Lab 03. Optimization

Today

1. 回顾梯度下降法、牛顿法、高斯牛顿法和列文伯格-马夸特法的发展脉络。
2. 介绍作业任务。

🤔 资料存在疑惑或者不明白的可以咨询→计算机视觉导论助教（沈泽弘），2024/10/10

1. 梯度下降法

利用函数的一阶导数来求解最优化问题。

多元函数求导实际上是求偏导，表示每个自变量方向上的变化率。让自变量沿着损失函数梯度的反方向变化，就是梯度下降法。

如果损失函数是凸函数，则梯度下降法得到的解一定是全局最优解。否则优化的结果和初始点相关，可能会得到局部最优解。

数学形式

- 假设一阶连续可导函数$f(x) \in \mathbb{R}$，定义域（参数）$x \in \mathbb{R}^n$。求解$\min\limits_{x \in \mathbb{R}^n} f(t; x)$。$t$是样本数据。
- Gradient Descent（GD）通过不断迭代 x ，达到减少函数值的目的。第 $k+1$ 次迭代值

  $$x^{(k+1)} \leftarrow x^{(k)} + \lambda_k p_k$$

  其中 p_k是搜索方向，在这里取梯度的负方向。 λ_k是搜索步长，是通过线性搜索，得到的能使得目标最小的步长。

- 当梯度不再变化时终止迭代。

变种

样本可以取“所有”、“部分”、“单个”。现在一般SGD（Stochastic-GD）指Mini-batch GD。

2. 牛顿法

相比较GD（一阶导数），进一步利用函数的二阶导数（Hessian矩阵）。

数学推导

假设参数 $x \in \mathbb{R}^n$，无约束优化目标二阶连续可导： $x* = \arg\min_x f(x)$

对多元函数 $f(x)$ 泰勒展开。

$$f\left(x\right) = f\left(x^{(k)}\right) + g_k^T\left(x - x^{(k)}\right) + \frac{1}{2}\left(x - x^{(k)}\right)^T H\left(x^{(k)}\right)\left(x - x^{(x)}\right)$$

> 💡 求解函数极值的常用套路：一阶、二阶泰勒展开（近似）后，再对等式求导。因为极值的必要条件是一阶导数为0。

对在 $x^{(k)}$点处的近似泰勒展开求导（用nabla算子 ∇）：

$$\nabla f(x) = g_k + H_k(x - x^{(k)})$$

此时，这里记 $H_k = H\left(x^{(k)}\right)$。因为Hessian矩阵内均是数值，所以求解此近似条件下的下一个 x 点：

$$g_k + H_k\left(x^{(k+1)} - x^{(k)}\right) = 0$$
$$x^{(k+1)} = x^{(k)} - H_k^{-1} g_k$$

优缺点

缺点：计算海森矩阵计算量大（二阶导数矩阵），不一定可逆。

变种：拟牛顿法

用一个矩阵代替海森矩阵的逆。经典的包括DFP(Davidon-Fletcher-Powell)、BFGS(Broyden-Fletcher-Goldfard-Shano)、Broyden等算法。

3. 高斯牛顿法

考虑牛顿法的特殊情况：目标函数是误差的二次函数。

数学推导

在 x_0展开误差函数 $\frac{1}{2}\|f(x)\|^2$

$$\begin{aligned}\frac{1}{2}\|f(x)\|^2 &= \frac{1}{2}(f(x_0)+J(x_0)^T(x-x_0))^2 \\ &= \frac{1}{2}(f^2(x_0)+2f(x_0)J(x_0)^T(x-x_0)+(x-x_0)^TJ(x_0)J(x_0)^T(x-x_0))\end{aligned}$$

▼ **Jacobian定义**

具体来说，Jacobian 矩阵是向量值函数的所有一阶偏导数组成的矩阵。假设有一个从 $\mathbb{R}^n$ 到 $\mathbb{R}^m$ 的向量值函数 $\mathbf{f}(\mathbf{x})=(f_1(\mathbf{x}),f_2(\mathbf{x}),\ldots,f_m(\mathbf{x}))$，其中 $\mathbf{x}=(x_1,x_2,\ldots,x_n)$ 是输入变量的向量，那么 Jacobian 矩阵 $J(\mathbf{x})$ 的定义为：

它是一个 $m\times n$ 的矩阵，表示的是向量函数在不同变量上的一阶偏导数。

- 如果 $m=1$（即标量值函数），那么 Jacobian 矩阵退化为梯度向量。
- 如果 $n=1$（即标量输入），那么 Jacobian 矩阵变成导数的列向量。

套路：求导为0得到

$$x_1=x_0-\frac{J^Tf_0}{JJ^T}$$

类比牛顿法，此时 $H=JJ^T$， $g=J^Tf$

优缺点

- 通过求偏导 J（一阶导数），避免了求海森矩阵（二阶导数），大大减少了计算量。
- 海森矩阵在凸函数时半正定的（在零特征值时不可逆），在非凸函数时存在负定或者不定的现象（不一定可逆）； JJ^T是半正定（目标函数是凸函数，在零特征值时不可逆），性质也不是特别好。因此下面引入了列文伯格-马夸特法。

4. 列文伯格-马夸特法

思想

高斯牛顿法中，计算步长时矩阵性质如果不好，容易导致锯齿形，因此引入“信赖区域”。制定一个评判指标来动态调整步长。引入阻尼量： $H\to H+\mu I$。

大致的思想是：当函数估计性质很好时，用高斯牛顿，i.e. 阻尼为0。当函数估计性质不好时，用最速下降，i.e. 增大阻尼因子。当然，也可以理解为，当阻尼很大的时候步长变小。

作业

- 上交完成好的Notebook导出的PDF

- 任务：提供753个数据点 (x, y, z)，已知这些数据点表示一个椭球。使用高斯牛顿法求解椭球参数。（下图为用scipy求解的结果，详情见notebook）