中国科学技术大学
University of Science and Technology of China

机器学习中的数学基础

顾言午

中国科学技术大学，大数据学院

2022 年 9 月 3 日

1. 线性代数
   - 基础概念
   - 矩阵的分解算法
2. 多元微积分
   - 凸集和凸函数
   - 线性规划问题
   - 非线性优化问题
   - KKT 条件
   - 梯度下降法
3. 概率论与数理统计
4. 其他

1 线性代数
- 基础概念
- 矩阵的分解算法

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

4 其他

- 向量、矩阵与张量
- 范数、距离

向量

$$\vec{a} = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}^T$$

or

$$\boldsymbol{a} = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}^T$$

矩阵

$$\boldsymbol{A} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ 1 & x_4 & x_4^2 & x_4^3 \end{bmatrix}$$

张量

$$\boldsymbol{A} = \begin{bmatrix} [0,0,0] & [256,256,0] \\ [0,256,256] & [256,256,256] \end{bmatrix}$$

向量范数:
正则化,防止过拟合

- L1 范数 $\|x\|_1 = \sum_{k=1}^{n} |x_k|$
- L2 范数 $\|x\|_2 = \sqrt{\sum_{k=1}^{n} x_k^2}$
- 无穷范数 $\|x\|_\infty = \max |x_k|$

矩阵范数：
防止过拟合

$$\|A\| = \max \frac{\|Ax\|}{\|x\|}, \|x\| \neq 0$$

- L1 范数 $\|A\|_1 = \max_j \sum_{i=1}^{m} |x_{ij}|$
- L2 范数 $\|A\|_2 = \max \lambda_i(A^H A)$
- 无穷范数 $\|A\|_1 = \max_i \sum_{j=1}^{n} |x_{ij}|$

距离：
验证算法效果

- 曼哈顿距离 $d=\sum_{k=1}^{n}|x_k-y_k|$
- 欧氏距离 $d=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$
- 闵可夫斯基距离 $d=\sqrt[p]{\sum_{k=1}^{n}(x_k-y_k)^p}$
- 切比雪夫距离 $d=\max(|x_k-y_k|)$
- 夹角余弦 $d=\frac{\sum_{k=1}^{n}x_k y_k}{\sqrt{\sum_{k=1}^{n}x_k^2}\sqrt{\sum_{k=1}^{n}y_k^2}}$
- 汉明距离，两串字符串中不同位数的数目

1 线性代数
- 基础概念
- 矩阵的分解算法

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

4 其他

- 特征值分解
- LU 分解
- SVD 分解
- Moore-Penrose 伪逆

通过旋转变换，将矩阵的主要信息转化到对角线上，主成分分析 (PCA)

$$\boldsymbol{A} = \boldsymbol{P}\Delta\boldsymbol{P}^{-1}$$

幂方法

L: 下三角矩阵，U: 上三角矩阵，便于求矩阵的逆，从而计算

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

Cholesky 分解与 Doolitle 分解

奇异值

$$\boldsymbol{A}_{m\times n} = \boldsymbol{U}_{m\times m}\boldsymbol{\Sigma}_{m\times n}\boldsymbol{V}_{n\times n}^{T}$$

其中

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}, \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$$

$\boldsymbol{\Sigma}$ 为对角元为 $\sigma_i(A)$ 的对角矩阵,$\sigma_i(A) = \lambda_i(A^TA)$

QR 分解

$$\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$$

$$\boldsymbol{A}^+ = \boldsymbol{V}\Sigma^{-1}\boldsymbol{U}^T$$

假设我们有一个以向量为自变量的函数

$$f(\boldsymbol{x}) = f(x_1, x_2, ..., x_n)$$

那么

$$\begin{aligned} df &= \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + ... + \frac{\partial f}{\partial x_n} dx_n \\ &= \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix} \begin{pmatrix} dx_1 \\ dx_2 \\ \vdots \\ dx_n \end{pmatrix} \end{aligned}$$

记

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix}$$

记

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial x_1 \partial^2 f}{\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

试试对 $f(\boldsymbol{x})$ 进行泰勒展开？

1 线性代数

2 多元微积分

- 基本概念
- 向量函数、方向梯度与海森矩阵
- 矩阵函数
- 一些优化方法
  - 凸集和凸函数
  - 线性规划问题
  - 非线性优化问题
  - KKT 条件
  - 梯度下降法

3 概率论与数理统计

4 其他

- $\frac{\partial f(\boldsymbol{A})}{\partial \boldsymbol{A}} = (\frac{\partial f(\boldsymbol{A})}{\partial A_{ij}})_{ij}$
- $\frac{\partial \boldsymbol{A}(x)}{\partial x} = (\frac{\partial A_{ij}}{\partial x})_{ij}$
- 请注意，求导的链式法则仍然满足

我们来推导 $\frac{\partial A^{-1}}{\partial x} = -A^{-1}\frac{\partial A}{\partial x}, \frac{\partial \ln det(A)}{\partial A} = A^{-T}$

设 $a, \boldsymbol{a}, \boldsymbol{A}$ 均与 $x, \boldsymbol{x}$ 无关， u, v 均有关, f, u, v 可导

- $\frac{\partial a\boldsymbol{u}}{\partial x} = a\frac{\partial \boldsymbol{u}}{\partial x}$
- $\frac{\partial \boldsymbol{A}\boldsymbol{u}}{\partial x} = \frac{\partial \boldsymbol{u}}{\partial x}\boldsymbol{A}^T$
- $\frac{\partial \boldsymbol{u}^T}{\partial x} = \left(\frac{\partial \boldsymbol{u}}{\partial x}\right)^T$
- $\frac{\partial f(\boldsymbol{u})}{\partial x} = \frac{\partial \boldsymbol{u}}{\partial x}\frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}}$

设 $a, \boldsymbol{a}, A$ 均与 $x, \boldsymbol{x}$ 无关, $f(\boldsymbol{u}), \boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x})$ 可导

- $\frac{\partial \boldsymbol{u}^T\boldsymbol{v}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}\boldsymbol{v} + \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}}\boldsymbol{u}$
- $\frac{\partial A\boldsymbol{u}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}A^T$
- $\frac{\partial \boldsymbol{x}^T A}{\partial \boldsymbol{x}} = A$
- $\frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}\frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}}$
- $\frac{\partial \boldsymbol{x}^T\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}$

尝试计算

$$\frac{\partial \boldsymbol{u}^T A\boldsymbol{v}}{\partial \boldsymbol{x}}, \frac{\partial a\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{b}}{\partial \boldsymbol{x}}$$

设 $a, \boldsymbol{a}, \boldsymbol{A}$ 均与 $x, \boldsymbol{X}$ 无关, $f(\boldsymbol{u}), \boldsymbol{u}(\boldsymbol{X}), \boldsymbol{v}(\boldsymbol{X})$ 可导

- $\frac{\partial \boldsymbol{u}^T \boldsymbol{v}}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{X}} \boldsymbol{v} + \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{X}} \boldsymbol{u}$
- $\frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{X}} \frac{\partial f(\boldsymbol{u})}{\partial \boldsymbol{u}}$
- $\frac{\partial \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^T$

尝试计算

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}}$$

1 线性代数

2 多元微积分
- 基本概念
- 向量函数、方向梯度与海森矩阵
- 矩阵函数
- 一些优化方法
  - 凸集和凸函数
  - 线性规划问题
  - 非线性优化问题
  - KKT 条件
  - 梯度下降法

3 概率论与数理统计

4 其他

凸集：给定集合 $C \subseteq \mathbb{R}^n$. 若 $\forall x, y \in C$ 满足

$$\forall t \in (0,1), tx + (1-t)y \in C$$

那么集合 C 为凸集

凸函数：给定一个函数 $f : \mathbb{R}^n \mapsto R$. 如果满足 $dom(f)$ 是凸集而且 $\forall x, y \subseteq dom(f)$,

$$\forall t \subset [0,1], f(tx + (1-t)y) \leq tf(x) + \ 1 - t)f(y)$$

那么函数 f 是凸函数

一阶条件：假设函数 f 可微，那么 f 是凸函数当且仅当 $\forall x, y \in dom(f)$，

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

二阶条件：假设函数 f 二阶可微，那么 f 是凸函数当且仅当 $\forall x \in dom(f)$

$$\nabla^2 f(x) \succcurlyeq 0,$$

即海森矩阵半正定

Thm：假设函数 f 可微凸函数，那么 x 是 f 的全局最优当且仅当

$$\nabla f(x) = 0$$

$$
\begin{aligned}
\min \quad & c^T x \\
s.t. \quad & A_e x_e = b_e \\
& A_i x_i \leq b_i
\end{aligned}
$$

可基于单纯形法或对偶问题求解

$$\begin{aligned} \min \quad & f(x) \\ s.t. \quad & g(x) = 0 \end{aligned}$$

转化为考虑

$$\min \quad f(x) + \lambda g(x)$$

，其中 λ 可以为任意值. 可以直接求导

$$
\begin{aligned}
\min \quad & c^T x \\
s.t. \quad & g_i(x) \geq 0, i = 1, ..., m \\
& h_i(x) = 0, i = 1, ..., l
\end{aligned}
$$

Kuhn-Tucker 条件:
设 $\bar{x}$ 为约束问题 (67) 的可行点,f 和 $g_i, i \in \mathcal{I}(\bar{x})$ 在点 $\bar{x}$ 可微,$g_i, i \notin \mathcal{I}(\bar{x})$ 在点 $\bar{x}$ 连续, h_j 在点

$\bar{x}$ 连续可微, 向量集 $\{\nabla g_i(\bar{x}), i \in \mathcal{I}(\bar{x}); \nabla h_j(\bar{x}), j = 1, ..., l\}$ 线性无关. 如果 $\bar{x}$ 是局部最优解, 则存在数 $\lambda_i \geq 0$ 和 μ_j 使得

$$\lambda_0 \nabla f(\bar{x}) - \sum_{i \in \mathcal{I}(\bar{x})} \lambda_i \nabla g_i(\bar{x}) - \sum_{j=1}^{l} \mu_j \nabla h_j(\bar{x}) = 0 \tag{82}$$

定义 Lagrange 函数 $L(x, \lambda, \mu) = f(x) - \sum_{i=1}^{m} \lambda_i g_i(x) - \sum_{j=1}^{l} \mu_j h_j(x)$.
若 $\bar{x}$ 为问题局部最优解, 则存在乘子向量 $\bar{\lambda} \geq 0, \bar{\mu}$ 使得

$$\nabla_x L(\bar{x}, \bar{\lambda}, \bar{\mu}) = 0.$$

此时, 一阶必要条件可表达为

$$(K-T) \begin{cases} \nabla_x L(x, \lambda, \mu) = 0 \\ g_i(x) \geq 0, i = 1, ..., m \\ \lambda_i g_i(x) = 0, i = 1, ..., m \\ \lambda_i \geq 0, i = 1, ..., m \\ h_j(x) = 0, j = 1, ..., l \end{cases} \tag{83}$$

$$
\begin{aligned}
f(x) &= f(x_0) + f'(x_0)(x - x_0) \\
0 &= f(x_0) + f'(x_0)(x - x_0) \\
x &= x_0 - \frac{f(x_0)}{f'(x_0)}
\end{aligned}
$$

$$
\begin{aligned}
f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \\
0 &= f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \\
x &= x_0 - \frac{f(\boldsymbol{x}_0)}{\nabla f(\boldsymbol{x}_0)}
\end{aligned}
$$

$$f(x^{(k)} + s) \approx f(x^{(k)}) + g^{(k)T} s + \frac{1}{2} s^T G_k s$$

$$g^{(k)} = \nabla f(x^{(k)}), G_k = \nabla^2 f(x^{(k)})$$

$$\hat{s} = -G_k^{-1} g^{(k)}$$

1 线性代数

2 多元微积分

- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

■ 组合、概率规则和公理 ■ 期望与方差 ■ 分布 ■ 贝叶斯公式、先验与后验 ■ 最大似然估计和最大后验估计

4 其他

请自主复习概率论与数理统计相关知识
大数定律是机器学习的基础

1 线性代数

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计
- 组合、概率规则和公理
- 期望与方差
- 分布
- 贝叶斯公式、先验与后验
- 最大似然估计和最大后验估计

4 其他

请自主复习概率论与数理统计相关知识
理解二者不可兼得

机器学习中比较重要的分布:

- 0-1 分布
- 几何分布
- 二项分布
- (多元) 高斯 (正态) 分布
- 指数分布
- 泊松分布
- 伽玛分布
- 贝塔分布
- 迪利克雷分布

1 线性代数

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计
■ 组合、概率规则和公理 ■ 期望与方差 ■ 分布 ■ 贝叶斯公式、先验与后验 ■ 最大似然估计和最大后验估计

4 其他

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

or

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{\mathbb{R}} f_{Y|X}(y|u)f_X(u)du}$$

先验：在考虑实验之前，我们首先通过经验给出参数的一个分布

后验：结合先验分布和实验数据，更新我们对先验分布的认知

一个例子

假设我们的观测值 x 服从关于 θ 的二项分布，
$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x = 0, 1, ..., n$
我们有先验知识，θ 服从参数为 α, β 的贝塔分布
$\pi(\alpha, \beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, 0 \le \theta \le 1$
如果我们观测到了一个值 x，那么 y 应该服从什么分布？

一个例子

1 线性代数

2 多元微积分

- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

■ 组合、概率规则和公理 ■ 期望与方差 ■ 分布 ■ 贝叶斯公式、先验与后验 ■ 最大似然估计和最大后验估计

4 其他

极大似然估计的核心思想是：认为当前发生的事件是概率最大的事件。因此就可以给定的数据集，使得该数据集发生的概率最大来求得模型中的参数。

$$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta), \theta = \arg\max_{\theta} L(\theta)$$

极大后验估计的核心思想是：允许引入参数的先验分布

$$\theta = \arg\max_{\theta} P(\theta|X) = \arg\max_{\theta} \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$= \arg\max_{\theta} P(X|\theta)P(\theta) = \arg\max_{\theta} L(\theta)P(\theta)$$

假设我们对 n 个观测点 x_i 进行观测得到结果 y_i, 且 $y \sim N(w^T x, \sigma^2)$, 试通过 MLE 和 MAP 去计算 $\hat{w}$.

一个例子

见 https://zhuanlan.zhihu.com/p/86009986

1 线性代数

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

4 其他
- 熵

1 线性代数

2 多元微积分
- 凸集和凸函数
- 线性规划问题
- 非线性优化问题
- KKT 条件
- 梯度下降法

3 概率论与数理统计

4 其他
- 熵

给出信息熵的公式

$$H(X) = -\sum_{i=1}^{n} p(x_i) log(p(x_i))$$

信息熵 H 作为对随机实验不确定程度的度量，满足三个规则：

- H 是 p 的连续函数；
- 对于等概结果为 n 的随机实验，H 是 n 的单调递增函数；
- 组合可加性

$$\begin{aligned} H_n(p_1, p_2, ..., p_n) =& H_{n-1}(p_1 + p_2, p_3, ..., p_n) \\ & + (p_1 + p_2) H_2(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}) \end{aligned}$$

假设现在有一个样本集中两个概率分布 p, q，其中 p 为真实分布，q 为非真实分布。假如，按照真实分布 p 来衡量识别一个样本所需要的编码长度的期望即位信息熵 $-\sum_{i=1}^{n} p(x_i) log(p(x_i))$.
如果采用错误的分布 q 来表示来自真实分布 p 的平均编码长度，则应该是交叉熵：

$$H(p, q) = -\sum_{i=1}^{n} p(x_i) log(q(x_i))$$

KL 散度公式为:

$$D(p\|q)=\sum_{i=1}^{n}p(x_i)log(\frac{p(x_i)}{q(x_i)})$$

- 不对称性
- 非负性

谢谢!