8.1.3 机器学习的工具

2023 年 9 月 18 日

机器学习的工具

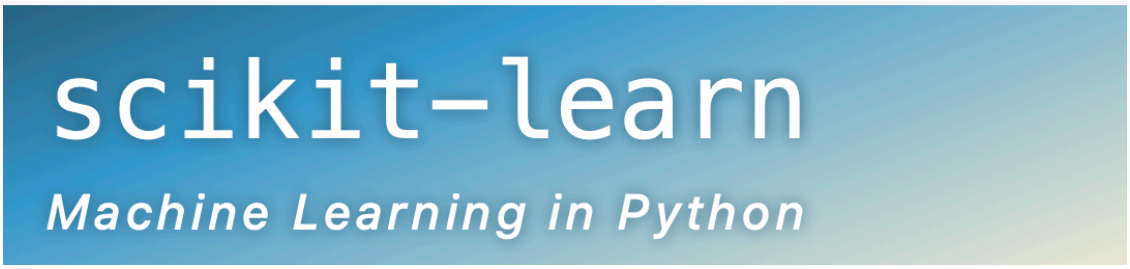
1. 开源的数据库

学习机器学习时，最好使用真实数据，而不是人工数据集。幸运的是，有上千个开源数据集可以进行选择，涵盖多个领域。以下是一些可以查找的数据的地方：

- 流行的开源数据仓库：
  UC Irvine Machine Learning Repository
  Kaggle datasets
  Amazon's AWS datasets

- 准入口（提供开源数据列表）
  http://dataportals.org/
  http://opendatamonitor.eu/
  http://quandl.com/
  https://github.com/awesomedata/awesome-public-datasets
- 其它列出流行开源数据仓库的网页：
  Wikipedia's list of Machine Learning datasets
  Quora.com question
  Datasets subreddit

2. 和 Python 相关的库



链接：scikit-learn.org

sklearn 提供基于 Python 语言的包含监督学习、无监督学习相关的机器学习模型，以及数据预处理方法等。

3. 机器学习的社区

Kaggle 是全球最大的数据科学社区，提供了数据科学相关的竞赛和数据集 [8]。很多著名企业，例如，在 Kaggle 平台上发布企业的数据和商业需求，并附带奖金。众多数据科学爱好者在网站上提交自己的解决方案，已赢取排名和奖金。此外，个人用户也可以上传和公开自己的数据集，开放给其他用户使用。这种竞争、开放和分享的氛围，让 Kaggle 在数据科学领域具有良好的口碑。链接：kaggle.com