机器学习大礼包——高质量数据集

常用的搜索网站

| UCI Machine Learning Repository

最著名的UCI数据集库，许多论文的数据均来源于此。

| AWS Public Datasets

亚马逊云服务提供的数据集，涵盖天文、生物、化学、天气、经济等多领域。

| YAHOO Webscope datasets

雅虎提供的数据集，包含图像、语言、排名分类等多领域数据。

| Kaggle datasets

Kaggle竞赛平台提供的数据集库，能在里面发现很多来自工业界有趣的数据，比如Uber、Netflix Prize、McDonald's等的数据。

计算机视觉

| ImageNet

图像处理最著名的数据集，可以根据你的项目需求搜索任一种类的图像，用来做对象识别，定位，分类和屏幕解析等问题。有14197122个不同尺寸的图像，总计140GB。

| MNIST

基本上是新提出的机器学习算法必跑的一个数据集。MNIST是一个手写数字数据库，它有60000个训练样本集和10000个测试样本集，是NIST数据库的一个子集。

| The CIFAR-10 dataset

32x32 彩色图像。

| Google Open Images

Google Open Images 是Google公司开放的大型图像标注数据集，包含 900万张图像中 7800种类别内容的标注。

自然语言处理

| 文本分类数据集

由 DBPedia、Amazon、Yelp、Yahoo!、Sogou 和 AG的文本分类数据整合成的一个大型数据集。样本大小从 120K 到 3.6M, 问题从 2 级到 14 级。

| WikiText

维基百科文章中的大型语言建模语料库。

| Billion Words

常用来训练如word2vec或Glove的分布式词表征

| Stanford Sentiment Treebank

用于情感分析的数据集

语音识别

| 2000 HUB5 English

英语的语音数据。

| CHIME

包含噪声的语音识别数据集

| TED-LIUM

TED演讲的语音数据集，有对应的全文本。

其它类

| UCR Time Series

时间序列界的“Imagnet”,发文章必跑。

| Million Song Dataset

做音乐推荐或分类的程序员可能会用到。

| Netflix 推荐系统数据

电影评价数据集，该数据集中包含随机挑选的 48万 Netflix客户，对 1.7万 部电影，超过 1百万 条评价，数据时间段为 1998.10 到 2005.11。评价以5分制评分为基准，每部电影评价为1-5分，客户信息进行了脱敏处理。

| Udacity 自动驾驶数据集

Udacity 学城开放的自动驾驶课程中的自动驾驶汽车数据集，旨在打造一个开源的自动驾驶项目。多个二进制压缩文件，总计100G左右