数值优化（Numerical Optimization）学习系列-概述

下一步　于 2015-12-27 18:43:01 发布

开篇

1. 数值优化通过迭代的方式解决优化问题，是数学建模中关键的一环。
2. Modeling过程，需要确定优化目标、目标所依赖的变量以及变量之间的约束关系，最后通过优化算法解决问题。

基础

1. 对于一个优化问题，通常有一个优化目标函数 f(x) x为参数变量，c(x)为约束。
2. 最优化问题的标注形式为

$$
\begin{aligned}
min\ f(x) \quad x \in \mathbb{R}^n \\
s.t.\ C_i(x) = 0\ i \in \mathcal{E} \\
C_i(x) \geq 0\ i \in \mathcal{I}
\end{aligned}
$$

3. 其中 $\mathcal{E}$ 表示等式集合，$\mathcal{I}$ 表示不等式集合
4. 其中满足约束的解称之为 **可行解**

问题分类

根据目标函数或者约束函数的不同，对于最优化问题可以分为：

- 连续/离散优化问题
- 约束/非约束优化问题
- 线性/非线性优化问题
- 全局/局部优化问题
- 随机/确定性优化问题

了解分类规则后可以根据建模后的形式选择不同的算法。

凸优化

对于凸优化需要了解一下几个概念，详细可以参考Stephen Boyd的《凸优化》，里面对凸优化问题进行了详细的介绍。

1. 凸集：如果集合S为凸集，当且仅当 $x \in S, \ y \in S$ 并且$\alpha(x) + (1-\alpha)(y)\ inS; \quad \alpha \in [0,1]$
2. 凸函数：如果函数f(x)为凸函数，当且仅当S为凸集，$x \in S, \ y \in S; \ \alpha f(x) + (1-\alpha)f(y) \geq f(\alpha x + (1-\alpha)y); \ \alpha \in [0,1]$
3. 严格凸函数，凸函数能够取到非等号，即$\alpha \in (0,1)$
4. 凸优化问题：对于标准形式目标函数为凸函数，等式约束为线性约束；不等式约束为凹函数。

无约束最优化问题

在 机器学习 中，有大量的问题可以归约为无约束最优化问题，例如线性回归、LR等。因此对于无约束问题的研究也很深入从简单的GD、SGD、TR到CG、Newton、(L-)BFGS等

1. 无约束最优化问题可以表示为 $min f(x); \ x \in \mathbb{R}^n$

2. 全局最优解 VS 局部最优解

* 全局最优简单理解为在整个定义域内解最小

* 局部最优：在某个邻域内解最小

3. 对于凸优化问题，任何局部最优解都是全局最优解。

局部最优解几个定理

1. 泰勒展开公式，根据泰勒公式对于函数f(x)可以近似为
   一阶展开近似：$f(x) \approx f(x_0) + \nabla f(x_0)^T(x-x_0)$
   二阶展开近似：$f(x) \approx f(x_0) + \nabla f(x_0)^T(x-x_0) + \frac{1}{2}(x-x_0)^T\nabla^2 f(x_0)(x-x_0)$
2. 局部最小值的一阶必要条件，如果 x^* 为局部最优解并且函数f一阶可导，则在 x^* 的邻域内 $\nabla f(x^*) = 0$
3. 局部最优解的二阶必要条件，如果 x^*为局部最优解并且一阶和二阶可导，则$\nabla f(x^*) = 0$ **并且** $\nabla^2 f(x)$正定
   证明：对于定理2，3的证明采用反证法。例如对于定理2. 假设 $\nabla f(x^*) \neq 0$， 则根据泰勒一阶展开则可以找到$\nabla f(x^*)^T(x-x^*) \leq 0$
4. 局部最优的二阶充分条件：如果函数f在 x^* 处满足$\nabla f(x^*) = 0$并且$\nabla^2 f(x)$正定，则 x^* 为局部最优解
5. 如果函数f为凸函数，则f的任何局部最优解都为全局最优解。

优化算法概述

在后面会介绍一系列解决该问题的算法，先介绍几个简单的概念。

1. 通过数值优化算法求解，一般会给定初始点 x_0, 找到一系列点$x_1, x_2, x_3 \ ... \ x_n \rightarrow x^*$

2. 通常有两大类比较重要的策略 **线搜索（Line Search）和信赖域（Trust Region）**
3. Line Search策略：假设在某点x_k，寻找方向p_k和步长α使得 $min\ f(x_k+\alpha p_k)$ 最小，如果p_k确定则只需要解决一维最优化问题就可以找到下一个搜索点，如何确定p_k后面介绍几类策略，最简单的策略就是当前点的梯度。
4. Trust Region策略：在某店x_k解决优化问题 $min\ m_k(x_k+p_k)$，其中m_k为函数f在x_k点的近似，同时为保证m_k为一个较好的近似，即当p_k 远离 x_k 时偏离太大，则需要保证x_k+p_k 在一个可信赖的区域内；
5. 通常情况下信赖域选择为椭圆、球或者盒状区域，即一个凸集容易找到最优解。
6. 模型m_k一般可以选择为函数f的泰勒二阶近似，即

$$m(x_k+p)\approx f(x_k)+\nabla f(x_k)^T p+\frac{1}{2}p^T\nabla^2 B_k p$$

其中 B_k为Hessian矩阵或者其近似
7. 以上两类策略的相同点是在某点 x_k通过解决一个优化问题找到下一个搜索点。LS首先选择方向 p_k通过解决一维最优化问题找到步长 α；TR首先对步长进行约束，通过解决简单的优化问题寻找搜索方向。

线搜索中搜索方向选择

1. 最速下降方向，即搜索方向选择为,负梯度方向： $p_k=-\nabla f_k$。由泰勒展开公式$f(x_k+\alpha p_k)\approx f(x_k)+\alpha\nabla f(x_k)^T p_k+\frac{1}{2}p_k^T\nabla^2 f_k p_k$， 由于$\nabla^2 f_k$满足正定，因此只需要$\nabla f(x_k)^T p_k$最小。即$min f(x_k)^T p_k\ s.t||p||=1$；可以推出$p=-\nabla f_k/||\nabla f_k||$。主要问题对于复杂问题效率较慢
2. 通用搜索方向： 从泰勒展开公式上可以看到，只要满足 $\nabla f_k p_k\leq 0$ 都可以选择为搜索方向，问题是相比最速下降效率可能会较低。
3. 牛顿方向(Nowton direction,p_k^N)，$p_k^N=-(\nabla^2 f_k)^{-1}\nabla f_k$，解释如下

$$\text{泰勒公式}: m_k(p)=f(x_k+p)\approx f(x_k)+\nabla f(x_k)^T p+\frac{1}{2}p^T\nabla^2 f_k p$$
$$min\ m_k(p)$$
$$\Rightarrow \nabla m_k(p)=0$$
$$\Rightarrow \nabla f_k+\nabla f_k^2 p=0$$
$$\Rightarrow p_k^N=-(\nabla^2 f_k)^{-1}\nabla f_k$$

关于 p_k^N 1) 当 $\nabla^2 f_k$正定时满足$p^T\nabla f_k=-p^T\nabla^2 f_k p\leq 0$ 满足函数值下降，为有效搜索方向。2) 当 $\nabla^2 f_k$非正定时，$-(\nabla^2 f_k)^{-1}$不一定存在，即使存在也不一定满足下降条件。

4. 伪牛顿方向(Quasi-Newton 方向)，$p_k=-B_k^{-1}\nabla f_k$，由于Hessian矩阵计算复杂度较高而且不一定能够满足正定，可进行近似。

$$\text{泰勒公式：}\nabla f(x_k+p)\approx \nabla f(x_k)+\nabla^2 f(x_k)p$$
$$\text{由于}x_{k+1}=x_k+p\text{，令}$$
$$s_k=x_{k+1}-x_k$$
$$y_k=\nabla f_{k+1}-\nabla f_k$$
$$\Rightarrow y_k=B_{k+1}s_k(\text{伪牛顿条件})$$

另外一种理解方式

$$\text{将}f(x)\text{在点}x_{k+1}\text{处进行泰勒展开}$$
$$f(x) \approx f(x_{k+1}) + \nabla f(x_{k+1})^T(x - x_{k+1})$$
$$+ \frac{1}{2}(x - x_{k+1})^T \nabla^2 f_{k+1}(x - x_{k+1})$$
$$\text{在}x = x_k\text{的梯度为}$$
$$\nabla f_k \approx \nabla f_{k+1} + \nabla^2 f_{k+1}(x_k - x_{k+1})$$
$$y_k = B_{k+1}s_k$$

在实际使用中一般还对B_{k+1}添加一些附加条件，例如对称、正定以及低秩等，两个比较常用的近似算法为*SR1* 和*BFGS*

5. 非线性共轭梯度方向：$p_k = -\nabla f_k + \beta_k p_{k-1}$，后面会详细介绍该算法。

信赖域模型

1. 对于LS中的共轭方向外，其他方向的模型均可以引入到TR中
2. 例如，牛顿方向$m_k(x_k + p)$ 中将$B_k = 0$ 对应于TR模型中

$$min f_k + \nabla_k^T p \quad s.t||p||_2 \leq \Delta_k$$
$$\Rightarrow p_k = -\frac{\Delta_k \nabla f_k}{||\nabla f_k||}$$

SCALING 问题

1. 一个poor scaled 问题是指函数f(x)在某个方向上的变化比其他方向的变化，带来更大的函数值改动。即某个方向的微小改动带来巨大函数响应，例如 $f(x) = 10^9 x_1^2 + x_2^2$ 对x1方向的变化比较敏感。
2. 可以通过变量重定义的方式解决问题。
3. 线搜索问题中的最速下降法是poor scaled算法，Newton算法也会受到影响。最速下降法对于条件数比较大的问题会带来之字迭代，收敛速度大幅下降

总结

几个重要的知识点
1. 优化问题的标准形式（后续的学习中以此为准）
2. 凸优化问题：凸集、凸函数
3. 全局最优解 VS 局部最优解
4. 局部最优解的一阶、二阶必要条件，可证明

5. 线搜索常用搜索方向；信赖域常用模型
6. poor scaled问题