# 机器学习中的数学第二期第7课:凸优化简介

七月在线 管老师

2018年3月

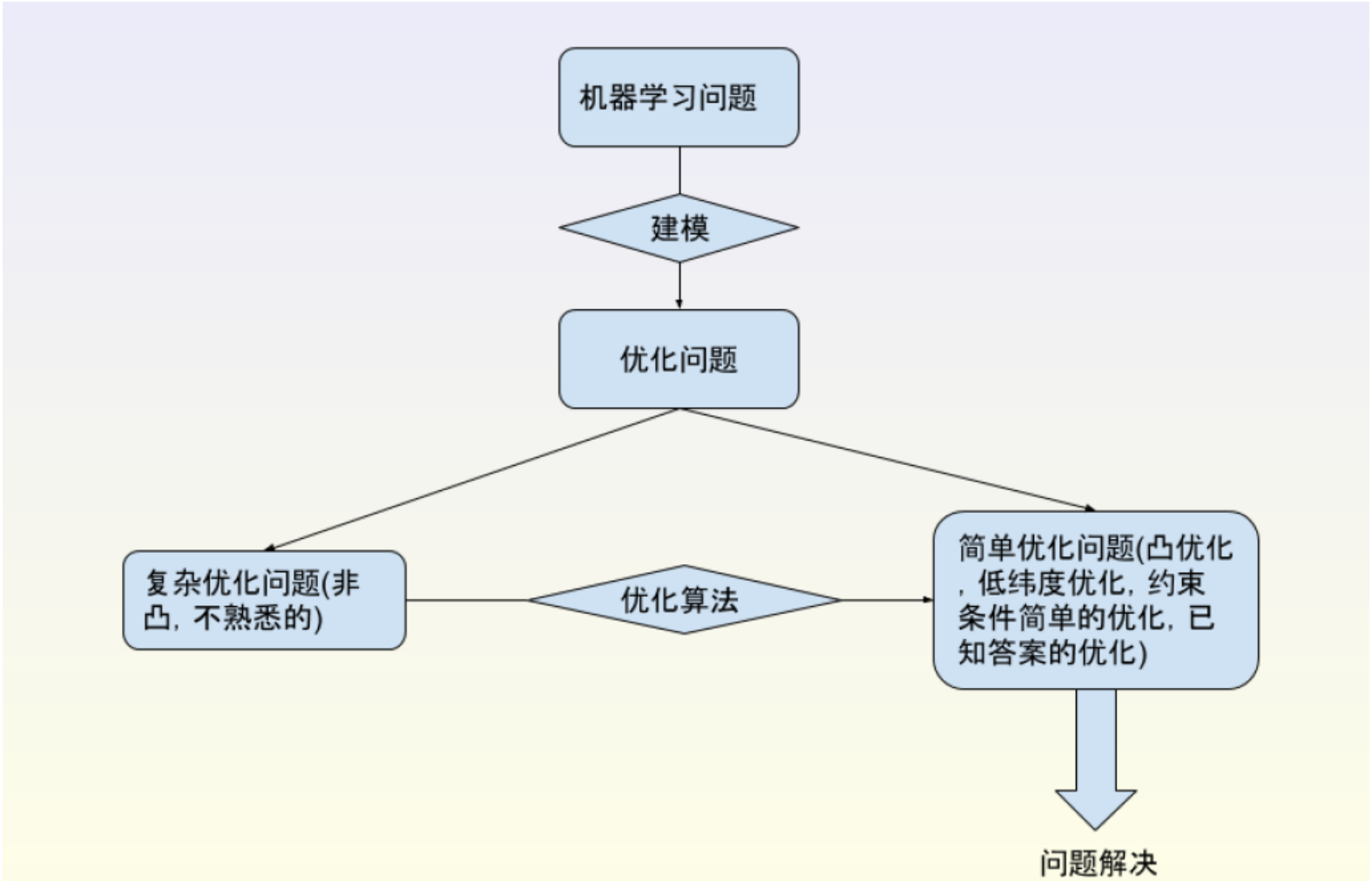
# 主要内容

1.优化问题简介

2.凸集合与凸函数
   a) 凸集合与凸函数的关系
   b) 琴生不等式的几何解释

3.优化与凸优化
   a) 凸优化问题
   b) 对偶问题
   c) 对偶性
   d) KKT条件
   e) 拉格朗日乘数法

# 1.优化问题简介


机器学习问题
建模
优化问题
复杂优化问题(非凸，不熟悉的)
优化算法
简单优化问题(凸优化，低纬度优化，约束条件简单的优化，已知答案的优化)
问题解决

# 1.优化问题简介

**优化问题**

优化问题的一般形式

$$\begin{aligned} \text{最小化：} & \quad f_0(x) \\ \text{条件：} & \quad f_i(x) \leq b_i, i = 1, \cdots, m. \end{aligned}$$

其中 $f_0(x)$ 为目标函数，条件里的不等式是限制条件.

# 1.优化问题简介

## 极大似然估计

如果 $L(\mu, \sigma)$ 是一个极大似然估计问题中的似然函数，其中 $\mu, \sigma$ 分别是期望和方差，那么极大似然估计的问题转化为

$$\text{最小化：} \quad -L(\mu, \sigma)$$
$$\text{条件：} \quad \sigma \geq 0$$

## 最小二乘

如果 $A_{n\times k}$ 是一个矩阵,$b \in \mathbb{R}^n$ 是一个向量, 对于 $x \in \mathbb{R}^k$

$$\text{最小化：} \quad f_0(x) = |Ax - b|^2$$

# 2.凸集合与凸函数

## 凸集合定义

如果一个集合 $\Omega$ 中任何两个点之间的线段上任何一个点还属于 $\Omega$, 那么 $\Omega$ 就是一个凸集合.i.e.

$$\lambda x_1 + (1-\lambda)x_2 \in \Omega, \forall x_1, x_2 \in \Omega, \lambda \in (0,1)$$

## 凸函数定义

如果一个函数 $f$ 定义域 $\Omega$ 是凸集，而且对于任何两点. 以及两点之间线段上任意一个点都有

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

$\forall x_1, x_2 \in \Omega, \lambda \in (0,1)$

# 2.凸集合与凸函数

**函数的上境图**

假设 $f$ 是一个定义在 $\Omega$ 上的函数，区域 $\{(x,y) : y \geq f(x), \forall x \in \Omega\}$ 就是 $f$ 的上境图.

上境图就是函数图像上方的部分区域.

**凸集合与凸函数**

一个函数是凸函数当且仅当 $f$ 的上境图是凸集合.

凸集合与凸函数有很多相对应的性质可以由这个结论来进行链接。

# 2.凸集合与凸函数

**凸组合**

对于任何 $n$ 个点 $\{x_i\}_{i=1}^n$，以及权重系数 $\{w_i\}_{i=1}^n$. 若权重系数非负 $w_i \leq 0$ 而且 $\sum_{i=1}^{n} w_i = 1$，则线性组合

$$S = \sum_{i=1}^{n} w_i x_i$$

为一个凸组合.

凸组合的物理意义可以理解成 $n$ 个重量为 $w_i$ 的点的整体重心.

# 2.凸集合与凸函数

**集合的凸包**

$n$ 个点 $\{x_i\}_{i=1}^n$ 的全部凸组合就构成 $\{x_i\}_{i=1}^n$ 的凸包.

**函数的凸闭包**

如果 $C$ 是函数 $f$ 的上境图，$\overline{C}$ 是 $C$ 的凸包，那么以 $\overline{C}$ 为上境图的函数称为 $f$ 的凸闭包.

# 2.凸集合与凸函数

## 凸集合性质

假设 $\Omega$ 是一个凸集合，那么 $\Omega$ 任何子集的凸包仍包含于 $\Omega$.

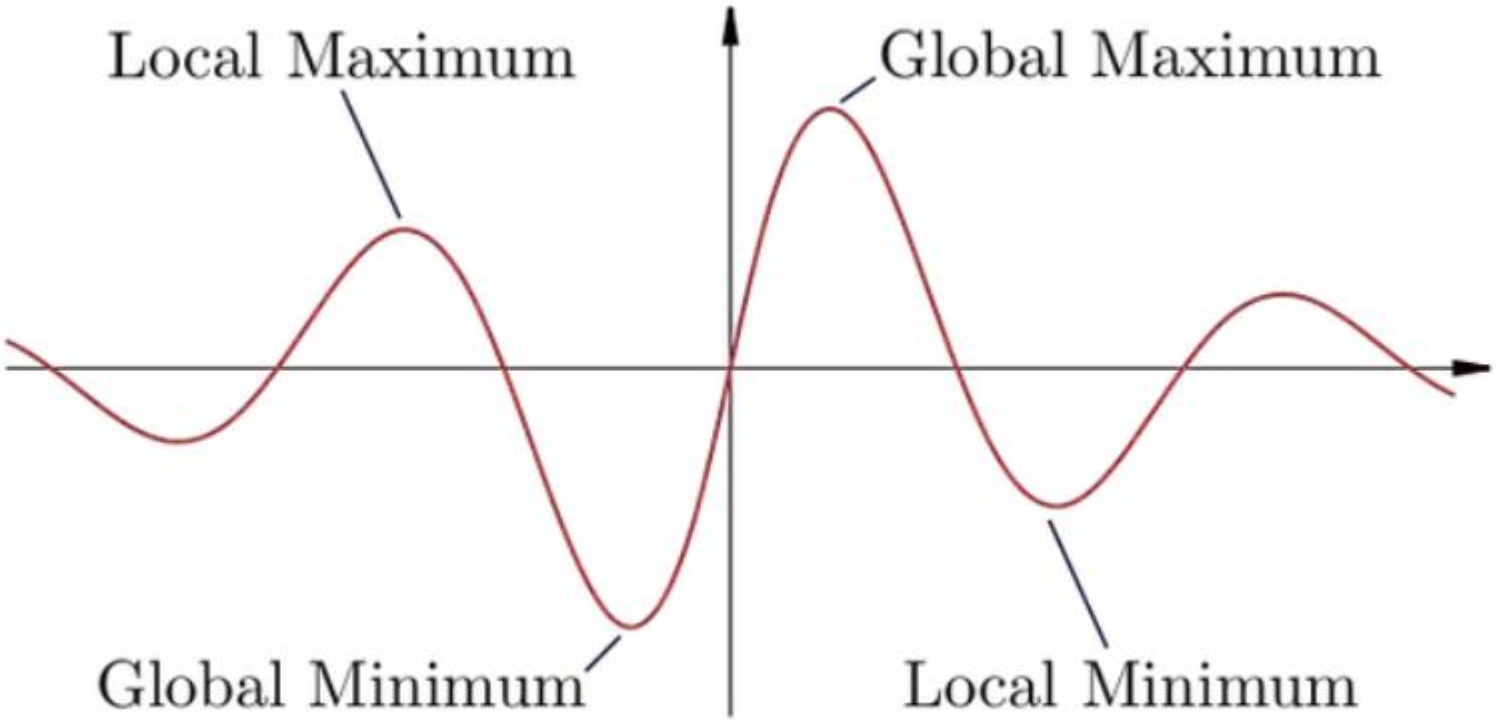
## 凸函数性质：琴生 (Jensen) 不等式

如果 $f:\Omega \to \mathbb{R}$ 是一个凸函数，则对于任何 $\{x_i \in \Omega\}_{i=1}^n$，以及凸组合 $\sum_{i=1}^n w_i x_i$ 都有

$$\sum_{i=1}^n w_i f(x_i) \geq f(\sum_{i=1}^n w_i x_i)$$

# 2.凸集合与凸函数

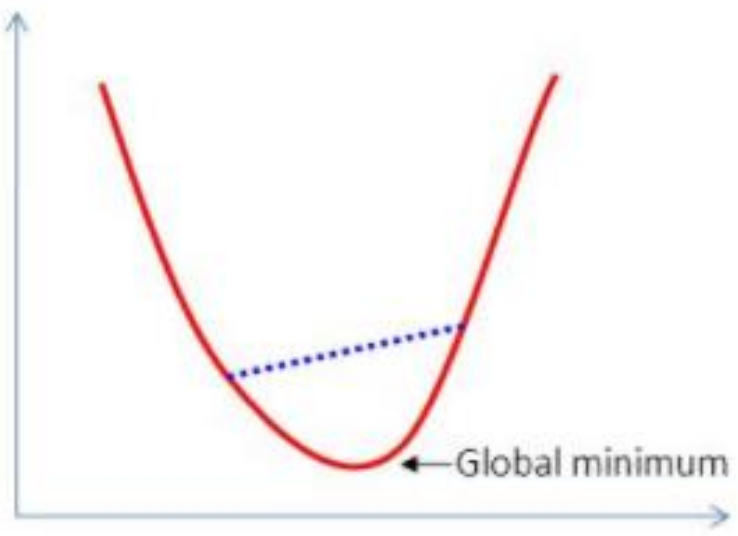
局部极值与全局极值



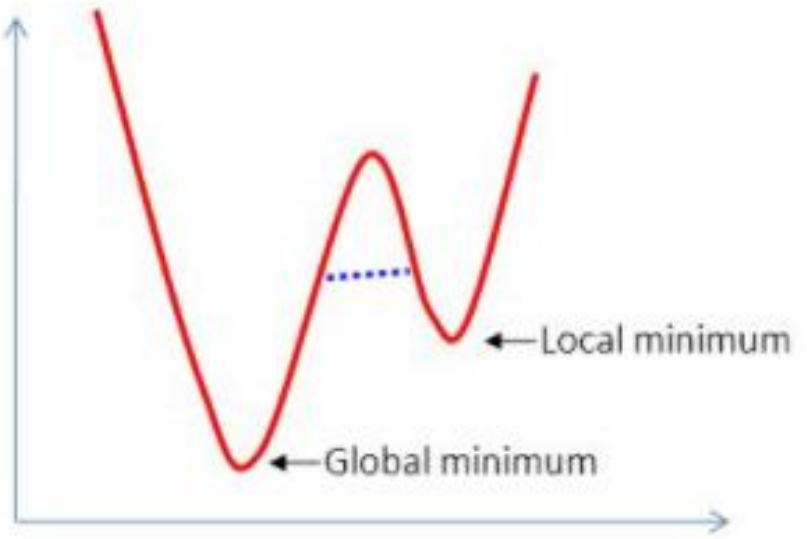
# 2.凸集合与凸函数

**凸函数的重要性质：**

- **局部极值一定是全局极值**




**凸函数**




**非凸函数**

# 3. 凸优化

## 凸优化问题

凸优化问题的一般形式

$$\begin{aligned} \text{最小化:} \quad & f_0(x) \\ \text{条件:} \quad & f_i(x) \leq b_i, i = 1, \cdots, m. \end{aligned}$$

其中 $f_0(x)$ 为目标函数，条件里的不等式是限制条件.

- 凸优化问题的条件,$f_0, f_1, \cdots, f_m$ 都是凸函数.
- 凸优化问题的特点, 局部最优等价于全局最优.
- 凸优化问题的求解, 几乎总有现成的工具来求解.

# 3 凸优化

## 凸优化的应用

- 凸优化问题逼近非凸优化问题，寻找非凸问题的初始点
- 利用对偶问题的凸性给原问题提供下界估计
- 凸优化问题可以给非凸问题带来一些启发

# 3. 凸优化：优化问题的对偶问题

## 优化问题

$$\text{最小化:} f_0(x)$$
$$\text{不等条件:} f_i(x) \leq b_i, i = 1, \cdots, m$$
$$\text{等式条件:} h_i(x) = 0, i = 1, \cdots, p$$
$$\text{定义域:} \mathcal{D} = \bigcap_{i=0}^{m} \mathrm{dom} f_i \cap \bigcap_{i=1}^{p} \mathrm{dom} h_i.$$

请注意定义域 $\mathcal{D}$ 指的是使得所有函数 $f_i, h_i$ 有定义的区域。而可行域指的是定义域中满足不等条件与等式条件的那些点. 本课中把这个优化问题称为原问题，优化点称为 $x^*$，最优化值为 $p^*$.

# 3. 凸优化：优化问题的对偶问题

根据原函数与限制条件我们定义拉格朗日量 $L(x,\lambda,\nu):\mathbb{R}^{n+m+p}\to\mathbb{R}$

**拉格朗日量**

$$L(x,\lambda,\nu)=f_0(x)+\sum_{i=1}^{m}\lambda_i f_i(x)+\sum_{i=1}^{p}\nu_i h_i(x),$$

根据拉格朗日函数我们定义拉格朗日对偶函数 $g(\lambda,\nu):\mathbb{R}^{m+p}\to\mathbb{R}$

**拉格朗日对偶函数**

$$\begin{aligned} g(\lambda,\nu) &= \inf_{x\in\mathcal{D}} L(x,\lambda,\nu) \\ &= \inf_{x\in\mathcal{D}} f_0(x)+\sum_{i=1}^{m}\lambda_i f_i(x)+\sum_{i=1}^{p}\nu_i h_i(x) \end{aligned}$$

# 3. 凸优化：优化问题的对偶问题

## 对偶函数为原问题提供下界

如果限制 $\lambda_i \geq 0, \forall i = 1, \cdots, m$, 则

$$g(\lambda, \nu) \leq p^*$$

## 证明

对任意一个 $x \in \mathcal{D}$, 如果 $x$ 在可行域中，那么

$$\begin{aligned} g(\lambda, \nu) &\leq f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \\ &= f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \\ &\leq f_0(x) \end{aligned}$$

# 3. 凸优化：优化问题的对偶问题

根据对偶函数, 定义对偶问题的一般形式

**对偶问题**

$$\text{最大化:} g(\lambda, \nu)$$
$$\text{不等条件:} \lambda_i \geq 0, i = 1, \cdots, m$$

我们把对偶问题的最大值点称为 $(\lambda^*, \nu^*)$, 相应的最大值称为 $d^*$, 这里面的对偶函数 $g$ 定义域为 $\mathrm{dom} g = \{(\lambda, \nu) : g(\lambda, \nu) > -\infty\}$. 在 $g$ 的定义域中满足 $\lambda_i \geq 0$ 的那些 $(\lambda, \nu)$ 全体, 叫做对偶可行域. 也就是对偶问题的可行域.

# 3. 凸优化：对偶性

根据对偶函数的性质我们已经知道在对偶可行域中，$g(\lambda, \nu)$ 总是不大于 $p^*$. 所以就有

**弱对偶性**

$$d* \leq p*$$

若对偶性总是对的. 相对而言的强对偶性是指一部分优化问题来说，有更好的结论.

**强对偶性**

$$d* = p*$$

强对偶性不总成立.

# 3. 凸优化：对偶性

**Slater 条件**

对于一个凸优化问题

$$\begin{aligned} &\text{最小化:} f_0(x) \\ &\text{不等条件:} f_i(x) \leq b_i, i = 1, \cdots, m \\ &\text{等式条件:} h_i(x) = 0, i = 1, \cdots, p \end{aligned}$$

如果存在一个可行域中的点 $x$ 使得 $f_i(x) < 0, i = 1, \cdots, m$, 那么这个凸优化问题就满足强对偶条件.

# 3. 凸优化：对偶性

## 满足强对偶性的例子

- 线性规划
- 最小二乘
- 最大熵问题

这种情况下我们如果发现对偶问题比原问题更容易解决，那么就可以使用对偶问题来解出 $d^* = p^*$

# 3. 凸优化：KKT条件

我们来看一下如果强对偶性满足的话，这些最优化点应该满足何种条件. 这一部分中我们假定所有的函数都是可微函数.
如果 $x^*, (\lambda^*, \nu^*)$ 分别是原问题与对偶问题的最优解，那么首先这些点应该满足可行域条件

- $f_i(x^*) \leq 0$
- $h_i(x^*) = 0$
- $\lambda_i^* \geq 0$

# 3. 凸优化：KKT条件

其次我们已经知道

$$
\begin{aligned}
d^* &= g(\lambda^*, \nu^*) \\
&\leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \\
&= f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) \\
&\leq f_0(x^*) \\
&= p^*
\end{aligned}
$$

于是 $d^* = p^*$ 意味着上述不等式全都是等式.

# 3. 凸优化：KKT条件

所以我们有

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$$

，以及

$$g(\lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*)$$

而因为

$$g(\lambda^*, \nu^*) = \inf(L(x^*, \lambda^*, \nu^*))$$

所以 $x^*$ 是拉格朗日函数在 $x$ 方向的驻点，所以有

$$\nabla_x L(x^*, \lambda^*, \nu^*) = 0$$

. 综上所述我们就的到了 KKT 条件.

# 3. 凸优化：KKT条件

**KKT 条件**

- $f_i(x^*) \leq 0, i = 1, \cdots, m$
- $h_i(x^*) = 0, i = 1, \cdots, p$
- $\lambda_i^* \geq 0, i = 1, \cdots, m$
- $\lambda_i^* f_i(x^*) = 0, i = 1, \cdots, m$
- $\nabla_x L(x^*, \lambda^*, \nu^*) = 0$

其中第四个条件是由 $\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$ 以及第一个和第三个条件共同得到的.

# 3. 凸优化：KKT条件

## KKT 条件使用

- 对于凸优化问题,KKT 条件是 $x^*, (\lambda^*, \nu^*)$ 分别作为原问题和对偶问题的最优解的充分必要条件.
- 对于非凸优化问题，KKT 条件仅仅是必要而非充分.

# 3. 凸优化：拉格朗日乘数法

当原问题只有等式约束而没有不等式约束时，KKT条件即为拉格朗日乘数法

例题：请在$x^2 + y^2 + z^2 = 1$的条件下最小化函数

$$x + 2y + 3z$$

# 总结

- **优化问题在机器学习的模型训练中有重要应用**
- **凸函数代数性质与凸集合的几何性质**
  - 琴生不等式的几何解释
- **凸优化是一类相对简单的优化问题**
  - 凸函数的局部最小值就是全局最小值
- **对偶方法的主要目的是处理原问题中的复杂边界条件**
  - 对偶问题永远是凸问题
  - 弱对偶性永远成立，可以为原问题提供下界
- **KKT条件可以用来求解一些优化问题**
  - 拉格朗日乘数法是KKT条件的一种特殊形式