梯度下降法调研结果

2313652郝硕洋

梯度下降法是机器学习和深度学习的重要优化算法，在搜索和比较了各种梯度下降法后，总结成此篇结果。

1. 共轭梯度法

共轭梯度法（Conjugate Gradient）是介于[最速下降法](https://baike.baidu.com/item/%E6%9C%80%E9%80%9F%E4%B8%8B%E9%99%8D%E6%B3%95/7186948?fromModule=lemma_inlink" \t "_blank)与[牛顿](https://baike.baidu.com/item/%E7%89%9B%E9%A1%BF/0?fromModule=lemma_inlink)法之间的一个方法，它仅需利用一阶[导数](https://baike.baidu.com/item/%E5%AF%BC%E6%95%B0/0?fromModule=lemma_inlink)信息，但克服了最速下降法收敛慢的缺点，又避免了[牛顿](https://baike.baidu.com/item/%E7%89%9B%E9%A1%BF/5463?fromModule=lemma_inlink)法需要存储和计算Hesse[矩阵](https://baike.baidu.com/item/%E7%9F%A9%E9%98%B5/0?fromModule=lemma_inlink)并求逆的缺点，共轭梯度法不仅是解决大型[线性方程组](https://baike.baidu.com/item/%E7%BA%BF%E6%80%A7%E6%96%B9%E7%A8%8B%E7%BB%84/5904308?fromModule=lemma_inlink)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。 在各种优化算法中，共轭梯度法是非常重要的一种。其优点是所需存储量小，具有步收敛性，稳定性高，而且不需要任何外来参数。

1. 牛顿法

梯度下降法是用梯度来建立迭代关系式的方法，而牛顿法则是用切线来建立迭代关系式的算法（所以也叫切线法）。牛顿法的迭代过程与梯度下降法有相似之处，只不过是用切线与x轴的交点来作为下一轮迭代的起点，第一次迭代是从f（x0）开始，沿着切线的相反方向一直前进到与x轴的交点x1处。第二次迭代从点x1的值f(x1)开始，前进到f（x1）处的切线与x轴的交点x2处。如此持续进行，逐步逼近x\*点。需要注意的是，牛顿法是用来求解方程的，因此f(x)与x轴必须有交点x\*，这是牛顿法应用的前提。牛顿法的精髓就是二阶收敛，不仅利用了损失函数的一阶偏导数，也用到了损失函数的二阶偏导数，即梯度变化的趋势，因此比梯度下降法更快的确定合适的搜索方向，具有二阶收敛速度。

1. SGD

随机梯度下降(SGD)是一种在凸损失函数(如(线性)[支持向量机](https://en.wikipedia.org/wiki/Support_vector_machine)和[Logistic回归](https://en.wikipedia.org/wiki/Logistic_regression))下拟合线性分类器和回归器的简单而有效的方法。

随机梯度下降的优点是：

* 高效
* 易于实现 (有大量优化代码的机会)
* 随机梯度下降的缺点包括：
* SGD需要一些超参数，例如正则化参数和迭代次数
* SGD对特征缩放非常敏感

1. ADAM

adam算法是一种基于“momentum”思想的随机梯度下降优化方法，通过迭代更新之前每次计算梯度的一阶moment和二阶moment，并计算滑动平均值，后用来更新当前的参数。这种思想结合了Adagrad算法的处理稀疏型数据，又结合了RMSProp算法的可以处理非稳态的数据。如今深度学习的应用愈加广泛，各种大型的计算模型比如[BERT](https://zhida.zhihu.com/search?content_id=108167056&content_type=Article&match_order=1&q=BERT&zhida_source=entity)，XLNET等参数数量级百万甚至上亿的模型，对于复杂模型的优化的收敛效率的研究成为了方向。ADAM算法的代码本身并不复杂，缺解决了由于数据分布复杂性带来的一些优化收敛慢的问题。

1. Momentum

Momentum是为解决SGD中出现的曲折摆动问题。 “之”字形的上下摆动，降低了损失函数到达最低点的速度。此情况下，若想减少摆动浮动，只能采用比较小的learning rate，这同样将导致寻优的速度较低。而Momentum就是为解决此问题而来。动量是衡量让运动物体停下难度的物理量。当动量momentum越大时，其转换为势能的能量也就越大，越易摆动，难以停下来。动量法几乎总是比标准的梯度下降法速度更快，算法的主要思想是计算梯度的指数加权平均，然后使用这个梯度来更新权重。可以理解为，在到达新的一点时，SGD会直接按照该点的负梯度方向去更新，而Momentum会考虑之前的梯度及方向，即动量。更新的时候考虑梯度均值（指数加权平均），指数衰减理解为摩擦力造成的损失。

为了克服传统梯度下降法的一些缺点，研究人员提出了许多改进算法，如随机梯度下降（SGD）通过随机选择样本计算梯度，减少了计算量并有助于跳出局部最优；Adagrad、Adadelta、Adam 等自适应学习率算法，能够根据参数的更新情况自动调整学习率，提高算法的收敛速度和稳定性。这些改进算法进一步拓展了梯度下降法在机器学习中的应用范围，使其能够更好地适应各种复杂的模型和数据。

总之，梯度下降法是机器学习领域中不可或缺的优化算法，为模型的训练和优化提供了重要的技术支持，推动了机器学习技术的不断发展和应用。

参考资料：

1. [深度学习之动量法-csdn](https://blog.csdn.net/weixin_46713695/article/details/123197350)
2. [Adam-知乎](https://zhuanlan.zhihu.com/p/90169812)
3. [随机梯度下降-scikit-learn](https://scikit-learn.org.cn/view/84.html)
4. [共轭梯度法-百度百科](https://baike.baidu.com/item/%E5%85%B1%E8%BD%AD%E6%A2%AF%E5%BA%A6%E6%B3%95/7139204)
5. [深度学习优化方法总结-csdn](https://blog.csdn.net/sunflower_sara/article/details/81321886)