分类算法数据集：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 数据用途 | 数据样例 | 维度 | 稠密度 | 非零特征总数 | 数据大小 | 样例个数 | 数据集  描述 |
| HIGGS  (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) | 分类（数据分为2类） | 0 (label)  1:1.343286752700805664e+00 2:-5.571643114089965820e-01 3:-1.376299738883972168e+00 | 28 | 92.10% | 283685620 | 7.9GB | 11000000 | 描述的是希格斯玻色子的振动情况。根据分类的结果发现异常粒子。 |
| Real-sim  (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) | 分类（数据分为2类） | -1 (label)  2:0.03275694851688871 12:0.03375173419589727 23:0.1455569855611168 | 20958 | 0.25% | 3709083 | 90.4M | 72309 | 数据集中包含了四组不同的数据：模拟的赛车数据、模型飞行数据、真实赛车数据、真实飞行数据。分类目的是把模拟数据和真实数据分开。 |
| Kdda  (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) | 分类（数据分为2类） | 1 (label)  618:1 3800:1 3813:1 3814:1 3815:1 3816:1 | 20216830 | 1.80E-4% | 305613510 | 2.7GB | 8407752 / 510302 | KDD CUP2010比赛数据,由当年获胜者national  Taiwan Univ大学提供。 |
| Kddb  (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) | 分类（数据分为2类） | 1 (label)  618:1 3800:1 3813:1 3814:1 3815:1 3816:1 | 29890095 | 9.84E-5% | 566345888 | 5.1GB | 19264097 / 748401 | KDD CUP2010比赛数据,由当年获胜者national  Taiwan Univ大学提供。 |
| Kdd2012  (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) | 分类（数据分为2类） | 1 (label)  618:1 3800:1 3813:1 3814:1 3815:1 3816:1 | 54686452 | 3.68E-5% | 1646030155 | 22.5GB/18GB(tr)/4.5GB(val) | 149639105 / 119705032 (tr) / 29934073 (val) | KDD CUP2012数据集 |
| Kdd-raw  (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) | 分类（数据分为2类） | 1 (label)  618:1 3800:1 3813:1 3814:1 3815:1 3816:1 | 1163024 | - | - | - | 19264097 / 748401 | KDD CUP2010原始数据 |
| url\_combined  (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) | 分类（数据分为2类） | -1 (label)  4:0.0995851 5:0.158621 6:0.117647 11:0.285714 | 3231961 | 0.004% | 277058644 | 2.2GB | 2396130 | 数据集包含不同的URL地址，分类的目的是找到不好(malicious)的链接。 |
| rcv1\_train.binary  (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) | 分类（数据分为2类） | -1 (label)  224:0.0981792780330669 915:0.10454237017157 2302:0.077889673671623 | 47236 | 0.16% | 1498952 | 36.5M | 20242 / 677399 | 路透社发布的700000条手工分类的文档集合。确定某个文档属于某个类别。 |
| Epsilon(http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) | 分类（数据分为2类） | 1 (label)  1:0.0117784 2:0.0185069 3:-0.0198577 4:-0.0276363 | 2000 | 99.99% | 800000000 | 11.3GB | 400000 / 100000 | 是Pascal large scale learning challenge比赛的模拟数据集 |
| Webspam  (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) | 分类（数据分为2类） | +1 (label)  592138:0.251252 592139:0.00946336 592142:0.00378534 | 16609143 | 0.02% | 1304697446 | 25GB | 350000 | 数据集收集不同的URL地址，判断URL是否是垃圾网页。 |
| SUSY  (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/) | 分类（数据分为2类） | 0 (label)  1:9.728614687919616699e-01 2:6.538545489311218262e-01 3:1.176224589347839355e+00 | 18 | 98.82% | 38805059 | 1.1GB | 5000000 | 数据集记录了在物理试验中的产生的新的超对称粒子，其中包括可见粒子和不可见粒子，这些粒子具有特定的物理学特征。 |

说明：1、数据用途是分类，上述数据集的类别标签只有两个。

2、

ALS算法数据集：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据集 | User | Item | Ratings | density | Rating Scale |
| Movielens 1M | 6040 | 3883 | 1000209 | 4.26% | [1 5] |
| Movielens 10M | 69878 | 10681 | 10000054 | 1.33% | [0.5 5] |
| Movielens 20M | 138493 | 27278 | 20000263 | 0.52% | [0.5 5] |
| Movielens latest | 260000 | 40000 | 24000000 | 0.23% | [0.5 5] |
| Netflix Prize | 2649429 | 17770 | |  | | --- | | 100480507 | | 0.21% | [1 5] |
| Book-Crossing | 92107 | 271379 | 1031175 | 0.004% | [1 10] and implicit |
| Last.fm | 1892 | 17632 | 92,834 | 0.28% | - |
| Wikipedia | 5583724 | 4936761 | 417996366 | 0.002% | - |
| OpenStreetMap (Azerbaijan) | 231 | 108330 | 205774 | 0.82% | - |
| Git (Django) | 790 | 1757 | 13165 | 0.95% | - |

说明：1、implicit

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>