**梯度下降对比试验**

**Data set:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Trainging  n | Features  d | 稀疏性 | Size |  |
| Rcv1 | 677,399 | 47,236 | 0.16% | 1.2G | N>>d |
| Susy | 2,181,583 | 18 |  | 1.2G | N>>d |
| Url | 2,396,130 | 3,231,961 |  | 2.2G | N~=d |
| News | 19,996 | 1,355,191 |  | 140M | N<<d |
| covtype | 581,012 | 54 |  | 58M | N<<d |

**实验１：对比　SVRG,Momentum,Mini-Batch SGD,Adagrad**

说明:

1.下面是多次实验，可以取得的最好的实验效果

2.每次实验都需要调节参数比如：SVRG的内部更新次数，采样比例等

3.时间单位 S

数据集1RCV

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Method | 1e-7 | 1e-6 | 1e-5 | 1e-4 | 1e-3 | 1e-2 | 1e-1 |
| SVRG | 2.6 | 2.5 | 2.5 | 2.1 | 2.0 | 1.5 | 1.5 |
| Momentum | 4.7 | 3.9 | 3.6 | 2.7 | 2.6 | 1.5 | 1.3 |
| SGD | 4.2 | 2.6 | 1.8 | 1.2 | 1.2 | 1.3 | 1.1 |
| Adagrad | 不收敛 | - |  |  |  |  |  |

数据集2Susy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | 1e-7 | 1e-6 | 1e-4 | 1e-2 |
| SVRG | 7.8 | 6.8/6.1/7.4/8.0 | 6.6 | 5.4/5.2 |
| Momentum | 8 | 6.5/6 | 4 | 3.4 |
| SGD | 10 | 7.9 | 4.3 | 3.4 |
| Adagrad | 24 | 25 | 22 | 20 |

说明：

1似乎SVRG 在 1e-6 波动挺大,但是在高精度范围内,效果还是好于SGD,在高精度收敛要求下,比SGD快 20%-30%

2:动量法在数据集RCV上不如,SGD,在数据集Susy上优于SGD

3:Adagrad在RCV数据集上不收敛,在Susy上,效果很差.

数据集3 Url size:2.2G # of data: 2,396,130 # of features: 3,231,961

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SVRG | 0.00001 |  |  |  |  |  |
| #iterations | 1 | 3 | 5 | **10** | 20 | 30 |
| Time(s) | 11 | 23 | 47 | **83** | 169 | 252 |
| AUC | 0.0 | 0.82 | 0.83 | **0.85** | 0.85 | 0.87 |
| Weight精度 | - | 0.87 | 0.29 | **0.11** | 0.06 | 0.002 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SGD |  |  |  |  |  |  |  |  |  |
| #iterations | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 100 |
| Time(s) | 40 | 76/78 | 109 | 141 | 178 | 202 | 251 | 279 | 343 |
| AUC | 0.0 | 0.80/0.66 | 0.76 | 0.81 | 0.82 | 0.89 | 0.87 | 0.86 | 0.86 |
| Weight精度 | 1.06 | 0.74/0.64 | 0.53 | 0.36 | 0.24 | 0.12 | 0.03 | 0.002 | 3.7e-6 |

SGD:迭代20次的时候,抖动严重,交替出现两种结果

SGD:迭代60次的时候,AUC较高,可以认为是抖动.

结论:获得一个不错的结果,SVRG可以比SGD快2-3倍(对比上面红色数据)

**实验二AdaGrad 和AdaDelta对不同数据集AUC对比,差别明显**

**下面是100轮迭代**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | News20 | Covtype | Rcv1 | Susy | Url |  |
| adagrad | 0.04 | 0.39 | 0.21 | 0.69 | 0.04 |  |
| adadelta | 0.38 | 1.0 | 1.0 | 1.0 | 1.0 |  |
| SGD | 0.78 | 0.64 | 0.92 | 0.69 | 0.86 |  |

**猜想:对于N<<D ,或者N~=D的情况,adadelta 效果很差;对于N>>D的情况,adadelta 效果很好**

**Leu:** #of data: 38#of features: 7129 #4M

**Real-sim:**# of data: 72,309# of features: 20,958 #90M

**Gisette-scale:# of data: 6,000# of features: 5,000**

**#253M**

**Madelon:# of data: 2,000 # of features: 500 #7.8M**

**100轮迭代**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | leu | real-sim | Gisette-  scale | madelon |  |  |
| adagrad | 0.0 | 0.24 | 0.0 | 0.0 |  |  |
| adadelta | 0.7 | 0.49 | 0.0 | 1.0 |  |  |
| SGD | 0.9 | 0.86 | 0.0 | 1.0 |  |  |

**实验三:单机模式**

数据集3 Url size:2.2G # of data: 2,396,130 # of features: 3,231,961

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SVRG | 0.00001 |  |  |  |  |  |
| #iterations | 1 | 3 | 5 | **10** | 20 | 30 |
| Time(s) | 11 | 23 | 47 | **83** | 169 | 252 |
| AUC | 0.0 | 0.82 | 0.83 | **0.85** | 0.85 | 0.87 |
| Weight精度 | - | 0.87 | 0.29 | **0.11** | 0.06 | 0.002 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SGD |  |  |  |  |  |  |  |  |  |
| #iterations | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 100 |
| Time(s) | 40 | 76/78 | 288 | 141 | 178 | 202 | 251 | 279 | 343 |
| AUC | 0.0 | 0.80/0.66 | 0.66 | 0.81 | 0.82 | 0.89 | 0.87 | 0.86 | 0.86 |
| Weight精度 | 1.06 | 0.74/0.64 | 0.44 | 0.36 | 0.24 | 0.12 | 0.03 | 0.002 | 3.7e-6 |