

机器学习基准训练数据集收集列表

商业敏感数据虽然难以获取，但好在仍有相当多有用数据可公开访问。它们中的不少常用来作为特定机器学习问题的基准测试数据。常见的有以下几个：

UCL机器学习知识库

包括近300个不同大小和类型的数据集，可用于分类、回归、聚类 and 推荐系统任务。数据集列表位于：<http://archive.ics.uci.edu/ml/>

Amazon AWS公开数据集

包含的通常是大型数据集，可通过Amazon S3访问。这些数据集包括人类基因组项目、Common Crawl网页语料库、维基百科数据和Google BooksNgrams。相关信息可参见：<http://aws.amazon.com/publicdatasets/>。

Kaggle

这里集合了Kaggle举行的各种机器学习竞赛所用的数据集。它们覆盖分类、回归、排名、推荐系统以及图像分析领域，可从Competitions区域下载，不过其中的许多数据集是收费的：<http://www.kaggle.com/competitions>

KDnuggets

这里包含一个详细的公开数据集列表，其中一些上面提到过的。该列表位于：<http://www.kdnuggets.com/datasets/index.html>

美国联邦政府数据集

这里提供了许多关于气候变化、教育以及海洋方面的数据：<http://www.data.gov>

来自infochimps公司的数据集

来自infochimps公司的数据集，已经开放了超过14000多个数据集，不过遗憾的是很多数据集是要购买的：<http://www.infochimps.com/>

百万歌曲数据库

百万歌曲数据库，来自哥伦比亚大学，默认的百万曲库达到了280GB。就连这个数据库的样本版（10,000首歌）都达到了1.8 GB（已经压缩过）：<http://labrosa.ee.columbia.edu/millionsong/>

蛋白质信息资源数据集

提供了基因序列，蛋白质和生物研究的数据集，很多关于DNA的机器学习的论文都用了这个网站的数据源：<http://pir.georgetown.edu>

手写数字图像数据集

训练数据集包含了60,000条手写数字图像，测试数据10,000条。在很多人的论文中都采用了这个数据集，并且用了各种算法比如KNN，SVM，NN等。是图像识别中比较好的数据集：<http://yan.n.lecun.com/exdb/mnist/>

人脸识别数据库

来自麻省大学：<http://vis-www.cs.umass.edu/fddb/>

一些贝叶斯网

<http://bndg.cs.aau.dk/html/>

kdd99数据集

<http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

数据集索引

<http://dmr.cs.umn.edu/datasets.html>

华盛顿大学

<http://www.cs.washington.edu/dm/vfml/>

搜狗数据资源

包括了评测集合、语料数据、新闻数据、图片数据以及自然语言处理相关数据：<http://www.sogou.com/labs/resources.htm>

本博客文章除特别声明，全部都是原创！
原创文章版权归过往记忆大数据（[过往记忆](#)）所有，未经许可不得转载。
本文链接：[【】（）](#)