8.3 二阶方法 8.4 随机梯度下降

纪一仑

Harbin Institute of Technology, Shen Zhen

Group Meeting, January 2023

目录

- 8.3 二阶方法
 - 8.3.1 牛顿法
 - 8.3.2 BFGS 和其他类牛顿法
 - 8.3.3 Trust region method
- 2 8.4 随机梯度下降
 - 8.4.1 有限和问题的应用
 - 8.4.2 实例: SGD 用于线性回归

纪一仑 (HITSZ) 优化算法 GM 2023 2/19

目录

- 8.3 二阶方法
 - 8.3.1 牛顿法
 - 8.3.2 BFGS 和其他类牛顿法
 - 8.3.3 Trust region method
- 2 8.4 随机梯度下降
 - 8.4.1 有限和问题的应用
 - 8.4.2 实例: SGD 用于线性回归

纪一仑 (HITSZ)

8.3.1 牛顿法

牛顿法

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{H}_t^{-1} \boldsymbol{g}_t$$

where

$$\mathbf{H}_{t} \triangleq \nabla^{2} \mathcal{L}(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}_{t}} = \nabla^{2} \mathcal{L}\left(\boldsymbol{\theta}_{t}\right) = \mathbf{H}\left(\boldsymbol{\theta}_{t}\right)$$

纪一仑 (HITSZ) 优化算法 GM 2023

8.3.1 牛顿法

推导

考虑用二阶泰勒级数近似在 θ_t 周围的 $\mathcal{L}(\theta)$

$$\mathcal{L}_{quad}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}_t) + \boldsymbol{g}_t^{\top}(\boldsymbol{\theta} - \boldsymbol{\theta}_t) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_t)^{\top}\mathbf{H}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t)$$

最小值点是 $\boldsymbol{\theta} = \boldsymbol{\theta}_t - \mathbf{H}_t^{-1} \boldsymbol{g}_t$,需要 \mathbf{H}_t 正定

线性回归应用

$$RSS(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (\mathbf{y}_n - \mathbf{w}^{\top} \mathbf{x}_n)^2 = \frac{1}{2} ||\mathbf{X} \mathbf{w} - \mathbf{y}||_2^2 = \frac{1}{2} (\mathbf{X} \mathbf{w} - \mathbf{y})^{\top} (\mathbf{X} \mathbf{w} - \mathbf{y})$$
$$= \frac{1}{2} \mathbf{y}^{\top} \mathbf{y} - \mathbf{y}^{\top} \mathbf{X} \mathbf{w} + \frac{1}{2} \mathbf{w}^{\top} \mathbf{X}^{\top} \mathbf{X} \mathbf{w}$$

损失函数就是二阶泰勒级数,使用牛顿法可以直接得到结果

5/19

计算 Hessian 矩阵 \mathbf{H}_t 的代价比较大,为了减小这个代价,类牛顿法使用 Pseudo Hessian 矩阵 \mathbf{B}_t 来近似 \mathbf{H}_t 。

Pseudo Hessian 的要求

- 1. **B**_t 正定对称
- 2. \mathbf{B}_{t+1} 由之前迭代步骤的梯度/更新方向得到
- \bullet 3. \mathbf{B}_{t+1} 需要逼近 \mathbf{B}_t ,来保证收敛
- 4. \mathbf{B}_{t+1} 的更新开销需要尽可能小
- 5. $\mathbf{B}_{t+1}(\boldsymbol{\theta}_t \boldsymbol{\theta}_{t-1}) = \boldsymbol{g}_t \boldsymbol{g}_{t-1}$

BFGS 公式的推导

求如下的优化问题

$$\min_{B_+} \|B_+ - B\|_{\textit{F}}$$

s. t.
$$\mathbf{B}_+ \boldsymbol{\delta} = \boldsymbol{\gamma}$$
 and $\mathbf{B}_+ = \mathbf{B}_+^{\top}$ and $\mathbf{B}_+ \succ 0$

参考链接: https://zhuanlan.zhihu.com/p/573703008

BFGS 公式

$$egin{aligned} \mathbf{B}_{t+1} &= \mathbf{B}_t + rac{oldsymbol{y}_t oldsymbol{y}_t^ op}{oldsymbol{y}_t^ op oldsymbol{s}_t} - rac{\left(\mathbf{B}_t oldsymbol{s}_t
ight) \left(\mathbf{B}_t oldsymbol{s}_t
ight)^ op}{oldsymbol{s}_t^ op \mathbf{B}_t oldsymbol{s}_t} \ oldsymbol{s}_t &= oldsymbol{ heta}_t - oldsymbol{ heta}_{t-1} \ oldsymbol{y}_t &= oldsymbol{g}_t - oldsymbol{g}_{t-1} \end{aligned}$$

7/19

一仑 (HITSZ) 优化算法 GM 2023

 \mathbf{B}_t 保持正定需要满足下面的条件

Wolfe Conditions

$$\bullet \ \mathcal{L}(\boldsymbol{\theta}_t + \eta \boldsymbol{d}_t) \leq \mathcal{L}(\boldsymbol{\theta}_t) + c \eta \boldsymbol{d}_t^\top \nabla \mathcal{L}(\boldsymbol{\theta}_t)$$

$$\bullet \ -\boldsymbol{d}_t^\top \nabla \mathcal{L}(\boldsymbol{\theta}_t + \eta \boldsymbol{d}_t) \leq -c_2 \boldsymbol{d}_t^\top \nabla \mathcal{L}(\boldsymbol{\theta}_t)$$

参考链接: https://www.zhihu.com/question/49600881

其中 $0 < c < c_2 < 1$

第一个条件确保函数单调下降。

第二个条件确保函数的梯度也下降。

4□ > 4□ > 4 = > 4 = > = 90

纪一仑 (HITSZ) 优化算法

8/19

其他需要注意的事情

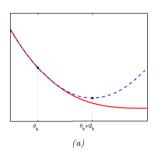
• 可以使用 SWM 公式直接维护 \mathbf{B}_{t+1} 的逆 \mathbf{C}_{t+1}

$$\mathbf{C}_{t+1} = \left(\mathbf{I} - \frac{\boldsymbol{s}_t \boldsymbol{y}_t^\top}{\boldsymbol{y}_t^\top \boldsymbol{s}_t}\right) \mathbf{C}_t \left(\mathbf{I} - \frac{\boldsymbol{y}_t \boldsymbol{s}_t^\top}{\boldsymbol{y}_t^\top \boldsymbol{s}_t}\right) + \frac{\boldsymbol{s}_t \boldsymbol{s}_t^\top}{\boldsymbol{y}_t^\top \boldsymbol{s}_t}$$

• 存储 \mathbf{B}_t 需要 $O(D^2)$ 的空间,对于很大的问题,可以只存储最近的 M 对 $(\mathbf{s}_t, \mathbf{y}_t)$,然后用这些 $(\mathbf{s}_t, \mathbf{y}_t)$ 直接计算 \mathbf{B}_t ,这样只需要 O(MD) 的空间。这就是 limited memory BFMS,即 L-BFMS。

纪一仑 (HITSZ) 优化算法 GM 2023

牛顿法得到的是多元二次函数的驻点,所以当 \mathbf{H}_t 不正定时, $\mathbf{d}_t = -\mathbf{H}_t^{-1}\mathbf{g}_t$ 可能不是函数下降的方向。



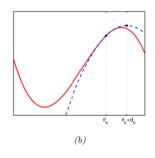


Figure: 牛顿法不一定向着函数下降的方向

纪一仑 (HITSZ) 优化算法 GM 2023 10/19

但是,这个多元二次函数在一定区域内肯定有最小值。

Trust-region optimization

$$\boldsymbol{\delta}^* = \operatorname*{arg\,min}_{\boldsymbol{\delta} \in \mathcal{R}_t} M_t(\boldsymbol{\delta})$$

其中 $\delta = \theta - \theta_t$, \mathcal{R}_t 是所取区域, $M(\delta)$ 是目标函数或者它的估计。

Tikhonov damping

通常情况下, 假设

$$M_t(\boldsymbol{\delta}) = \mathcal{L}_{quad}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}_t) + \boldsymbol{g}_t^{ op} \boldsymbol{\delta} + \frac{1}{2} \boldsymbol{\delta}^{ op} \mathbf{H}_t \boldsymbol{\delta}$$

$$\mathcal{R}_t = \{ \boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \le r \}$$

使用拉格朗日乘数法将有约束问题转化为无约束问题

$$\boldsymbol{\delta}^* = \operatorname*{arg\,min}_{\boldsymbol{\delta}} \boldsymbol{\mathit{M}}(\boldsymbol{\delta}) + \lambda \|\boldsymbol{\delta}\|_2^2 = \operatorname*{arg\,min}_{\boldsymbol{\delta}} \boldsymbol{\mathit{g}}^\top \boldsymbol{\delta} + \frac{1}{2} \boldsymbol{\delta}^\top (\mathbf{H} + \lambda \mathbf{I}) \boldsymbol{\delta}$$

当 λ 足够大,这个函数的 Hessian,即 $\mathbf{H} + \lambda \mathbf{I}$ 肯定是正定的,于是有最小值点

$$\boldsymbol{\delta} = -(\mathbf{H} + \lambda \mathbf{I})^{-1} \boldsymbol{g}$$

- 4 ロ ト 4 個 ト 4 恵 ト 4 恵 ト - 恵 - 夕 Q C

12 / 19

纪一仓 (HITSZ) 优化算法 GM 2023

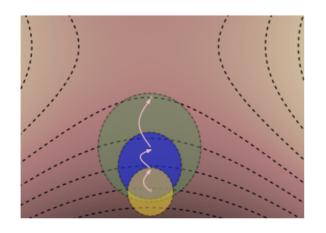


Figure: trust region method

目录

- 1 8.3 二阶方法
 - 8.3.1 牛顿法
 - 8.3.2 BFGS 和其他类牛顿法
 - 8.3.3 Trust region method
- 2 8.4 随机梯度下降
 - 8.4.1 有限和问题的应用
 - 8.4.2 实例: SGD 用于线性回归

纪一仑 (HITSZ) 优化算法 GM 2023 14/19

8.4 随机梯度下降

目标函数:

$$\mathcal{L}(oldsymbol{ heta}) = \mathbb{E}_{oldsymbol{q(z)}}[\mathcal{L}(oldsymbol{ heta}, oldsymbol{z})]$$

这里 \mathbf{z} 是随机输入,且 $\mathbf{z} \sim \mathbf{q}$ 。 在每次迭代中,假设我们观测到 $\mathcal{L}_t(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}, \mathbf{z}_t)$,如果分布 $\mathbf{q}(\mathbf{z})$ 独立于我们优化的参数 $\boldsymbol{\theta}$,我们可以使用 $\mathbf{g}_t = \nabla_{\boldsymbol{\theta}} \mathcal{L}_t(\boldsymbol{\theta}_t)$ 来近似 $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_t)$

于是, 迭代公式为:

$$oldsymbol{ heta}_{t+1} = oldsymbol{ heta}_t - \eta_t
abla \mathcal{L}(oldsymbol{ heta}_t, oldsymbol{z}_t) = oldsymbol{ heta}_t - \eta_t oldsymbol{g}_t$$

8.4.1 有限和问题的应用

经验风险最小化的损失函数:

$$\mathcal{L}\left(\boldsymbol{\theta}_{t}\right) = \frac{1}{N} \sum_{n=1}^{N} \ell\left(\boldsymbol{y}_{n}, f(\boldsymbol{x}_{n}; \boldsymbol{\theta}_{t})\right) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{n}\left(\boldsymbol{\theta}_{t}\right)$$

此损失函数的梯度:

$$\boldsymbol{g}_{t} = \frac{1}{N} \sum_{n=1}^{N} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{n}(\boldsymbol{\theta}_{t}) = \frac{1}{N} \sum_{n=1}^{N} \nabla_{\boldsymbol{\theta}} \ell\left(\boldsymbol{y}_{n}, f(\boldsymbol{x}_{n}; \boldsymbol{\theta}_{t})\right)$$

可以看到, 想要精确计算出梯度, 需要在所有 N 个样本点上计算。

纪一仑 (HITSZ) 优化算法 GM 2023 16/

8.4.1 有限和问题的应用

当 N 很大的时候,计算梯度会很慢。幸运的是,我们可以用少量样本上的梯度估计所有样本上的梯度。

在所有 N 个样本中随机取 $B \ll N$ 个样本,作为一个 minibatch,计算 minibatch 上的梯度:

$$\mathbf{g}_{t} pprox \frac{1}{|\mathcal{B}_{t}|} \sum_{n \in \mathcal{B}_{t}} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{n}(\boldsymbol{\theta}_{t}) = \frac{1}{|\mathcal{B}_{t}|} \sum_{n \in \mathcal{B}_{t}} \nabla_{\boldsymbol{\theta}} \ell\left(\mathbf{y}_{n}, f(\mathbf{x}_{n}; \boldsymbol{\theta}_{t})\right)$$

因为 minibatch 是随机取的,所以这个是无偏估计。

8.4.2 实例: SGD 用于线性回归

线性回归的损失函数:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2N} \sum_{n=1}^{N} \left(\mathbf{x}_{n}^{\top} \boldsymbol{\theta} - y_{n} \right)^{2} = \frac{1}{2N} \| \mathbf{X} \boldsymbol{\theta} - \mathbf{y} \|_{2}^{2}$$

梯度:

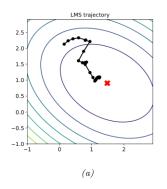
$$\mathbf{g}_t = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{\theta}_t^{\top} \mathbf{x}_n - y_n) \mathbf{x}_n$$

使用 SGD, minibatch 的大小取 B=1, 迭代公式为:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t (\boldsymbol{\theta}_t^{\mathsf{T}} \boldsymbol{x}_n - \boldsymbol{y}_n) \boldsymbol{x}_n$$

纪一仑 (HITSZ) 优化算法 GM 2023 18 / 19

8.4.2 实例: SGD 用于线性回归



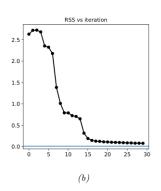


Figure: SGD for fitting linear regression