Poll的笔记

- [三叶草精神] what hurts mor
pain of hard work or the pai
regret?

首页
管理 博

随笔 - 63 文章 - 1 评论

00361312

昵称：Poll的笔记
园龄：2年1个月
粉丝：516
关注：14
+加关注

‹ 2017年7月

日	一	二	三	四	五
25	26	27	28	29	30
2	3	4	5	6	7
9	10	11	12	13	14
16	17	18	19	20	21
23	24	25	26	27	28
30	31	1	2	3	4

最新随笔

1. [Machine Learning] 深度学习中的梯度
2. [Machine Learning] logistic函oftmax函数
3. [Machine Learning & Algorith经网络基础
4. [Machine Learning] Active Leg
5. [Machine Learning & Algorith ML机器学习系列2：深入浅出ML ropy-Based家族
6. [Machine Learning & Algorith ML机器学习系列1：深入浅出ML ression家族
7. [Data Structure] LCSs——最共子序列和最长公共子串
8. [Algorithm & NLP] 文本深度表型——word2vec&doc2vec词向量
9. [Algorithm] 机器学习算法常用总结
10. [Linux] Linux常用文本操作命理

[Machine Learning] 梯度下降法的三种形式BGD、SGD以及MBGD

阅读目录

在应用机器学习算法时，我们通常采用梯度下降法来对采用的算法进行训练。其实，常用的梯度下降法还具体包含有三种不同的形式，它们也各自有着不同的优缺点。

下面我们以线性回归算法来对三种梯度下降法进行比较。

一般线性回归函数的假设函数为：

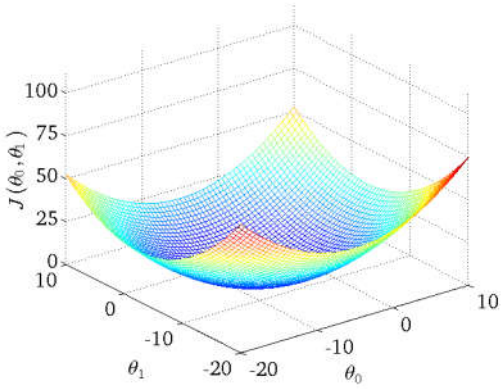
$$h_\theta = \sum_{j=0}^{n} \theta_j x_j$$

对应的能量函数（损失函数）形式为：

方差作为LOSS函数？

$$J_{train}(\theta) = 1/(2m) \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

下图为一个二维参数（θ_0和θ_1）组对应能量函数的可视化图：



回到顶部

1. 批量梯度下降法BGD

批量梯度下降法（Batch Gradient Descent，简称BGD）是梯度下降法最原始的形式，它的具体思路是在更新每一参数时都使用所有的样本来进行更新，其数学形式如下：

(1) 对上述的能量函数求偏导：

所有样本点的每一个点都需要算一次，累加和

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}(y^i - h_\theta(x^i))x_j^i$$

(2) 由于是最小化风险函数，所以按照每个参数θ的梯度负方向来更新每个θ：

$$\theta_j' = \theta_j + \frac{1}{m}\sum_{i=1}^{m}(y^i - h_\theta(x^i))x_j^i$$

具体的伪代码形式为：

repeat{

21 0

随笔分类

Algorithm(23)

Bash(1)

C/C++(6)

Computational Advertising(1)

Data Structure(6)

Database(3)

Evolutionary Algorithm(2)

Hadoop(4)

Linux(6)

Machine Learning(15)

Math(2)

Network(2)

Operate System

Python(11)

Recommendation System(1)

Search Engine(3)

Social Network Analysis(1)

Web Development(2)

生活杂谈(1)

随笔档案

2017年1月 (1)

2016年7月 (1)

2016年6月 (1)

2016年5月 (4)

2016年4月 (2)

2016年3月 (2)

2016年2月 (2)

2016年1月 (1)

2015年12月 (5)

2015年11月 (3)

2015年10月 (1)

2015年9月 (5)

2015年8月 (8)

2015年7月 (8)

2015年6月 (19)

My Team

OMEGA team

每一次都使用所有的样本点，并且使用更新后的参数继续计算梯度，继续迭代更新参数

$$\theta_j' = \theta_j + \frac{1}{m}\sum_{i=1}^{m}(y^i - h_\theta(x^i))x_j^i$$

(for every j=0, … , n)

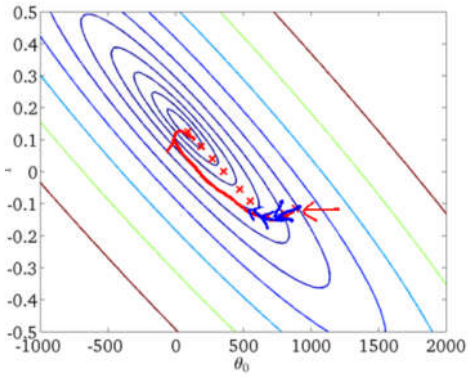
}

从上面公式可以注意到，它得到的是一个全局最优解，但是每迭代一步，都要用到训练集所有的数据，如果样本数目m很大，那么可想而知这种方法的迭代速度！所以，这就引入了另外一种方法，随机梯度下降。

优点：全局最优解；易于并行实现；

缺点：当样本数目很多时，训练过程会很慢。

从迭代的次数上来看，BGD迭代的次数相对较少。其迭代的收敛曲线示意图可以表示如下：



回到顶部

2. 随机梯度下降法SGD

由于批量梯度下降法在更新每一个参数时，都需要所有的训练样本，所以训练过程会随着样本数量的加大而变得异常的缓慢。随机梯度下降法（Stochastic Gradient Descent，简称SGD）正是为了解决批量梯度下降法这一弊端而提出的。

将上面的能量函数写为如下形式：

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(y^i - h_\theta(x^i))^2 = \frac{1}{m}\sum_{i=1}^{m}cost(\theta,(x^i,y^i))$$

$$cost(\theta,(x^i,y^i)) = \frac{1}{2}(y^i - h_\theta(x^i))^2$$

使用每个点误差直接更新一次参数，不再使用所有点的误差平均值。

利用每个样本的损失函数对θ求偏导得到对应的梯度，来更新θ：

$$\theta_j' = \theta_j + (y^i - h_\theta(x^i))x_j^i$$

具体的伪代码形式为：

1. Randomly shuffle dataset ;

2. **repeat**{

for i=1, … , m {

$$\theta_j' = \theta_j + (y^i - h_\theta(x^i))x_j^i$$

(for j=0, … , n)

}

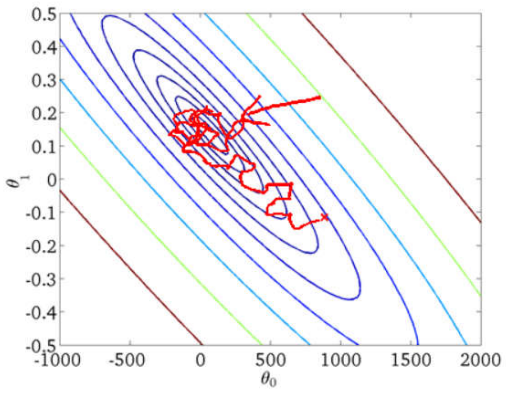
}

随机梯度下降是通过每个样本来迭代更新一次，如果样本量很大的情况（例如几十万），那么可能只用其中几万条或者几千条的样本，就已经将theta迭代到最优解了，对比上面的批量梯度下降，迭代一次需要用到十几万训练样本，一次迭代不可能最优，如果迭代10次的话就需要遍历训练样本10次。但是，SGD伴随的一个问题是噪音较BGD要多，使得SGD并不是每次迭代都向着整体最优化方向。

优点：训练速度快；

缺点：准确度下降，并不是全局最优；不易于并行实现。

从迭代的次数上来看，SGD迭代的次数较多，在解空间的搜索过程看起来很盲目。其迭代的收敛曲线示意图可以表示如下：



回到顶部

3. 小批量梯度下降法MBGD

有上述的两种梯度下降法可以看出，其各自均有优缺点，那么能不能在两种方法的性能之间取得一个折衷呢？即，算法的训练过程比较快，而且也要保证最终参数训练的准确率，而这正是小批量梯度下降法（Mini-batch Gradient Descent，简称MBGD）的初衷。

MBGD在每次更新参数时使用b个样本（b一般为10），其具体的伪代码形式为：

Say b=10, m=1000.

Repeat{

for i=1, 11, 21, 31, ... , 991{

$$\theta_j := \theta_j - \alpha\frac{1}{10}\sum_{k=i}^{i+9}(h_\theta(x^{(k)}) - y^{(k)})x_j^{(k)}$$

(for every j=0, ... , n)

}

}

回到顶部

4. 总结

Batch gradient descent: Use all examples in each iteration ;

Stochastic gradient descent: Use 1 example in each iteration ;

Mini-batch gradient descent: Use b examples in each iteration.

阮一峰的网络日志
算法，数学，文学，科技，创业...
石山园
Hadoop入门进阶课程系列
淘宝技术部
淘宝技术介绍
王路情
Hadoop研究和R实战
小坦克
网络协议介绍

积分与排名

积分 - 132162
排名 - 1820

分类: Algorithm,Machine Learning

标签: Machine Learning

« 上一篇：[Network Analysis] 复杂网络分析总结
» 下一篇：博客目录

posted @ 2015-12-30 19:46 Poll的笔记 阅读(14909) 评论(1) 编辑 收藏

评论列表

#1楼 2017-03-20 23:01 Cppowboy

网上好多介绍随机梯度下降的，这篇是我读过的最清楚的一篇，伪代码讲得很明白，还有互相的对比和评价，非常棒的一篇博客，收藏了！

支持(0) 反对(0)