基本概念及理解

Atri

日期：May 26, 2023

无论是听课还是自学, 我发现大部分同学都存在一个问题, 就是对某些基础概念相当陌生, 这对听课效率乃至后面自学都产生了较大的影响. 因此, 我打算专门做个 LaTeX 的 note 来总结一下.

1 超平面 (hyperplane)

课件所给的超平面的形式:

$$L = \left\{ \boldsymbol{x} \in \mathbb{R}^k | A\boldsymbol{x} = \boldsymbol{b}, A \in \mathbb{R}^{d \times k} \right\}.$$

显然, 我完全不知道为什么长这样. 问了 TA 表示也不知道. 那我们就暂且先不管这个定义.

给出维基百科对 hyperplane 的定义: In geometry a hyperplane is a subspace of one dimension less than its ambient space. 也就是说, 超平面是比当前空间少一个维度的子空间, 这个子空间把该空间分成了两个部分.

设 $\boldsymbol{x}_0$ 是超平面上的点, $\boldsymbol{\omega}$ 为超平面的法向量. 根据法向量正交于任何超平面上的向量的性质, 可以得出, 对超平面上任意一点 $\boldsymbol{x}$, 有:

$$\boldsymbol{\omega}^\top (\boldsymbol{x} - \boldsymbol{x}_0) = 0.$$

即

$$\boldsymbol{\omega}^\top \boldsymbol{x} = \boldsymbol{\omega}^\top \boldsymbol{x}_0 = \boldsymbol{b}.$$

因此可以得出超平面的一般形式:

$$\boldsymbol{\omega}^\top \boldsymbol{x} = \boldsymbol{b}.$$

可以认为, $\boldsymbol{\omega}$ 决定超平面的姿态, 而 $\boldsymbol{b}$ 决定超平面的位置.

2 投影

对于空间内的一条直线, 可以使用点向式定义:

$$S = \{ \boldsymbol{y} | \boldsymbol{y} = \boldsymbol{p} + t\boldsymbol{v}, t \in \mathbb{R} \}.$$

这里主要研究点到直线的投影.

根据生活常识, 投影的定义是直线上与空间中的点相对应的一个点, 二者满足距离最小, 即

$$\min_{t \in \mathbb{R}} \| \boldsymbol{x} - (\boldsymbol{p} + t\boldsymbol{v}) \|^2 = \| \boldsymbol{x} - \boldsymbol{p} \|^2 + t^2 \| \boldsymbol{v} \|^2 - 2t\boldsymbol{v}^\top (\boldsymbol{x} - \boldsymbol{p}).$$

为了方便, 令方向向量 $\boldsymbol{v}$ 为单位向量, 因此我们只需求出放缩系数 t :

$$t = \arg\min_{t \in \mathbb{R}} t^2 - 2t\boldsymbol{v}^\top (\boldsymbol{x} - \boldsymbol{p}) + \| \boldsymbol{x} - \boldsymbol{p} \|^2.$$

这是关于 t 的二次函数, 极小值点 $t^* = \boldsymbol{v}^\top (\boldsymbol{x} - \boldsymbol{p})$, 代入可得投影点:

$$\widehat{\boldsymbol{x}} = \boldsymbol{p} + (\boldsymbol{v}^\top (\boldsymbol{x} - \boldsymbol{p}))\boldsymbol{v}.$$

若直线过原点, 可得

$$\widehat{\boldsymbol{x}} = (\boldsymbol{v}^\top \boldsymbol{x})\boldsymbol{v}.$$

3 协方差矩阵

在概率论中, 随机变量 $\mathbf{X}$ 和 $\mathbf{Y}$ 的协方差为:

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}\left\{[\mathbf{X} - \mathbb{E}(\mathbf{X})][\mathbf{Y} - \mathbb{E}(\mathbf{Y})]\right\}.$$

拓展到高维的情况, 其协方差矩阵 C 的第 i 行第 j 列为: $\mathrm{Cov}(\mathbf{X}_i, \mathbf{X}_j)$. 显然协方差矩阵是一个对称矩阵.
然而, 在概率论中, 我们对样本协方差的定义是一个统计量, 有以下两种形式 (假设样本均值为 0):

$$S = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top \quad \text{or} \quad S = \frac{1}{N-1}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

前者是课件中所用的形式, 在概率论中可作为协方差的最大似然估计；相对应的, 后者是协方差的一种无偏估计.
其中, 以下秩一矩阵的和, 被称为散布矩阵, 它是个半正定矩阵:

$$\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

在 FLDA(线性判别分析) 的推导中也有用到相关概念: https://zhuanlan.zhihu.com/p/625837046.

4 矩阵求导的本质

高等数学的时候我们学过偏导, 它是将一个 function 对所有自变量分别求导.
矩阵求导也是一样的, 本质就是 function 中的每个 f 分别对变元中的每个元素逐个求偏导, 只不过写成了向量、矩阵的形式而已.

4.1 向量变元的实值标量函数布局

直观上来看, **分子布局**, 就是分子是列向量的形式, 分母是行向量的形式:

$$\frac{\partial \boldsymbol{f}_{2\times 1}(\boldsymbol{x})}{\partial \boldsymbol{x}_{3\times 1}^\top} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} \end{bmatrix}_{2\times 3}$$

而**分母布局**, 就是分子是行向量的形式, 分母是列向量的形式:

$$\frac{\partial \boldsymbol{f}_{2\times 1}^\top(\boldsymbol{x})}{\partial \boldsymbol{x}_{3\times 1}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} \\ \frac{\partial f_1}{\partial x_3} & \frac{\partial f_2}{\partial x_3} \end{bmatrix}_{3\times 2}$$

4.2 矩阵变元的实值标量函数布局

这里只介绍矩阵的形式.

Jocabian 矩阵形式, 即先把矩阵变元 $\boldsymbol{X}$ 进行转置, 然后逐分量求导:

$$\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}_{m\times n}^{\top}} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{21}} & \cdots & \frac{\partial f}{\partial x_{m1}} \\ \frac{\partial f}{\partial x_{12}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{m2}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f}{\partial x_{1n}} & \frac{\partial f}{\partial x_{2n}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix}_{n\times m}$$

梯度向量形式, 即直接逐分量求导, 为上述形式的转置.

5 链式法则

在**动手学深度学习**这门课中, 给出的链式法则公式如下:

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

然而, Lect3(or cmu-10315) 给出的微分的链式法则: (令 $\boldsymbol{u} = g(\boldsymbol{x})$)

$$\nabla f \circ g(\boldsymbol{x}) = \frac{\partial f \circ g(\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}\frac{\partial f}{\partial \boldsymbol{u}}.$$

可以发现二者顺序相反. 个人认为应该是**动手学深度学习**这门课所选用的求导为分子布局, 而本课程用的是分母布局.

However, **动手学深度学习**所给链式法则更合乎常人理解, 其也是计算图上实现自动微分的原理.

6 凸

如果一个优化问题可以被转化为一个凸优化问题, 那么这个问题就算解决了（

6.1 凸集

定义 6.1. 集合 C 被称为凸集, 当且仅当对任意的 $\boldsymbol{x}, \boldsymbol{y} \in C$ 及 $0 \le \theta \le 1$, 都有

$$\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y} \in C.$$

可以对比一下锥的概念: $\theta_1, \theta_2 > 0$, 不限制 $(\theta_1 + \theta_2 = 1)$, 都有

$$\theta_1 \boldsymbol{x} + \theta_2 \boldsymbol{y} \in C.$$

可以理解为两向量所**夹住**的一块锥形区域.

However, 在 COP 中用到的切锥的概念, 跟锥似乎一点关系没有?

6.2 凸函数

主定义: $f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \le \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$, 严格凸对于 $\lambda \in (0,1)$ 不等式严格成立.

中点凸: $f((\boldsymbol{x}+\boldsymbol{y})/2) \le (f(\boldsymbol{x}) + f(\boldsymbol{y}))/2$.

上面二者等价.

保凸运算: 设所有凸函数的集合为 L.

- $f_1 \in L, f_2 \in L \Rightarrow f_1 + f_2 \in L$
- $f \in L, \alpha > 0 \Rightarrow \alpha f \in L$
- $f \in L \Rightarrow f(A\boldsymbol{x} + b) \in L$.

6.2.1 凸函数的判定

- 定义.
- 一阶条件: $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})$.
- 二阶条件: $\nabla^2 f(\boldsymbol{x}) \succeq 0$, 若正定为严格凸函数.
- 构造函数: $g = f(\boldsymbol{x} + t\boldsymbol{v})$ 为凸.
- 上方图为凸集.

6.2.2 凸函数的性质

- 所有的下水平集 $S_\alpha = \{\boldsymbol{x} \in \mathrm{dom}(f) | f(\boldsymbol{x}) \leq \alpha\}$ 为凸集.
- 一、二阶条件.
- 上方图为凸集.

6.2.3 强凸

μ-强凸定义: $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$.
另一个定义: $h = f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|^2$ 为凸函数.
一个性质: $(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^\top (\boldsymbol{x} - \boldsymbol{y}) \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|^2$.
对比一下 L-光滑函数:

$$
\begin{aligned}
&f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 \\
&h = \frac{L}{2}\|\boldsymbol{x}\|^2 - f(\boldsymbol{x}) \text{ is convex} \\
&(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^\top (\boldsymbol{x} - \boldsymbol{y}) \leq L\|\boldsymbol{x} - \boldsymbol{y}\|^2
\end{aligned}
$$

7 切空间

7.1 曲面

工科数学分析中, 我们在三维空间中常见的曲面有柱面、球面、旋转抛物面、双曲面等等. 一般来讲, 如果想用什么东西来表示一个曲面, 我们可能会去挖掘坐标 (x, y, z) 的信息.
例如单位球面:

$$x^2 + y^2 + z^2 = 1.$$

然而, 事实上, 对于三维空间内的一条曲线, 也可以把它归结为曲面. 例如二次函数曲线:

$$
\begin{cases}
y = x^2 \\
z = 0
\end{cases}
$$

因此, 我们需要用维度来衡量一个曲面. 然而, 无论是上述的简单曲面和简单曲线, 都是处于 $\mathbb{R}^3$ 空间内的.
对于一个高维曲面, 我们或许可以用一种映射来表示:

$$M : \mathbb{R}^n \to \mathbb{R}^m, \boldsymbol{x} \mapsto F = (F_1(\boldsymbol{x}), \cdots, F_m(\boldsymbol{x})).$$

也就是说, 对于三维空间即 $m = 3$, 我们或许可以用最少的相互独立的 n 个变量来确定空间, 每一个 $F_i(\boldsymbol{x})$ 或许代表一个坐标内的一个分量.
例如单位球面, 我们可以用两个角度 θ, α 来表示:

$$\begin{cases} x = \sin\theta\sin\alpha \\ y = \sin\theta\cos\alpha \\ z = \cos\theta \end{cases}$$

以及柱面, 可以用角度和高度表示:

$$\begin{cases} x = \sin\theta \\ y = \cos\theta \\ z = h \end{cases}$$

(以上皆为个人 yy, 毕竟找不到什么好的参考材料).
另外, 或许你熟悉子空间的维度的定义, 因此需要给出切空间的说明, 然后你会发现, 曲面上某点的切空间的维度就是曲面的维度.

7.2 切空间

定义 7.1. 设向量 $\boldsymbol{x} \in \mathbb{R}^n$, $V \subset \mathbb{R}^n$ 是一个线性子空间, 如果 $\forall \boldsymbol{y} \in V$, 都有

$$\boldsymbol{x}^\top \boldsymbol{y} = 0.$$

则称向量 $\boldsymbol{x}$ 正交于子空间 V.

假设用坐标表示的曲面为 $h(\boldsymbol{x}) = 0, \boldsymbol{x} \in \mathbb{R}^n$, $t \mapsto \boldsymbol{x}(t)$ 为曲面上的一条曲线 (一维曲面). 因此曲线可以写成 $h(x_1(t), \cdots, x_n(t)) = 0$. 两边对 t 求导:

$$h^{'}(t) = \nabla h(\boldsymbol{x})^\top \boldsymbol{x}^{'}(t) = 0.$$

即, 切线 $\boldsymbol{x}^{'}(t)$ 正交于曲面 $h(\boldsymbol{x}) = 0$.
我们把点 $\boldsymbol{x}$ 上所有的曲面的曲线的切线的集合称之为切空间 $T_{\boldsymbol{x}}M$.
在有约束的最优化问题中, 对于判断一个 KKT 点是否为极值点, 我们通常需要用到上述切空间的定义:

$$T(\boldsymbol{x}) = \left\{ \boldsymbol{y} \in \mathbb{R}^n | \nabla h(\boldsymbol{x})^\top \boldsymbol{y} = 0 \right\}.$$