|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| yin数据集 | 数据样例 | 维度 | 稠密度 | 非零特征总数 | 数据大小 | 记录个数 | 数据用途 |
| HIGGS | 0(label)  1:1.343286752700805664e+00 2:-5.571643114089965820e-01 3:-1.376299738883972168e+00 | 28 | 92.1057124059742% | 283685620 | 7.9GB | 11000000 | 分类  (2类) |
| Real-sim | -1(label)  2:0.03275694851688871 12:0.03375173419589727 23:0.1455569855611168 | 20,958 | 0.24474755069536114% | 3709083 | 90.4M | 72,309 | 分类(2类) |
| kdda | 1 618:1 3800:1 3813:1 3814:1 3815:1 3816:1 | 20,216,830 | 1.7979580103397633E-4% | 305613510 | 2.7GB | 8,407,752 / 510,302 | 分类2 |
| kddb | 1 618:1 3800:1 3813:1 3814:1 3815:1 3816:1 | 29,890,095 | 9.835711404175401E-5 | 566345888 | 5.1GB | 19,264,097 / 748,401 | 分类2 |
| Kdd2012 | 1 618:1 3800:1 3813:1 3814:1 3815:1 3816:1 | 54,686,452 | 3.680148867539499E-5% | 1646030155 | 22.5GB/18GB(tr)/4.5GB(val) | 149,639,105 / 119,705,032 (tr) / 29,934,073 (val) | 分类2 |
| Kdd-raw | 1 618:1 3800:1 3813:1 3814:1 3815:1 3816:1 | 1,163,024 |  |  |  | 19,264,097 / 748,401 | 分类2 |
| url\_combined | -1 4:0.0995851 5:0.158621 6:0.117647 11:0.285714 | 3,231,961 | 0.0035776267979986433% | 277058644 | 2.2GB | 2,396,130 | 2 |
| rcv1\_train.binary | -1 224:0.0981792780330669 915:0.10454237017157 2302:0.0778896736716234 | 47,236 | 0.15676161783119918% | 1498952 | 36.5M | 20,242 / 677,399 | 2 |
| epsilon | 1 1:0.0117784 2:0.0185069 3:-0.0198577 4:-0.0276363 | 2,000 | 99.999750000625% | 800000000 | 11.3GB | 400,000 / 100,000 | 2 |
| webspam | +1 592138:0.251252 592139:0.00946336 592142:0.00378534 | 16,609,143 | 0.022443640457610985% | 1304697446 | 25GB | 350,000 | 2 |
| SUSY | 0 1:9.728614687919616699e-01 2:6.538545489311218262e-01 3:1.176224589347839355e+00 | 18 | 98.81973936890431% | 38805059 | 1.1GB | 5,000,000 | 2 |

ALS算法数据集：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | User | Item | Ratings | density | Rating Scale |
| Movielens 1M | 6040 | 3883 | 1,000,209 | 4.26% | [1 5] |
| Movielens 10M | 69,878 | 10,681 | 10,000,054 | 1.33% | [0.5 5] |
| Movielens 20M | 138,493 | 27,278 | 20,000,263 | 0.52% | [0.5 5] |
| Movielens  latest | 260,000 | 40,000 | 24,000,000 | 0.2307692% | [0.5 5] |
| Jester | 124,113 | 150 | 5,865,235 | 31.50% | [10, 10] |
| Book-Crossing | 92,107 | 271,379 | 1,031,175 | 0.0041% | [1, 10], and implicit |
| Last.fm | 1892 | 17632 | 92,834 | 0.28% | - |
| Wikipedia | 5,583,724 | 4,936,761 | 417,996,366 | 0.0015% | - |
| OpenStreetMap (Azerbaijan) | 231 | 108,330 | 205,774 | 0.82% | - |
| Git (Django) | 790 | 1757 | 13,165 | 0.95% | - |
| Netflix Prize | 2649429 | 17770 | |  | | --- | | 100480507 | | 0.2134236% | [1 5] |

推荐系统常用的数据集下载：

1. Million Song Dataset Challenge：

<https://www.kaggle.com/c/msdchallenge#description>

1. Amazon product data

<http://jmcauley.ucsd.edu/data/amazon/links.html>

1. MovieLens - Movie Recommendation Data Sets

<http://www.grouplens.org/node/73>

1. Yahoo! - Movie, Music, and Images Ratings Data Sets

<http://webscope.sandbox.yahoo.com/catalog.php?datatype=r>

1. Jester - Movie Ratings Data Sets (Collaborative Filtering Dataset)

<http://www.ieor.berkeley.edu/~goldberg/jester-data/>

1. Cornell University - Movie-review data for use in sentiment-analysis experiments

<http://www.cs.cornell.edu/people/pabo/movie-review-data/>

1. Last.fm - Music Recommendation Data Sets

<http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/index.html>

1. Yahoo! - Movie, Music, and Images Ratings Data Sets

<http://webscope.sandbox.yahoo.com/catalog.php?datatype=r>

1. Audioscrobbler - Music Recommendation Data Sets

<http://www-etud.iro.umontreal.ca/~bergstrj/audioscrobbler_data.html>

1. Amazon - Audio CD recommendations

<http://131.193.40.52/data/>

1. Institut für Informatik, Universität Freiburg - Book Ratings Data Sets

<http://www.informatik.uni-freiburg.de/~cziegler/BX/>

1. National University of Singapore - Scholarly Paper Recommendation

<http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html>

1. the [Jester](http://shadow.ieor.berkeley.edu/humor/) Online Joke Recommender System

<http://goldberg.berkeley.edu/jester-data/>

1. Retailrocket recommender system dataset

<https://www.kaggle.com/retailrocket/ecommerce-dataset>

15、amazon-reviews-1995-2013\_archive.torrent(迅雷种子)

<https://ia601306.us.archive.org/24/items/amazon-reviews-1995-2013/amazon-reviews-1995-2013_archive.torrent>

16、Book-Crossing Dataset

<http://www2.informatik.uni-freiburg.de/~cziegler/BX/>

17 Netflix Prize Data Set (迅雷种子)

<http://academictorrents.com/details/9b13183dc4d60676b773c9e2cd6de5e5542cee9a>