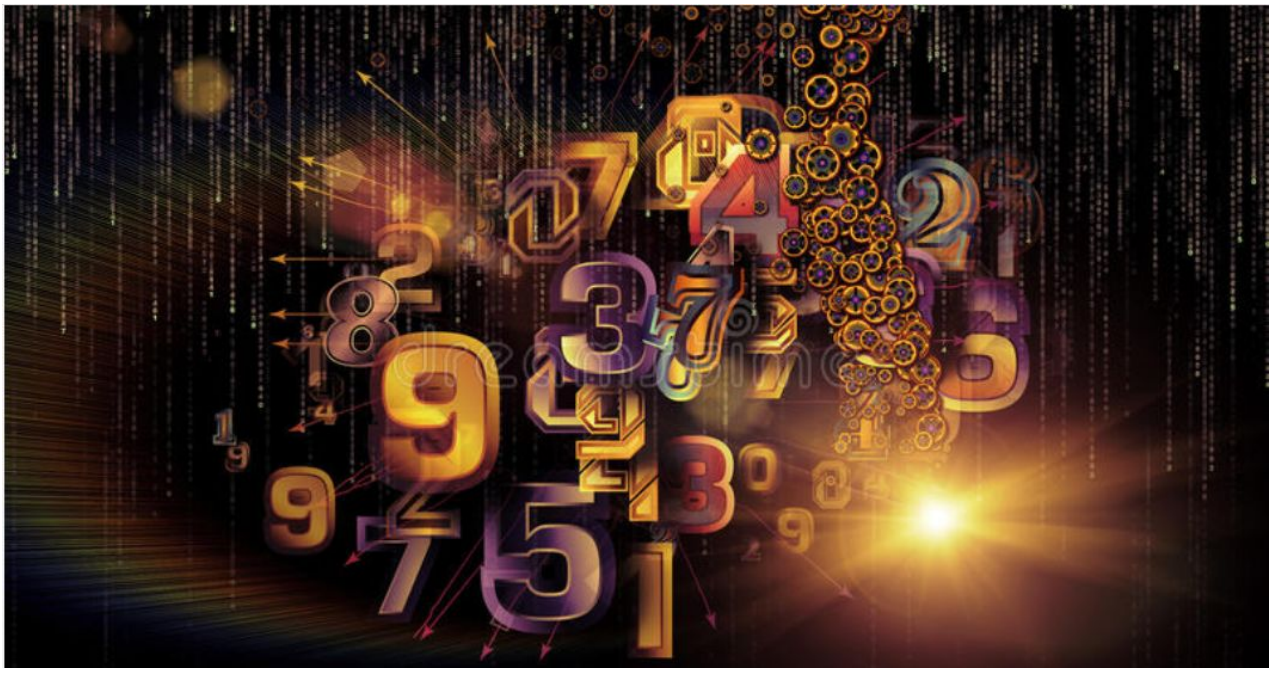
推荐收藏：50个最佳机器学习公共数据集

mlmemoirs 机器学习初学者 2022-05-03 12:00

> 外国自媒体mlmemoirs根据github、福布斯、CMU官网等信息，整理了一张50个最佳机器学习公共数据集的榜单，为大家分享一下~



作者：mlmemoirs 郭一璞 编译

外国自媒体**mlmemoirs**根据github、福布斯、CMU官网等信息，整理了一张50个最佳机器学习公共数据集的榜单，为大家分享一下~

提前说下须知：

一、寻找数据集的意义

根据CMU的说法，寻找一个好用的数据集需要注意一下几点：

数据集不混乱，否则要花费大量时间来清理数据。

数据集不应包含太多行或列，否则会难以使用。

数据越干净越好，清理大型数据集可能非常耗时。

应该预设一个有趣的问题，而这个问题又可以用数据来回答。

二、去哪里找数据集

- **Kaggle**：爱竞赛的盆友们应该很熟悉了，Kaggle上有各种有趣的数据集，拉面评级、篮球数据、甚至西雅图的宠物许可证。
  https://www.kaggle.com/
- **UCI机器学习库**：最古老的数据集源之一，是寻找有趣数据集的第一站。虽然数据集是用户贡献的，因此具有不同的清洁度，但绝大多数都是干净的，可以直接从UCI机器学习库下载，无需注册。
  http://mlr.cs.umass.edu/ml/
- **VisualData**：分好类的计算机视觉数据集，可以搜索~
  https://www.visualdata.io/

好了，下面就是那50个数据集了，由于后期加上了一些补充，所以总数已经超过了50。

机器学习数据集

图片

- Labelme：带注释的大型图像数据集。
  http://labelme.csail.mit.edu/Release3.0/browserTools/php/dataset.php

- ImageNet：大家熟悉的ImageNet，女神李飞飞参与创建，同名比赛影响整个计算机视觉界。
  http://image-net.org/
- LSUN：场景理解与许多辅助任务（房间布局估计，显着性预测等）
  http://lsun.cs.princeton.edu/2016/
- MS COCO：同样也是知名计算机视觉数据集，同名比赛每年都被中国人屠榜。
  http://mscoco.org/
- COIL 100 ：100个不同的物体在360度旋转的每个角度成像。
  http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php
- 视觉基因组：非常详细的视觉知识库。
  http://visualgenome.org/
- 谷歌开放图像：在知识共享下的900万个图像网址集合“已经注释了超过6000个类别的标签”。
  https://research.googleblog.com/2016/09/introducing-open-images-dataset.html
- 野外标记面：13000张人脸标记图像，用于开发涉及面部识别的应用程序。
  http://vis-www.cs.umass.edu/lfw/
- 斯坦福狗子数据集：20580张狗子的图片，包括120个不同品种。
  http://vision.stanford.edu/aditya86/ImageNetDogs/
- 室内场景识别：包含67个室内类别，15620个图像。
  http://web.mit.edu/torralba/www/indoor.html

情绪分析

- 多域情绪分析数据集：一个稍老一点的数据集，用到了来自亚马逊的产品评论。
  http://www.cs.jhu.edu/~mdredze/datasets/sentiment/
- IMDB评论：用于二元情绪分类的数据集，不过也有点老、有点小，有大约25000个电影评论。
  http://ai.stanford.edu/~amaas/data/sentiment/
- 斯坦福情绪树库：带有情感注释的标准情绪数据集。
  http://nlp.stanford.edu/sentiment/code.html
- Sentiment140：一个流行的数据集，它使用160,000条预先删除表情符号的推文。
  http://help.sentiment140.com/for-students/
- Twitter美国航空公司情绪：2015年2月美国航空公司的Twitter数据，分类为正面，负面和中性推文。
  https://www.kaggle.com/crowdflower/twitter-airline-sentiment

自然语言处理

- HotspotQA数据集：具有自然、多跳问题的问答数据集，具有支持事实的强大监督，以实现更易于解释的问答系统。
  https://hotpotqa.github.io/
- 安然数据集：来自安然高级管理层的电子邮件数据。
  https://www.cs.cmu.edu/~./enron/
- 亚马逊评论：包含18年来亚马逊上的大约3500万条评论，数据包括产品和用户信息，评级和文本审核。
  https://snap.stanford.edu/data/web-Amazon.html
- Google Books Ngrams：Google Books中的一系列文字。
  https://aws.amazon.com/datasets/google-books-ngrams/
- Blogger Corpus：收集了来自blogger.com的681,288篇博文，每篇博文至少包含200个常用英语单词。
  http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm
- 维基百科链接数据：维基百科的全文，包含来自400多万篇文章的近19亿个单词，可以按段落、短语或段落本身的一部分进行搜索。
  https://code.google.com/p/wiki-links/downloads/list

- Gutenberg电子书列表：Gutenberg项目中带注释的电子书书单。
  http://www.gutenberg.org/wiki/Gutenberg:Offline_Catalogs

- Hansards加拿大议会文本：来自第36届加拿大议会记录的130万组文本。
  http://www.isi.edu/natural-language/download/hansard/

- Jeopardy：来自问答节目Jeopardy的超过200,000个问题的归档。
  http://www.reddit.com/r/datasets/comments/1uyd0t/200000_jeopardy_questions_in_a_json_file/

- 英文垃圾短信收集：由5574条英文垃圾短信组成的数据集。
  http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/

- Yelp评论：Yelp，就是美国的“大众点评”，这是他们发布的一个开放数据集，包含超过500万条评论。
  https://www.yelp.com/dataset

UCI的Spambase：一个大型垃圾邮件数据集，对垃圾邮件过滤非常有用。
https://archive.ics.uci.edu/ml/datasets/Spambase

自动驾驶

- Berkeley DeepDrive BDD100k：目前最大的自动驾驶数据集，包含超过100,000个视频，其中包括一天中不同时段和天气条件下超过1,100小时的驾驶体验。其中带注释的图像来自纽约和旧金山地区。
  http://bdd-data.berkeley.edu/

- 百度Apolloscapes：度娘的大型数据集，定义了26种不同物体，如汽车、自行车、行人、建筑物、路灯等。
  http://apolloscape.auto/

- Comma.ai：超过7小时的高速公路驾驶，细节包括汽车的速度、加速度、转向角和GPS坐标。
  https://archive.org/details/comma-dataset

- 牛津的机器人汽车：这个数据集来自牛津的机器人汽车，它于一年时间内在英国牛津的同一条路上，反反复复跑了超过100次，捕捉了天气、交通和行人的不同组合，以及建筑和道路工程等长期变化。
  http://robotcar-dataset.robots.ox.ac.uk/

- 城市景观数据集：一个大型数据集，记录50个不同城市的城市街景。
  https://www.cityscapes-dataset.com/

- CSSAD数据集：此数据集对于自动驾驶车辆的感知和导航非常有用。不过，数据集严重偏向发达国家的道路。
  http://aplicaciones.cimat.mx/Personal/jbhayet/ccsad-dataset

- KUL比利时交通标志数据集：来自比利时法兰德斯地区数以千计的实体交通标志的超过10000条注释。
  http://www.vision.ee.ethz.ch/~timofter/traffic_signs/

- MIT AGE Lab：在AgeLab收集的1,000多小时多传感器驾驶数据集的样本。
  http://lexfridman.com/automated-synchronization-of-driving-data-video-audio-telemetry-accelerometer/

- LISA：UC圣迭戈智能和安全汽车实验室的数据集，包括交通标志、车辆检测、交通信号灯和轨迹模式。
  http://cvrr.ucsd.edu/LISA/datasets.html

- 博世小交通灯数据集：用于深度学习的小型交通灯的数据集。
  https://hci.iwr.uni-heidelberg.de/node/6132

- LaRa交通灯识别：巴黎的交通信号灯数据集。
  http://www.lara.prd.fr/benchmarks/trafficlightsrecognition

- WPI数据集：交通灯、行人和车道检测的数据集。
  http://computing.wpi.edu/dataset.html

临床

- MIMIC-III：MIT计算生理学实验室的公开数据集，标记了约40000名重症监护患者的健康数据，包括人口统计学、生命体征、实验室测试、药物等维度。
  https://mimic.physionet.org/

一般数据集

除了机器学习专用的数据集，还有一些其他的一般数据集，可能很有趣~

公共政府数据集

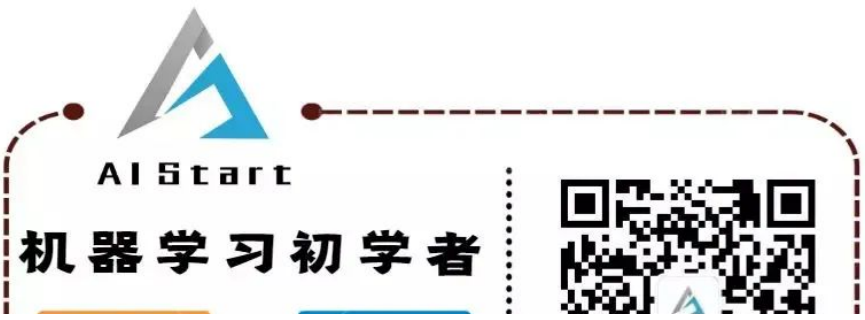
- Data.gov：该网站可以从多个美国政府机构下载数据，包括各种奇怪的数据，从政府预算到考试分数都有。不过，其中大部分数据需要进一步研究。
  https://www.data.gov/
- 食物环境地图集：本地食材如何影响美国饮食的数据。
  https://catalog.data.gov/dataset/food-environment-atlas-f4a22
- 学校财务系统：美国学校财务系统的调查。
  https://catalog.data.gov/dataset/annual-survey-of-school-system-finances
- 慢性病数据：美国各地区慢性病指标数据。
  https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi-e50c9
- 美国国家教育统计中心：教育机构和教育人口统计数据，不仅有美国的数据，也有一些世界上其他地方的数据。
  https://nces.ed.gov/
- 英国数据服务：英国最大的社会、经济和人口数据集。
  https://www.ukdataservice.ac.uk/
- 数据美国：全面可视化的美国公共数据。
  http://datausa.io/
- 中国国家统计局。
  http://www.stats.gov.cn/

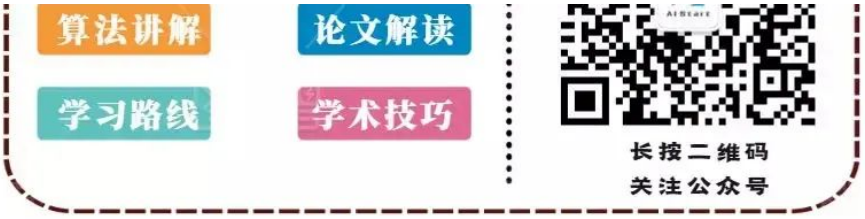
金融与经济

- Quandl：经济和金融数据的良好来源，有助于建立预测经济指标或股票价格的模型。
  https://www.quandl.com/
- 世界银行开放数据：全球人口统计数据，还有大量经济和发展指标的数据集。
  https://data.worldbank.org/
- 国际货币基金组织数据：国际货币基金组织公布的有关国际金融，债务利率，外汇储备，商品价格和投资的数据。
  https://www.imf.org/en/Data
- 金融时报市场数据：来自世界各地的金融市场的最新信息，包括股票价格指数，商品和外汇。
  https://markets.ft.com/data/
- Google Trends：世界各地的互联网搜索行为和热门新闻报道的数据。
  http://www.google.com/trends?q=google&ctab=0&geo=all&date=all&sort=0
- 美国经济协会：美国宏观经济数据。
  https://www.aeaweb.org/resources/data/us-macro-regional

备注：有一些网址需要科学上网才能打开。

暂时手头没有工具怎么办？先收藏呀！





往期精彩回顾

适合初学者入门人工智能的路线及资料下载

(图文+视频)机器学习入门系列下载

中国大学慕课《机器学习》（黄海广主讲）

机器学习及深度学习笔记等资料打印

《统计学习方法》的代码复现专辑

AI基础下载

机器学习交流qq群955171419，加入微信群请扫码：

喜欢此内容的人还喜欢

号称最强深度学习笔记本电脑，雷蛇与Lambda公司推出，售价超2万

机器之心

我是吴恩达：人在美国，刚上知乎，先答个「如何系统学习机器学习」

量子位

Transformer论文引用破4万，两位作者离开谷歌创业

机器之心