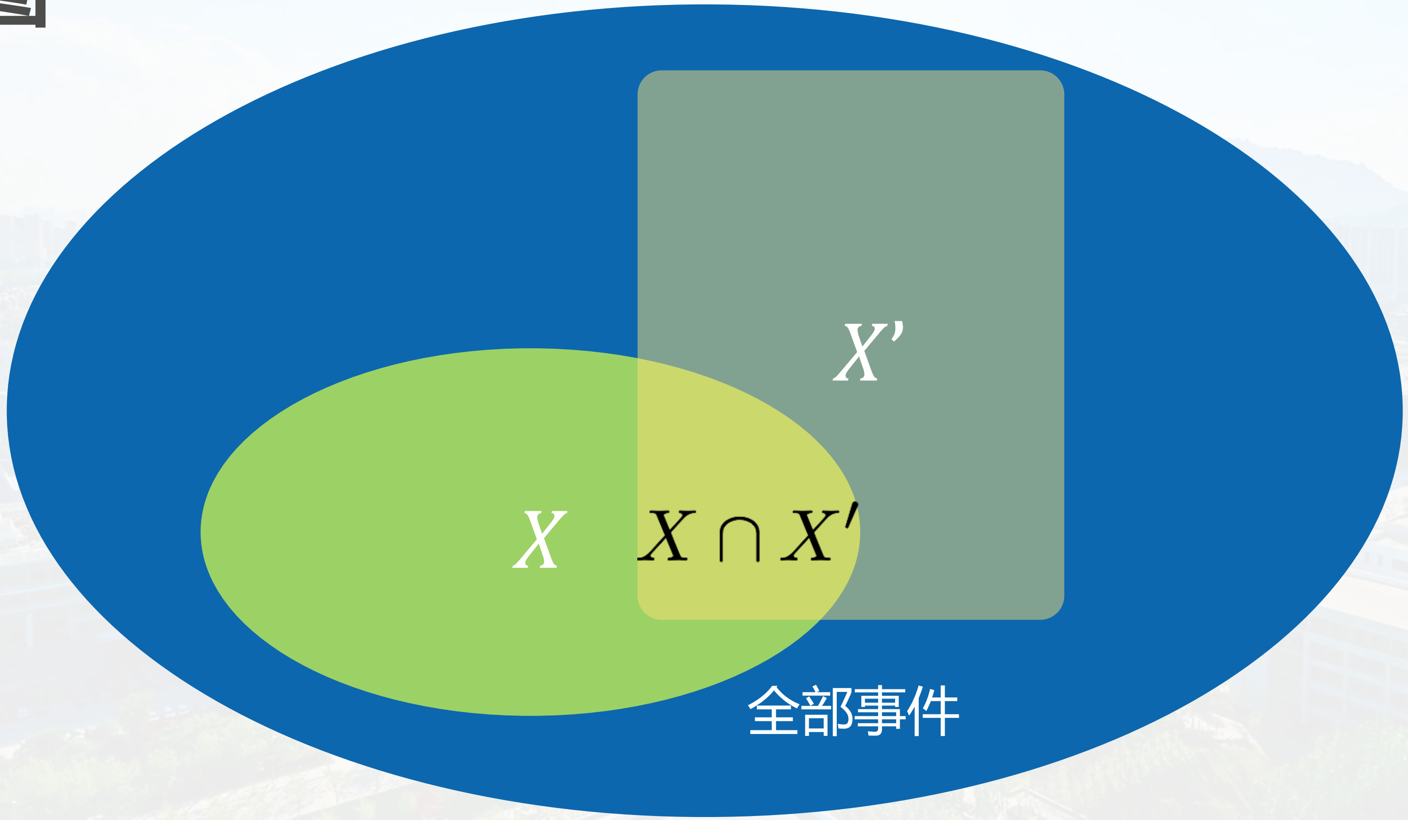
$$P(X=x_i)=p(x_i), \qquad \forall i \in \{1,2,\ldots,n\}$$

  - 并满足

$$\sum_{i=1}^{n} p(x_i)=1$$

$$p(x_i)\geq 0, \qquad \forall i \in \{1,2,\ldots,n\}$$

韦恩图



$$\Pr(X \cup X') = \Pr(X) + \Pr(X') - \Pr(X \cap X')$$

独立性与相关性

- 独立事件
  - 第一次抛骰子，和第2次抛骰子，两次抛出的点数没有影响
  - 袋中有3个球，第1次取完放回，则对第2次随机取1个概率没有影响

$$\Pr(x, y) = \Pr(x) \cdot \Pr(y)$$

- 相关事件
  - 邮件
  - 搜索
  - 新闻流
  - 即时通讯

Everywhere

$$\Pr(x, y) \neq \Pr(x) \cdot \Pr(y)$$

不确定性和条件作用

- **不确定性**
  - ❑扔硬币 （正, 反, 边）
  - ❑彩票
- **条件作用**
  - ❑ （如果信息相关,）更多信息使事情更加确定。
    - $p(y|x)$ 而不是 $p(y)$
  - ❑我们可以建立分类器，回归量等等。

贝叶斯法则

- 联合概率

$$\Pr(X, Y) = \Pr(X|Y)\Pr(Y) = \Pr(Y|X)\Pr(X)$$

- 贝叶斯法则

$$\Pr(X|Y) = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)}$$

- 假设检验
- 逆向假设

- 伯努利分布（Bernoulli Distribution）
  - 在一次试验中，事件A出现的概率为μ，不出现的概率为$1-\mu$。若用变量X表示事件A出现的次数，则X的取值为0和1，其相应的分布为

$$p(x)=\mu^{x}(1-\mu)^{(1-x)}$$

- 二项分布（Binomial Distribution）
  - 在n次伯努利分布中，若以变量X表示事件A出现的次数，则X的取值为$\{0,\dots,n\}$，其相应的分布

$$P(X=k)=\binom{n}{k}\mu^{k}(1-\mu)^{(n-k)},k=1,2,\dots,n$$

二项式系数，表示从n个元素中取出k个元素而不考虑其顺序的组合的总数。

均匀分布

- 在一个区间内恒定，在区间外为零

$$p(x) = \frac{1}{U - L} \text{ if } L \le x \le U$$

正态分布

- 概率密度函数（PDF）

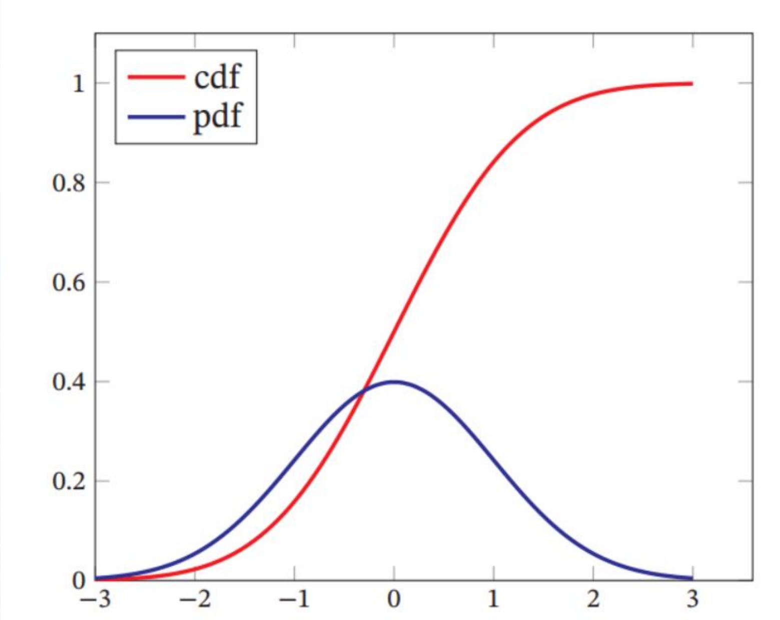
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

累积分布函数

$$\mathrm{CDF}(x) = P(X \le x)$$

连续随机变量X，累积分布函数：

$$\mathrm{CDF}(x) = \int_{-\infty}^{x} p(t)dt$$



标准正态分布的概率密度函数和累积分布函数

- 期望:
  - 离散

$$\mathbb{E}[X] = \sum_{n=1}^{N} x_n\, p(x_n)$$

  - 连续:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, p(x) \mathrm{d}x$$

- 方差:

$$\mathrm{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- 协方差:

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

- 协方差矩阵:

$$\mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])^T]$$

熵（Entropy）

- 在信息论中，熵用来衡量一个随机事件的不确定性。
  - 自信息（Self Information）$I(x) = -\log(p(x))$
  - 熵

$$
\begin{aligned}
H(X) &= \mathbb{E}_X[I(x)] \\
&= \mathbb{E}_X[-\log p(x)] \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x)
\end{aligned}
$$

  - 熵越高，则随机变量的信息越多（不确定性越大）；
  - 熵越低，则随机变量的信息越少（不确定性越小）．
  - 在对分布 $q(y)$ 的符号进行编码时，熵 $I(q)$ 也是理论上最优的平均编码长度，这种编码方式称为熵编码（Entropy Encoding）

交叉熵（Cross Entropy）

- 交叉熵是按照概率分布q的最优编码对真实分布为p的信息进行编码的长度。

$$H(p,q)=\mathbb{E}_p[-\log q(x)]$$

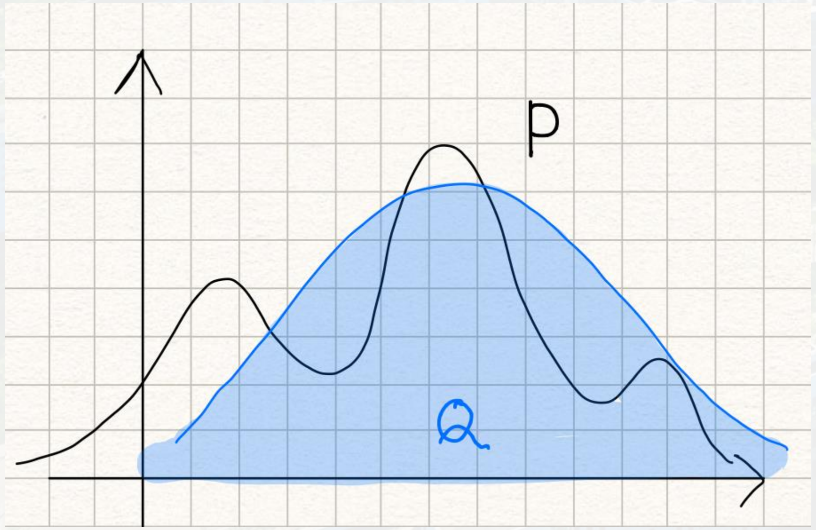
$$=-\sum_x p(x)\log q(x)$$

- 在给定 q 的情况下，如果 p和 q 越接近，交叉熵越小；
- 如果 p 和 q 越远，交叉熵就越大。

KL散度（K-L Divergence）

- KL散度是用概率分布q来近似p时所造成的信息损失量。

  KL散度是按照概率分布q的最优编码对真实分布为p的信息进行编码，其平均编码长度（即交叉熵）$H(p,q)$和p的最优平均编码长度（即熵）$H(p)$之间的差异。



$$\mathrm{KL}(p,q) = H(p,q) - H(p)$$

$$= \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$\int p(x) \log \frac{p(x)}{q(x)} dx$$

交叉熵损失

$$\int p_r(y|x)\log\frac{p_r(y|x)}{p_\theta(y|x)}dy$$

$$\mathrm{D}_{KL}(p_r(y|x)||p_\theta(y|x))$$

$$=\sum_{y=0}^{k}p_r(y|x)\log\frac{p_r(y|x)}{p_\theta(y|x)}$$

KL散度

$$\propto -\sum_{y=0}^{k}p_r(y|x)\log p_\theta(y|x)$$

交叉熵损失

y为x的真实标签

$$\propto -\sum_{y=0}^{k}y_i\log p_\theta(y_i|x)$$

信息论

- 负对数似然损失函数

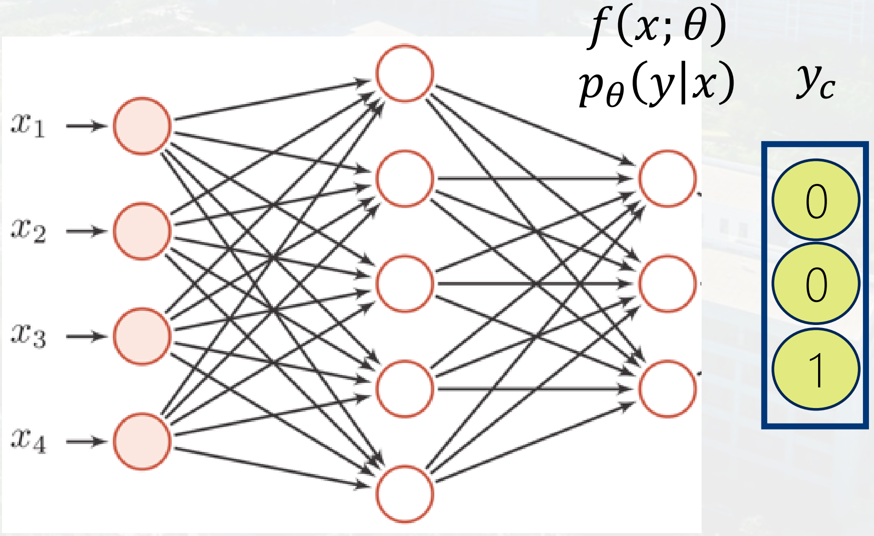
$$\mathcal{L}\big(y, f(x,\theta)\big) = -\sum_{c=1}^{C} y_c \log f_c(x,\theta)$$

- 对于一个三类分类问题，类别为[0,0,1]，

预测类别概率为[0.3,0.3,0.4]，则

$$\mathcal{L}(\theta) = -\begin{pmatrix} 0 \times \log 0.3 + \\ 0 \times \log 0.3 + 1 \times \log 0.4 \end{pmatrix}$$

$$= -\log 0.4$$

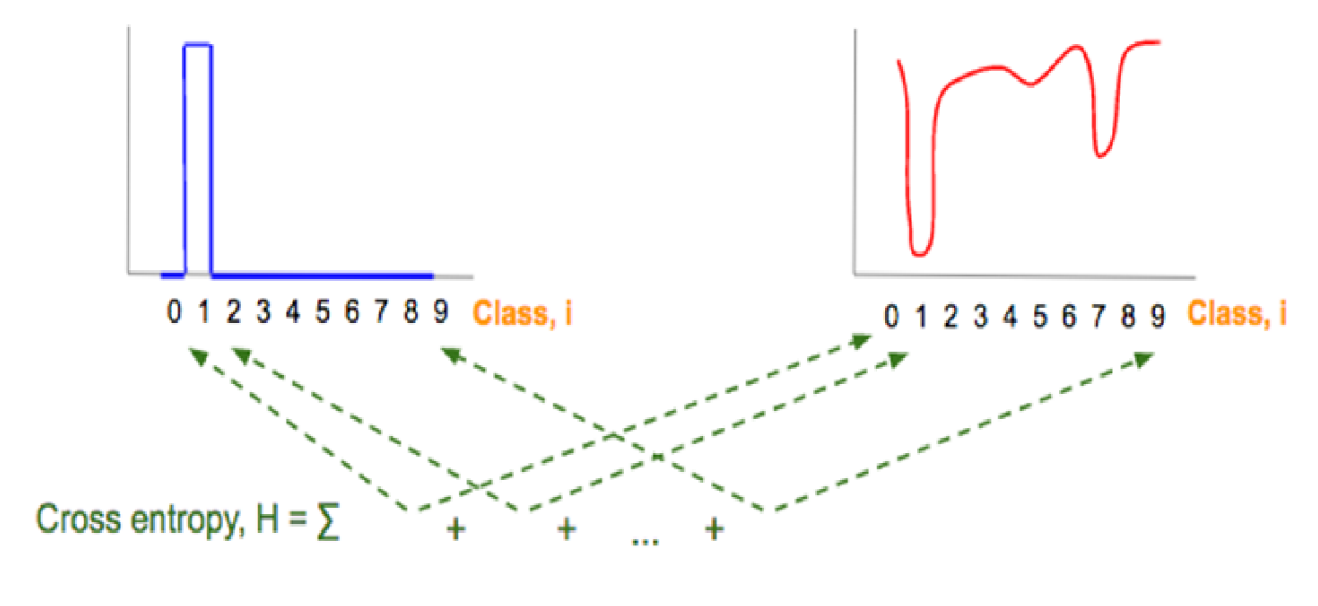


交叉熵损失

$$-\sum_{y=1}^{C} p_r(y|x)\log p_\theta(y|x)$$

真实概率$p_r(y|x)$

预测概率的负对数$-\log p_\theta(y|x)$



THANKS

谢谢大家

中国石油大学（华东）
CHINA UNIVERSITY OF PETROLEUM

汇报人 张琛　　青岛软件学院、计算机科学与技术学院　　2024/2/26