牛顿类算法

芦星宇

April 2021

目录

1 牛顿-共轭梯度法

1.1 牛顿法

牛顿法是梯度下降的进一步发展，梯度下降法利用了目标函数的一阶导数信息，以负梯度方向为搜索方向，只考虑目标函数在迭代点的局部性质，而牛顿法为代表的二阶近似方法除了一阶导数信息外，还利用二阶导数信息掌握了梯度变化的趋势，因而能够确定更合适的搜索方向加快收敛，具有二阶收敛速度。

1. 牛顿法对目标函数有比较高的要求，必须一阶二阶可导，海森矩阵必须**正定**
2. 计算量比较大，除了计算梯度外，还要计算海森矩阵及其逆矩阵
3. 当目标函数不是完全的凸函数时，容易陷入**鞍点**，导致更新朝着错误的方向移动。

1.2 共轭梯度法

共轭梯度法介于梯度下降法和牛顿法之间。共轭梯度法把共轭性与最速下降法相结合，利用迭代点处的梯度构造一组共轭方向，并沿共轭方向进行搜索，当前方向上的极小值搜索不会影响已经搜索过的方向的极值，因此共轭梯度法就有二次终止性。

什么是共轭?

H 是海森矩阵，如果满足 $d_t^T H d_{t-1} = 0$，则两个方向 d_t 和 d_{t-1} 共轭。

在应用共轭梯度法的时候，我们寻求一个和先前性搜索方向共轭的搜索方向，即不会撤销该方向上的进展，在训练迭代 t 时，下一步的搜索方向 d_t 的形式如下：

$$d_t = \bigtriangledown_\theta J(\theta) + \beta_t d_{t-1}$$

其中，β_t 控制我们应沿 d_{t-1} 加回多少到当前搜索方向上。适应共轭的直接方法会涉及到 H 特征向量的计算来选择 β_t，计算量是非常大的。为了避免这些计算，产生了以 FR 为代表的两种流行的算法来计算 β_t。

神经网络及其他深度学习模型的目标函数比二次函数复杂得多，但经验上，共轭梯度法在这种情况下仍然适用。当目标函数是高于二次的连续函数 (即目标函数的梯度存在) 时，其对应的解方程是 ** 非线性 ** 方程，非线性问题的目标函数可能存在局部极值，并且破坏了二次截止性，共轭梯度法需要在两个方面加以改进后，仍然可以用于实际的反演计算，但共轭梯度法不能确保收敛到全局极值。

1) 首先是共轭梯度法不能在 n 维空间内依靠 n 步搜索到达极值点，需要重启共轭梯度法，继续迭代，以完成搜索极值点的工作。

2) 在目标函数复杂，在计算时，由于需要局部线性化，需计算 Hessian 矩阵，且计算工作量比较大，矩阵 A 也有可能是病态的。Fletcher-Reeves 算法最为常用，抛弃了矩阵的计算。

共轭梯度法仅需一阶导数信息，但克服了最速下降法收敛慢的缺点，又避免了牛顿法需要存储和计算 Hesse 矩阵并求逆的缺点，共轭梯度法不仅是解决大型线性方程组最有用的方法之一，也是解大型非线性最优化最有效的算法之一。

对于规模较大的问题，精确求解牛顿方程组的代价比较高。事实上，牛顿方程求解等于无约束二次优化问题：

$$\min_{d^k} \frac{1}{2}(d^k)^\top \nabla^2 f(x^k) d^k + (\nabla f(x^k))^\top d^k,$$

其可以通过共轭梯度法来进行求解。

2 逻辑回归问题

如果面试官问你熟悉哪个机器学习模型，可以说 SVM，但千万别说 LR，因为细节真的太多。

Logistic Regression 虽然被成为回归, 但其实际上是分类模型, 并常用于二分类。Logistic Regression 因其简单、可并行化，可解释性强深受工业界喜爱。

本质：假设数据服从这个分布，然后使用极大似然估计做参数的估计。

2.1 Logistic 分布

Logistic 分布是一种连续型的概率分布，其 ** 分布函数 ** 和 ** 密度函数 ** 分别为：

$$F(x) = P(X \le x) = \frac{1}{1+e^{-(x-\mu)/\gamma}} f(x) = F'(X \le x) = \frac{e^{-(x-\mu)/\gamma}}{\gamma(1+e^{-(x-\mu)/\gamma})^2}$$

其中 μ 表示位置参数, $\gamma > 0$ 为形状参数。

Logistic 分布是由其位置和尺度参数定义的连续分布。Logistic 分布的形状与正态分布的形状相似, 但是 Logistic 分布的尾部更长, 所以我们可以使用 Logistic 分布来建模比正态分布具有更长尾部和更高波峰的数据分布。在深度学习中常用到的 Sigmoid 函数就是 Logistic 的分布函数在 $\mu = 0, \gamma = 1$ 的特殊形式。

2.2 Logistic 回归

以二分类为例，对于所给数据集假设存在一条直线可以将数据完成线性可分。

有时候我们只要得到一个类别的概率，那么我们需要一种能输出 [0, 1] 区间的值的函数，考虑二分类模型，我们利用判别模型，希望对 $p(C|x)$ 建模，利用贝叶斯定理：

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_2)p(C_2)}$$

取 $a=\ln\frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}$，于是：

$$p(C_1|x)=\frac{1}{1+\exp(-a)}$$

上面的式子叫 Logistic Sigmoid 函数，其参数表示了两类联合概率比值的对数。在判别式中，不关心这个参数的具体值，模型假设直接对 a 进行。

Logistic 回归的模型假设是：

$$a=w^Tx$$

于是，通过寻找 w 的最佳值可以得到在这个模型假设下的最佳模型。概率判别模型常用最大似然估计的方式来确定参数。

对于一次观测，获得分类 y 的概率为（假定 $C_1=1,C_2=0$）：

$$p(y|x)=p_1^yp_0^{1-y}$$

那么对于 N 次独立全同的观测 MLE 为：

$$\hat{w}=\underset{w}{argmax}\,J(w)=argmax*w\sum*i=1^N(y_i\log p_1+(1-y_i)\log p_0)$$

注意到，这个表达式是交叉熵表达式的相反数乘 N，MLE 中的对数也保证了可以和指数函数相匹配，从而在大的区间汇总获取稳定的梯度。

对这个函数求导数，注意到：

$$p_1'=(\frac{1}{1+\exp(-a)})'=p_1(1-p_1)$$

则：

$$J'(w)=\sum_{i=1}^Ny_i(1-p_1)x_i-p_1x_i+y_ip_1x_i=\sum_{i=1}^N(y_i-p_1)x_i$$

由于概率值的非线性，放在求和符号中时，这个式子无法直接求解。于是在实际训练的时候，和感知机类似，也可以使用不同大小的批量随机梯度上升（对于最小化就是梯度下降）来获得这个函数的极大值。

3 总结

牛顿法使用了二阶梯度，梯度下降仅仅是一阶梯度，对于梯度下降而言，把它比做下山问题，那么梯度下降站在当前的位置要找到梯度最大的点，这样也就是坡度最大下山最快，但是牛顿法他不仅要找当前下降最快的方向，还要确保下下步的坡度更大，下山更快。

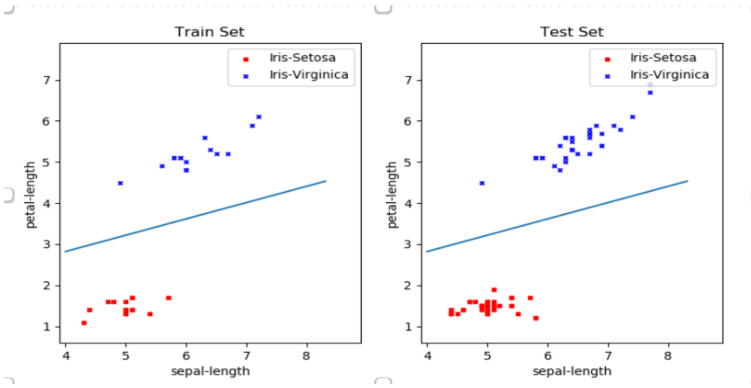

图 1: 牛顿法解决逻辑回归问题 (Iris 数据集)

参考文献

[1]. 拟牛顿法、共轭梯度法 - 知乎

[2]. 逻辑回归问题

[3]. 【机器学习】【白板推导系列】shuhuai008

[4]. 《Deep Learning》