凸优化 [1]

July 15, 2020

1 几何体的向量表示

在介绍凸集等概念之前，首先介绍一下空间几何体的向量表示，下面在定义凸集概念时使用到了线段的线段表示。先通过一个例子来认识一下如何使用向量表示线段。

已知二维平面上两定点 $A(5,1)$、$B(2,3)$，给出线段 AB 的方程表示如下：

$$\begin{cases} x_1 = \theta \times 5 + (1-\theta) \times 2 \\ x_2 = \theta \times 1 + (1-\theta) \times 3 \end{cases} \qquad \theta \in [0,1]$$

如果将点 A 看成向量 $\vec{a}$，点 B 看成向量 $\vec{b}$，则线段 AB 的向量表示为：

$$\vec{x} = \theta\vec{a} + (1-\theta)\vec{b}, \quad \theta \in [0,1]$$

而直线的向量表示是：

$$\vec{x} = \theta\vec{a} + (1-\theta)\vec{b}, \quad \theta \in \mathbb{R}$$

由此衍生推广到高维，可得以下几何体的向量表示，三角形的向量表示：

$$\vec{x} = \theta_1\vec{a_1} + \theta_2\vec{a_2} + \theta_3\vec{a_3}, \quad \theta_i \in [0,1], \sum \theta_i = 1$$

三维平面的向量表示：

$$\vec{x} = \theta_1\vec{a_1} + \theta_2\vec{a_2} + \theta_3\vec{a_3}, \quad \theta_i \in \mathbb{R}, \sum \theta_i = 1$$

超几何体的向量表示：

$$\vec{x} = \theta_1\vec{a_1} + \theta_2\vec{a_2} + \cdots + \theta_k\vec{a_k}, \quad \theta_i \in [0,1], \sum \theta_i = 1$$

超平面的向量表示：

$$\vec{x} = \theta_1\vec{a_1} + \theta_2\vec{a_2} + \cdots + \theta_k\vec{a_k}, \quad \theta_i \in \mathbb{R}, \sum \theta_i = 1$$

2 凸集凸函数定义

2.1 凸集

集合 C 内任意两点间的线段也均在集合 C 内，则称集合 C 为凸集，数学定义为：

$$\forall x_1, x_2 \in C, \forall \theta \in [0,1], 则 x = \theta \cdot x_1 + (1-\theta) \cdot x_2 \in C$$

k 个点的版本：

$$\forall x_1, x_2, \cdots x_k \in C, \theta_i \in [0,1], 且 \sum \theta_i = 1, 则 x = \sum_{i=1}^{k} \theta_i x_i \in C$$

上面凸集定义中便用到了线段的向量表示，含义是如果点 x_1 和点 x_2 在集合 C 内，则线段 x_1x_2 上所有点都在集合 C 内，凸集的交集仍是凸集。

2.2 凸函数

凸函数定义为，设 $f \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^1$，$C$ 是凸集，对于 $x_1, x_2 \in C$ 都有：

$$f(\alpha_1 x_1 + \alpha_2 x_2) \leq \alpha_1 f(x_1) + \alpha_2 f(x_2) \quad \sum \alpha_i = 1, \alpha_i \geq 0$$

则称 $f(x)$ 为定义在凸集 C 上的凸函数。

严格凸函数定义为，设 $f \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^1$，$C$ 是凸集，对于 $x_1, x_2 \in C$ 都有：

$$f(\alpha_1 x_1 + \alpha_2 x_2) < \alpha_1 f(x_1) + \alpha_2 f(x_2) \quad \sum \alpha_i = 1, \alpha_i \geq 0$$

则称 $f(x)$ 为定义在凸集 C 上的严格凸函数。

凸函数的等价定义：设 $f \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^1$，$C$ 是凸集，对于 $x_1, x_2, x_3 \in C$ 且 $x_1 < x_2 < x_3$，下式成立则 $f(x)$ 为凸函数 (?? 待证明??)：

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_3) - f(x_1)}{x_3 - x_1} \leq \frac{f(x_3) - f(x_2)}{x_3 - x_2}$$

3 凸函数各种性质及其证明

3.1 性质-下水平集

设 $f \subseteq \mathbb{R}^n \to \mathbb{R}^1$，$C$ 是凸集，若 f 是凸函数，则对于 $\forall \beta$，证明下面水平集 D_β 是凸集

$$D_\beta = \{x | f(x) \leq \beta, x \in C\}$$

证明：若对于 $\forall x_1, x_2 \in D_beta$，则线段 x_1x_2 为：

$$x_1x_2 = \alpha_1 x_1 + \alpha_2 x_2, \alpha_i > 0, \sum \alpha_i = 1$$

且因为 $f(x)$ 是凸函数，所以：

$$f(\alpha_1 x_1 + \alpha_2 x_2) \leq \alpha_1 f(x_1) + \alpha_2 f(x_2) \leq \alpha_1 \beta + \alpha_2 \beta = \beta$$

所以原命题得证，水平集 D_β 是凸集。

3.2 性质-全局最优

凸优化问题的局部极小值是全局极小值。

这个性质是凸优化问题一个良好的性质，在机器学习任务中我们只需将非凸问题转化为凸优化问题，便可直接求出问题的全局极值，下面给出证明：

反证法证明：假定 x^* 是局部极小点，不是全局极小点，则一定存在 x' 使得：

$$f(x') < f(x^*)$$

构造凸组合点 $x = \alpha_1 x' + \alpha_2 x^*$，$\sum \alpha_i = 1$，有：

$$f(x) = f(\alpha_1 x' + \alpha_2 x^*) \leq \alpha_1 f(x') + \alpha_2 f(x^*) < \alpha_1 (f^*) + \alpha_2 f(x^*) = f(x^*)$$

当 $\alpha_2 \to 1$ 时，x 点无限趋近于 x^*，即 x 是 x^* 邻域中的一点，但仍然存在 $f(x) < f(x^*)$，与 x^* 是局部极小点矛盾；因此原问题得证。

3.3 性质-梯度

设 $f \subseteq \mathbb{R}^n \to \mathbb{R}^1$，$C$ 是凸集，对于 $x_1, x_2 \in C$，

(1) f 为凸函数的充要条件是：对于 $\forall x_1, x_2 \in C$ 且 $x_1 \neq x_2$ 都有：

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1)$$

(2) f 为严格凸函数的充要条件是：对于 $\forall x_1, x_2 \in C$ 且 $x_1 \neq x_2$ 都有：

$$f(x_2) > f(x_1) + \nabla f(x_1)^T (x_2 - x_1)$$

证明：先证明充分性，即 $f(x)$ 是凸函数 $\Rightarrow f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1)$

因为 $f(x)$ 是凸函数，所以 $\alpha_1, \alpha_2 \in [0,1], \sum \alpha_i = 1$，有：

$$f(\alpha_1 x_1 + \alpha_2 x_2) \leq \alpha_1 f(x_1) + \alpha_2 f(x_2)$$

令 $\alpha_1 = 1 - t, \alpha_2 = t$，有：

$$f((1-t)x_1 + t x_2) \leq (1-t) f(x_1) + t f(x_2)$$

$$f(x_1 + t(x_2 - x_1)) \leq f(x_1) + t(f(x_2) - f(x_1))$$

$$\frac{f(x_1 + t(x_2 - x_1)) - f(x_1)}{t} \leq f(x_2) - f(x_1)$$

用极限的思想：

$$\lim_{t \to 0} \frac{f(x_1 + t(x_2 - x_1)) - f(x_1)}{t} = \nabla f(x_1)^T (x_2 - x_1)$$

所以：

$$\nabla f(x_1)^T (x_2 - x_1) \leq f(x_2) - f(x_1)$$

故充分性得证；

再证明必要性，即 $f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) \Rightarrow f(x)$ 是凸函数

$\forall x_1, x_2 \in C$，有 $x = \alpha_1 x_1 + \alpha_2 x_2$，其中 $\alpha_1, \alpha_2 \in [0,1], \sum \alpha_i = 1$，所以:

$$f(x_1) \geq f(x) + \nabla f(x)^T (x_1 - x)$$

$$f(x_2) \geq f(x) + \nabla f(x)^T (x_2 - x)$$

所以：

$$\alpha_1 f(x_1) \geq \alpha_1 f(x) + \nabla f(x)^T \alpha_1 (x_1 - x)$$

$$\alpha_2 f(x_2) \geq \alpha_2 f(x) + \nabla f(x)^T \alpha_2 (x_2 - x)$$

$$\alpha_1 f(x_1) + \alpha_2 f(x_2) \geq (\alpha_1 + \alpha_2) f(x) + \nabla f(x)^T (\alpha_1 x_1 + \alpha_2 x_2 - (\alpha_1 + \alpha_2) x)$$

因为 $\alpha_1 + \alpha_2 = 1$，所以有：

$$\alpha_1 f(x_1) + \alpha_2 f(x_2) \geq f(x) = f(\alpha_1 x_1 + \alpha_2 x_2)$$

故必要性得证。

3.4 性质-Hessian 矩阵

凸函数的 Hessian 矩阵半正定。

该性质描述了凸函数二阶可微时满足的性质，即凸函数的 Hessian 矩阵半正定，此性质可通过泰勒公式进行，在给出该性质证明之前，先给出 Hessian 矩阵和泰勒公式定义。

Hessian 矩阵是一个多元函数的二阶偏导数构成的方阵，一个多元函数 Hessian 矩阵定义如下：

$$H(f) = J(J(f)) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

一般情况下，多元函数的混合二阶偏导数与求导次序无关 [2]，即：

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

因此 Hessian 矩阵是一个对称矩阵，它可以看作二阶导数对多元函数的推广。Hessian 矩阵简写为 $\nabla^2 f(x)$ 。对于如下多元函数：

$$f(x, y, z) = 2x^2 - xy + y^2 - 3z^2$$

它的 Hessian 矩阵为：

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{bmatrix} = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & -6 \end{bmatrix}$$

根据多元函数极值判别法，假设多元函数在点 M 的梯度为 0，即 M 是函数的驻点，则有：

- 如果 Hessian 矩阵正定，函数在该点有极小值
- 如果 Hessian 矩阵负定，函数在该点有极大值
- 如果 Hessian 矩阵不定，则不是极值点（鞍点）

这可以看做是一元函数极值判别法对多元函数对推广，Hessian 矩阵正定类似于二阶导数大于 0，其他的以此类推。对于 n 阶矩阵 A，对于任意非 0 的 n 维向量 x 都有：

$$x^T Ax > 0$$

则称矩阵 A 为正定矩阵。判定矩阵正定的常用方法有以下几种：

- 矩阵的特征值全大于 0；
- 矩阵的所有顺序主子式都大于 0；
- 矩阵合同于单位阵 I；

泰勒公式是用若干项连加来表示一个函数，这些相加的项由函数在某一点的导数求得，下面给出一个函数 $f(x)$ 在 $x = a$ 点的泰勒展开式：

$$f(x) = \frac{f(a)}{0!} + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n(x)$$

证明：Hessian 矩阵可看做：$\nabla^2 f(x)$

对于凸函数 $f(x)$，有 $f(x) \geq f(x_0) + \nabla f(x_0)^T (x - x_0)$，

将 $f(x)$ 进行泰勒展开，有

$$f(x) = f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)^T (x - x_0) + R(n)$$

其中 $R(n)$ 是余项，关于 $(x - x_0)^2$ 的高阶无穷小，即 $R(n) \to 0$

因此：

$$\frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)^T (x - x_0) \geq 0$$

其中 $x - x_0 \in \mathbb{R}^n, H(f) = \nabla^2 f(x_0)^T$，所以：

$$(x - x_0)^T \nabla^2 f(x_0)^T (x - x_0) \geq 0$$

因此 H 半正定，问题得证。

3.5 性质

若 $x \in \mathbb{R}^n, y \in \mathbb{R}^n$（即 x, y 为 n 维空间向量），Q 为半正定对称阵，则 $f(x) = X^T QX$ 为凸函数

证明：要证明 $f(x) = X^T QX$ 为凸函数，需要证明：

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$$

其中，$x \in \mathbb{R}^n, y \in \mathbb{R}^n, \alpha \in [0,1]$

令

$$
\begin{aligned}
F &= f(\alpha x + (1-\alpha)y) - \alpha f(x) - (1-\alpha)f(y) \\
&= (\alpha x + (1-\alpha)y)^T Q(\alpha x + (1-\alpha)y) - \alpha x^T Qx - (1-\alpha)y^T Qy \\
&= \alpha^2 x^T Qx + (1-\alpha)^2 y^T Qy + \alpha(1-\alpha)x^T QY + \alpha(1-\alpha)y^T Qx - \alpha x^T Qx - (1-\alpha)y^T Qy
\end{aligned}
$$

因为 Q 是对称阵，即 $Q^T = Q$，所以：

$$
x^T Qy = (y^T Q^T x)^T = (y^T Qx)^T = y^T Q^T x
$$

最后一步等号是因为 $(y^T Qx)^T \in \mathbb{R}^1$，所以 $(y^T Qx)^T = y^T Q^T x$

所以：

$$
F = \alpha(\alpha-1)(x^T Qx - 2x^T Qy + y^T Qy) = \alpha(\alpha-1)(x-y)^T Q(x-y)
$$

因为 Q 是半正定阵，所以对 $(x-y) \in \mathbb{R}^n$，有

$$
(x-y)^T Q(x-y) \geq 0
$$

所以 $F \leq 0$，所以 $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$，即问题得证

3.6 性质-Jensen 不等式前奏

凸函数 $f(x)$，其中 $Q_1 + Q_2 + ... + Q_n = 1, 0 <= Q_i <= 1$，证明下面不等式：

$$
f(Q_1x_1 + Q_2x_2 + \cdots + Q_nx_n) \leq Q_1f(x_1) + Q_2f(x_2) + \cdots + Q_nf(x_n)
$$

使用数学归纳法证明：

假设，$n = 2$ 时成立，即：$f(Q_1x_1 + Q_2x_2) \leq Q_1f(x_1) + Q_2f(x_2)$

假设，$n = k$ 时成立，即：$f(Q_1x_1 + Q_2x_2 + \cdots + Q_kx_k) \leq Q_1f(x_1) + Q_2f(x_2) + \cdots + Q_kf(x_k)$

则当 $n = k+1$ 时，有：

$$
\begin{aligned}
F &= Q_1f(x_1) + Q_2f(x_2) + \cdots + Q_kf(x_k) + Q_{k+1}f(x_{k+1}) \\
&= (1-Q_{k+1})\frac{1}{(1-Q_{k+1})}(Q_1f(x_1) + Q_2f(x_2) + \cdots + Q_kf(x_k)) + Q_{k+1}f(x_{k+1})
\end{aligned}
$$

根据 $n=k$ 时，$\frac{\sum_{i=1}^{k} Q_i}{1-Q_{k+1}}=1$, 有

$$
\begin{aligned}
F &\geq (1-Q_{k+1}) f(\frac{Q_1 f(x_1)+Q_2 f(x_2)+\cdots+Q_k f(x_k)}{1-Q_{k+1}})+Q_{k+1} f(x_{k+1}) \\
&\geq f(Q_1 x_1+Q_2 x_2+\cdots+Q_k x_k+Q_{k+1} x_{k+1})
\end{aligned}
$$

所以命题得证。

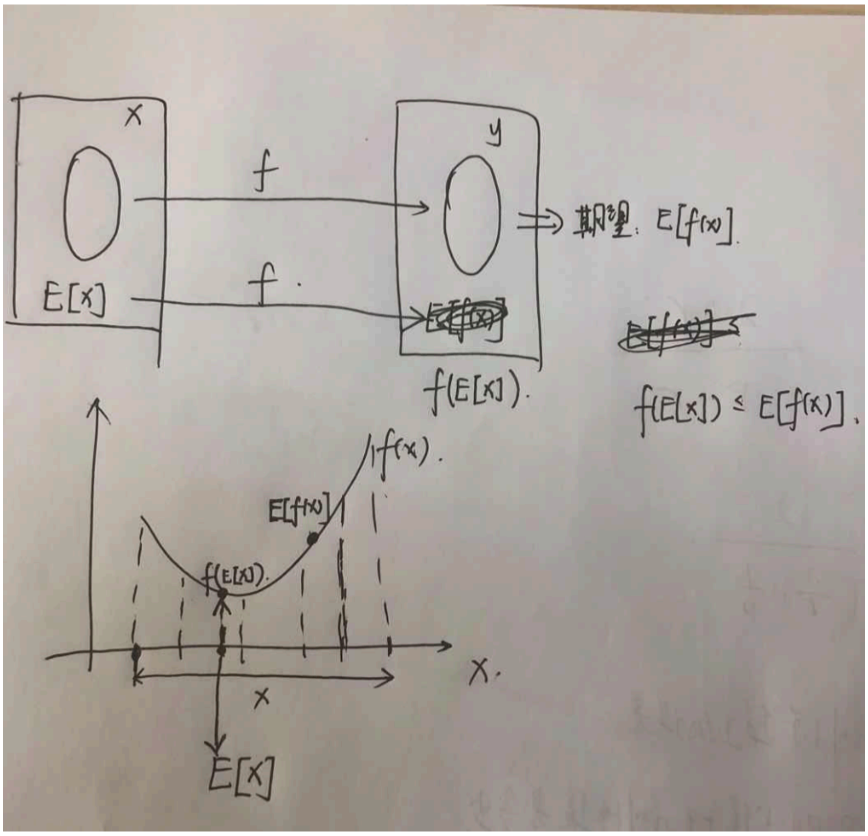
3.7 性质-Jensen 不等式

Jessen 不等式，$f(x)$ 为凸函数，其中 $E(x)$ 是 x 的期望，则:

$$
f(E[x]) \leq E[f(x)]
$$

根据上述性质，可以快速得证。

Jensen 不等式的直观理解:



4　凸优化

4.1　凸优化问题定义

一个凸优化问题可描述为:

$$f(x) \quad s.t. \quad x \in C$$

$$s.t. \quad g_i(x) <= 0, h_i(x) = 0$$

通过以下凸优化性质便可理解何为凸优化问题:

(1) 目的是求解目标函数的最小值;

(2) 目标函数 $f(x)$ 和不等式约束函数 $g(x)$ 都是凸函数，定义域是凸集;

(3) 若存在等式约束函数，则等式约束函数 $h(x)$ 为仿射函数；仿射函数指的是最高次数为 1 的多项式函数，一般形式为 $f(x) = Ax + b$，A 是 $m * k$ 矩阵，x 是一个 k 向量，b 是一个 m 向量

(4) 凸优化问题有一个良好的性质即：局部最优解便是全局最优解

4.2　常见凸优化问题

4.2.1　线性规划 LinearProgramming(LP)

如果目标函数和不等式约束函数都是仿射函数，则凸优化问题称为线性规划，数学表示为:

$$C^T x + d$$

$$s.t. \quad Gx <= h, Ax = b$$

4.2.2　二次规划 QuadraticProgramming(QP)

如果目标函数是凸二次函数，而不等式约束仍是仿射函数，则凸优化问题称为二次规划，数学表示为:

$$\frac{1}{2} x^T P x + c^T x + d$$

$$s.t. \quad Gx <= h, Ax = b$$

4.2.3　二次约束的二次规划 Quadratically Constrained Quadratic Programming(QCQP)

如果目标函数和不等书约束均为凸二次函数，则凸优化问题称为二次约束的二次规划，数学表示为:

$$\frac{1}{2} x^T P x + c^T x + d$$

$$s.t. \quad \frac{1}{2} x^T Q_i x + r_i^T x + s_i <= 0, \quad i = 1, 2, \cdots, m, \quad Ax = b$$

4.2.4 半正定规划 Semidefinite Programming(SDP)

半正定规划较前面的复杂，在机器学习中也经常用到，下面给出数学描述：

$$tr(CX)$$

$$s.t. \quad tr(A_i x) = b_i \quad i = 1, 2, \cdots, p, \quad X >= 0$$

其中符号 $tr(A)$ 表示矩阵 A 的迹，矩阵 A 的迹是指 A 的对角线上各个元素的总和

5 浅谈凸优化问题为何如此重要

1、凸优化具有良好性质，如局部最优解是全局最优解，且凸优化问题是多项式时间可解问题，如：线性规划问题；

2、很多非凸优化或 NP-Hard 问题可以转化成凸优化问题，方法：对偶、松弛 (扩大可行域，去掉部分约束条件)，在 SVM 算法中，为了对目标函数进行优化，便使用了拉格朗日乘子法、对偶问题、引入松弛因子等。

References

[1] “机器学习概念篇：一文详解凸函数和凸优化，干货满满.” [Online]. Available: https://zhuanlan.zhihu.com/p/51127402

[2] “理解凸优化.” [Online]. Available: https://zhuanlan.zhihu.com/p/37108430