

随机梯度下降（Stochastic gradient descent）和 批量梯度下降（Batch gradient descent）的公式对比、实现对比

原创 玉心sober 最后发布于2013-05-25 21:21:45 阅读数 181806 ☆ 收藏

居

梯度下降（GD）是最小化风险函数、损失函数的一种常用方法，随机梯度下降和批量梯度下降是两种迭代求解思路，下面从公式和实现的角度对两者进行分析，有哪个方面写的不对，希望网友纠正。

下面的h(x)是要拟合的函数，J(theta)损失函数，theta是参数，要迭代求解的值，theta求解出来了那最终要拟合的函数h(theta)就出来了。其中m是训练集的记录条数，j是参数的个数。

$$h(\theta)=\sum_{j=0}^n\theta_jx_j$$
$$J(\theta)=\frac{1}{2m}\sum_{i=1}^m(y^i-h_{\theta}(x^i))^2$$

1、批量梯度下降的求解思路如下：

（1）将J(theta)对theta求偏导，得到每个theta对应的的梯度

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

（2）由于是要最小化风险函数，所以按每个参数theta的梯度负方向，来更新每个theta

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

（3）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m很大，那么可想而知这种方法的迭代次数会很多，以，这就引入了另外一种方法，随机梯度下降。

2、随机梯度下降的求解思路如下：

（1）上面的风险函数可以写成如下这种形式，损失函数对应的是训练集中每个样本的粒度，而上面批量梯度下降对应的是所有的训练样本：

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

（2）每个样本的损失函数，对theta求偏导得到对应梯度，来更新theta

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

（3）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

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。

3、对于上面的linear regression问题，与批量梯度下降对比，随机梯度下降求解的会是最优解吗？

(1) 批量梯度下降---最小化所有训练样本的损失函数，使得最终求解的是全局的最优解，即求解的参数是使得风险函数最小。

(2) 随机梯度下降---最小化每条样本的损失函数，虽然不是每次迭代得到的损失函数都向着全局最优方向，但是大的整体的方向是向全局最优解的，最终的结果往往是在全局最优解附近。

4、梯度下降用来求最优解，哪些问题可以求得全局最优？哪些问题可能局部最优解？

对于上面的linear regression问题，最优化问题对theta的分布是unimodal，即从图形上面看只有一个peak，所以梯度下降最终求得的是全局最优解。然而对于multimodal的问题，因为存在多个peak值，很有可能梯度下降的最终结果是局部最优。

5、随机梯度和批量梯度的实现差别

以前一篇博文中NMF实现为例，列出两者的实现差别（注：其实对应python的代码要直观的多，以后要练习多写python！）

```
1 // 随机梯度下降，更新参数
2 public void updatePQ_stochastic(double alpha, double beta) {
3     for (int i = 0; i < M; i++) {
4         ArrayList<Feature> Ri = this.dataset.getDataAt(i).getAllFeature();
5         for (Feature Rij : Ri) {
6             // eij=Rij.weight-PQ for updating P and Q
7             double PQ = 0;
8             for (int k = 0; k < K; k++) {
9                 PQ += P[i][k] * Q[k][Rij.dim];
10            }
11            double eij = Rij.weight - PQ;
12
13            // update Pik and Qkj
14            for (int k = 0; k < K; k++) {
15                double oldPik = P[i][k];
16                P[i][k] += alpha
17                    * (2 * eij * Q[k][Rij.dim] - beta * P[i][k]);
18                Q[k][Rij.dim] += alpha
19                    * (2 * eij * oldPik - beta * Q[k][Rij.dim]);
20            }
21        }
22    }
23 }
24
25 // 批量梯度下降，更新参数
26 public void updatePQ_batch(double alpha, double beta) {
27
28     for (int i = 0; i < M; i++) {
29         ArrayList<Feature> Ri = this.dataset.getDataAt(i).getAllFeature();
30
31         for (Feature Rij : Ri) {
32             // Rij.error=Rij.weight-PQ for updating P and Q
33             double PQ = 0;
34             for (int k = 0; k < K; k++) {
35                 PQ += P[i][k] * Q[k][Rij.dim];
36             }
37             Rij.error = Rij.weight - PQ;
38         }
39     }
40 }
```



举报

```
41 |         for (int i = 0; i < M; i++) {
42 |             ArrayList<Feature> Ri = this.dataset.getDataAt(i).getAllFeature();
43 |             for (Feature Rij : Ri) {
44 |                 // 对参数更新的累积项
45 |                 double eq_sum = 0;
46 |                 double ep_sum = 0;
47 |
48 |                 for (int ki = 0; ki < M; ki++) { // 固定k和j之后,对所有i项加和
49 |                     ArrayList<Feature> tmp = this.dataset.getDataAt(i).getAllFeature();
50 |                     for (Feature Rj : tmp) {
51 |                         if (Rj.dim == Rij.dim)
52 |                             ep_sum += P[ki][k] * Rj.error;
53 |                     }
54 |                 }
55 |                 for (Feature Rj : Ri) { // 固定k和i之后,对多有j项加和
56 |                     eq_sum += Rj.error * Q[k][Rj.dim];
57 |                 }
58 |
59 |                 // 对参数更新
60 |                 P[i][k] += alpha * (2 * eq_sum - beta * P[i][k]);
61 |                 Q[k][Rij.dim] += alpha * (2 * ep_sum - beta * Q[k][Rij.dim]);
62 |             }
63 |         }
64 |     }
65 | }
66 | }
```

👍 点赞 47 ☆ 收藏 🔄 分享 ...



玉心sober

发布了10 篇原创文章 · 获赞 33 · 访问量 39万+

私信

关注



想对作者说点什么...



哈煌 2年前

这里有一个疑问，可以这样解释吧：传统的GD，就是文中所说的BGD，每次迭代更新都使用了全部的训练样本。为了加快训练速度，有了SGD，其中可分为mini-batch SGD：也就是每次使用小批量的样本来训练；以及文中描述的mini-batch为1的SGD。



xijinping1_ 2年前

这个cost写得实在是妙，误导了多少人以为是 $\cos(t)$ 还有，这两个公式其实是一回事，主要的区别是更新权重的时机问题吧



驼房营汽车工程师 1年前

文中的拟合函数写的有歧义，theta是参数，不是自变量，不应该写在h()括号内

登录 查看 39 条热评

机器学习：随机梯度下降法

阅读数 750

1.梯度下降 1) 什么是梯度下降？因为梯度下降是一种思想，没有严格的定义，所以用一个比喻来解释什么是梯度下... 博文 来自： weixin_30622181...

学习笔记13：随机梯度下降法（Stochastic gradient descent, SGD）

阅读数 2875

假设我们提供了这样的数据样本（样本值取自于 $y=3 \times x_1+4 \times x_2$ ）： x_1 x_2 y 1 4 192 5 265 1 194 2 29 x_1 和 x_2 是样本值，... 博文 来自： Softdiamonds的博客

【转载】梯度下降算法的参数更新公式

阅读数 8367

NN这块的公式，前馈网络是矩阵乘法。损失函数的定义也是一定的。但是如何更新参数看了不少描述，下面的叙述... 博文 来自： rickhuan的博客

梯度下降法的推导（非常详细、易懂的推导）

阅读数 4万+

梯度下降算法的公式非常简单，”沿着梯度的反方向（坡度最陡）“是我们日常经验得到的，其本质的原因到底是什么... 博文 来自： liupc的学习笔记

ML笔记：随机梯度下降法(Stochastic gradient descent, SGD)、BGD、MSGD+Momentum！

阅读数 2025

随机梯度下降法(Stochastic gradient descent, SGD)+python实现！ 文章目录一、设定样本二、梯度下降法原理一、... 博文 来自： Deving Zhang

随机梯度下降法概述与实例

阅读数 8895

机器学习算法中回归算法有很多，例如神经网络回归算法、蚁群回归算法，支持向量机回归算法等，其中也包括本篇... 博文 来自： 不清不白的博客

批量梯度下降法（BGD）、随机梯度下降法（SGD）和小批量梯度下降法（MBGD）

阅读数 8103

梯度下降法作为机器学习中较常使用的优化算法，其有着三种不同的形式：批量梯度下降（BatchGradientDescent）... 博文 来自： Andyato的博客

梯度下降与随机梯度下降概念及推导过程

阅读数 1万+

接前一章:常用算法一多元线性回归详解2(求解过程)同这一章的梯度下降部分加起来,才是我们要讲的如何求解多元线... 博文 来自： 激进的蜗牛

梯度下降法（Gradient Descent）推导和示例

阅读数 8776

梯度下降法（Gradient Descent）推导和示例注：作者在其他文献的基础上进行整理，形成本文的基本脉络，并希望... 博文 来自： weixin_42278173...

...法(Stochastic Gradient Descent)和批量梯度下降法..._CSDN博客

随机梯度下降(Stochastic gradient descent)和 批量梯..._CSDN博客

1:

随机梯度下降(SGD)与经典的梯度下降法的区别

阅读数 3530

随机梯度下降(SGD)与经典的梯度下降法的区别经典的优化方法，例如梯度下降法，在每次迭代过程中需要使用所有... 博文 来自： 米兰小子SHC

随机梯度下降(Stochastic gradient descent)和 批量梯..._CSDN博客

...Stochastic gradient descent)和 批量梯度下降(Batc..._CSDN博客

机器学习（四）：批量梯度下降法（BGD）、随机梯度下降法（SGD）和小批量梯度下降法（MBGD）

阅读数 1304

本文基于吴恩达老师的机器学习课程。看了吴恩达老师的机器学习课程，收获很多，想把课上学做的笔记结合自己的... 博文 来自： Sakuya__的博客



weixin_30622181

4482篇文章

关注 排名:千里之外



Softdiamonds

49篇文章


关注 排名:千里之外



rickhuan08

8篇文章

关注 排名:千里之外



/home/liupc

489篇文章

关注 排名:6000+

...随机梯度下降(Stochastic gradient descent)和 批量..._CSDN博客

...批量梯度下降) 和 stochastic gradient descent(随..._CSDN博客

:

随机梯度下降法

阅读数 5万+

一、考虑一下线性方程组 博文 来自： Forever-守望

https://blog.csdn.net/lilyth_lilyth/article/details/8973972

Page 4 of 8

随机梯度下降（SGD）、批量梯度下降（BGD）、小批量梯度下降（MSGD）	阅读数 364
转自：https://www.2cto.com/net/201610/557111.html接触过神经网络的人都知道，网络的训练是其核心，本人在读... 博文 来自： 小白白	
...批量梯度下降)和 stochastic gradient descent(随..._CSDN博客	
...批量梯度下降)和 stochastic gradient descent(随..._CSDN博客	11
基于随机梯度下降的矩阵分解算法	阅读数 833
import pandas as pdimport numpy as npimport os def difference(left,right,on): #求两个dataframe的差集 df... 博文 来自： qq_25628891的博客	
随机梯度下降法(Stochastic gradient descent)和 批量梯度下降...	11
随机梯度下降求解矩阵分解的sample（M=UV类型分解）	阅读数 2353
以下是代码，3小时搞定，完成的一刻，非常喜悦。原始矩阵可以理解为5个用户对5件衣服的点评，其中用户1,2对第... 博文 来自： xiaolu的专栏	
梯度下降学习率的设定策略	阅读数 2002
发现一篇写的很好的关于学习率的文章本文转载自卢明冬的博客-梯度下降学习率的设定策略1.学习率的重要性1）学... 博文 来自： 得克特	
随机梯度下降SGD算法原理和实现	阅读数 172
backpropagationbackpropagation解决的核心问题损失函数c与w,b求偏导，(c为cost(w,b))整体来说，分两步1.z=w*a'... 博文 来自： mercies的博客	
几种梯度下降方法对比（Batch gradient descent、Mini-batch gradient descent 和 stochastic gradient des...	阅读数 1万+
几种梯度下降方法对比（Batch gradient descent、Mini-batch gradient descent 和 stochastic gradient descent）&a... 博文 来自： 天泽28的专栏	
梯度下降与随机梯度下降	阅读数 6983
梯度下降法先随机给出参数的一组值，然后更新参数，使每次更新后的结构都能够让损失函数变小，最终达到最小即... 博文 来自： Marshall的专栏	
梯度下降法(steepest descent)和共轭梯度法(conjugate gradient)	阅读数 1343
写一篇自己的理解，算不上严格意义的证明，事实上很多熟悉的公式和推导方式都没有摆上来。推导的过程没有参考... 博文 来自： tanmx219的博客	
梯度下降法原理完全式手推	阅读数 148
梯度下降法我手写在了纸上了，思路是先这样，然后再那样，之后这样，最后再那样，是不是很简单？函数单调导数与... 博文 来自： qq_23016555的博客	
梯度下降 和反向传播推导（公式）	阅读数 694
1、训练算法几乎都是使用梯度来使得代价函数下降，大多数都是对随机梯度下降算法的改进。目标函数关于的梯度... 博文 来自： qxsunshine的博客	
梯度下降及具体计算方式	阅读数 1万+
阅读目录1. 批量梯度下降法BGD2. 随机梯度下降法SGD3. 小批量梯度下降法MBGD4. 总结 在应用机器学习算法... 博文 来自： 翻墙的老王	
随机梯度下降	阅读数 227
随机梯度下降(SGD)是一种简单但又非常高效的方法，主要用于凸损失函数下线性分类器的判别式学习，例如(线性)... 博文 来自： Iceforest的博客	
梯度下降算法以及随机梯度下降算法的原理及python代码	阅读数 314
梯度下降算法梯度下降，依照所给数据，判断函数，随机给一个初值w，之后通过不断更改，一步步接近原函数的方... 博文 来自： Dr.sen的博客	
基于批量随机梯度下降的非负矩阵分解	阅读数 578
非负矩阵分解(NMF)NMF的基本思想为什么分解的矩阵式非负？为什么要运用非负矩阵分解？NMF的基本思想：对于... 博文 来自： qq_25628891的博客	
随机梯度下降(SGD)和批量梯度下降(BGD)的区别	阅读数 853
我的机器学习教程「美团」算法工程师带你入门机器学习 以及「三分钟系列」数据结构与算法已经开始更新了，欢... 博文 来自： Machine Learning ...	

- 批**梯度下降法**(Batch Gradient Descent), 小批**梯度下降** (Mini-Batch GD), **随机梯度下降** (Stochastic GD)

阅读数 6294

一、梯度下降法 在机器学习算法中，对于很多监督学习模型，需要对原始的模型构建损失函数，接下来便是通过... [博文](#) 来自: [cs24k1993的博客](#)
- 处理soup.select()中的填写以及爬取信息出现空列表的情况

阅读数 3493

soup.select以及爬取信息出现空列表的情况举例一、先说soup.select()中的填写方法一方法二headers的修改方法， ... [博文](#) 来自: [Prodigal](#)
- 手推机器学习--**梯度下降法** 与 最小二乘法

阅读数 196

只是做个笔记，思路说明不是很详细，想看详细推导与说明的请参考凸优化求解的书籍，各位路过的兄弟不要见怪。... [博文](#) 来自: [刘宏宇的博客](#)
- 深度学习中的**随机梯度下降**(SGD)简介

阅读数 6568

随机梯度下降(StochasticGradientDescent,SGD)是梯度下降算法的一个扩展。 机器学习中反复出现的一个... [博文](#) 来自: [网络资源是无限的](#)
- 批量梯度下降法**与 **随机梯度下降** 区别

阅读数 499

1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... [博文](#) 来自: [JH_Zhai的博客](#)
- 随机梯度下降法**（SGD）

阅读数 1502

有一组数据，需要进行拟合，（拟合后可以去做很多事，做很多事都需要数据拟合，比如机器学习，从样本中学习也就... [博文](#) 来自: [Whiteleaf3er的博客](#)
- 批量梯度下降**、**随机梯度下降**与**小批量梯度下降**算法之间的比较

阅读数 882

这三种算法都用于反向传播的优化损失函数算法。在每轮迭代中更新一次权重w，根据多次迭代，最终无限的靠近我... [博文](#) 来自: [lcczzu的专栏](#)
- 随机梯度下降**算法

阅读数 9318

BP神经网络是神经网络的基础，其中可用随机梯度下降算法来加快更新权值和偏置，但是随机梯度下降算法总是忘... [博文](#) 来自: [qqzj_bupt的博客](#)
- 三种**梯度下降**的方式：**批量梯度下降**、**小批量梯度下降**、**随机梯度下降**

阅读数 3万+

在机器学习领域中，梯度下降的方式有三种，分别是：批量梯度下降法BGD、随机梯度下降法SGD、小批量梯度下... [博文](#) 来自: [UESTC_C2_403的...](#)
- Coursera-AndrewNg(吴恩达)机器学习第一周笔记

阅读数 5709

引言Introduction1 Welcome2 什么是机器学习What is Machine Learning3 监督学习Supervised Learning4 无监督学... [博文](#) 来自: [scrueit的博客](#)
- 梯度下降**算法-R语言

阅读数 5739

#读取自变量、因变量数据x [博文](#) 来自: [yuanhangzhegogo...](#)
- 初探**梯度下降**之**随机梯度下降**（SGD）

阅读数 98

随机梯度下降算法先解释一些概念。1.什么是梯度下降我们先从一张图来直观解释这个过程。如上图假设这样一个场... [博文](#) 来自: [kwame211的博客](#)
- 深度学习理论——**随机梯度下降**法(SGD) & 反向传播

阅读数 4298

大家好，一直在用深度学习，但是感觉理论并不扎实，打算开始补点理论基础，在CSDN上记录下来。今天介绍随机... [博文](#) 来自: [Miss_yuki的博客](#)
- Java C语言 Python C++ C# Visual Basic .NET JavaScript PHP SQL Go语言 R语言 Assembly language Swift Ruby

MATLAB PL/SQL Perl Visual Basic Objective-C Delphi/Object Pascal Unity3D



玉心sober

TA的个人主页 >

原创

粉丝

获赞

评论

访问

10

500

33

107

39万+

等级: 博客 4

周排名: 14万+

积分: 1205

总排名: 6万+

关注

私信



Kindergarten

≤ 6th

1st Grade

7th

2nd Grade

≈ 8th

3rd Grade

x² Alg

4th Grade

△ Geo

5th Grade

Σ Pre

最新文章

CTR预估中GBDT与LR融合方案

对数线性模型之一(逻辑回归), 广义线性模型学习总结

c++ 学习总结

linux、Hadoop相关的常用东西总结

大规模数据相似度计算时，解决数据倾斜的问题的思路之一（分块思想）

分类专栏

 矩阵分解

2篇

 梯度下降

2篇

 最优化

2篇

 正态分布

1篇

 NMF

1篇

展开

归档

2015

8月

1篇

2013

8月

1篇

7月

5篇

5月

3篇

热门文章

随机梯度下降（Stochastic gradient descent）和 批量梯度下降（Batch

阅读数 181749

CTR预估中GBDT与LR融合方案

阅读数 97469

对数线性模型之一(逻辑回归), 广义线性模型学习总结

阅读数 45517

正态分布具有很多好的性质，很多模型假设数据服从正态分布。但是如果数据不服从正

阅读数 29256

NMF(非负矩阵分解)的SGD（随机梯度下降）实现

阅读数 12326

最新评论

CTR预估中GBDT与LR融合方案

qq_39651918: [reply]oSuYan1234[/reply]你好，我也对这个有疑惑。请问你明白了吗

CTR预估中GBDT与LR融合方案

bingcore: [reply]yi_yunfei[/reply] 看文章，并没有以id为特征建树。就是把id拉出来，按id划分数 ...

随机梯度下降（Stochastic...

distanter: 文中的拟合函数写的有歧义，theta是参数，不是自变量，不应该写在 h() 括号内

CTR预估中GBDT与LR融合方案

GitzLiu: 写的真是好。

CTR预估中GBDT与LR融合方案

u014512961: [reply]yi_yunfei[/reply] 兄弟知道怎么建了吗

亿速云香港服务器免备案1

亿速云香港服务器稳定可免费换ip 免备CN2线路低至29元-支持试用，24小时售

打开

QQ客服

kefu@csdn.net

客服论坛

400-660-0108

工作时间 8:30-22:00

关于我们 招聘 广告服务 网站地图

京ICP备19004658号 经营性网站备案信息

公安备案号 11010502030143

京网文〔2020〕1039-165号

©1999-2020 北京创新乐知网络技术有限公司

网络110报警服务

北京互联网违法和不良信息举报中心

中国互联网举报中心 家长监护

版权与免责声明 版权申诉