K-MEANS聚类算法

机器学习概论-第二次实验

2024-春

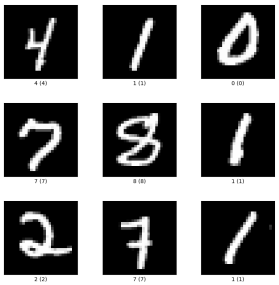
目标

- 实现K-Means聚类算法，并在真实数据集上进行评测

K-Means算法：

1. 给定聚类数K，初始化K个聚类中心:$\{g_1,\ldots,g_k\}$;
2. 把每个样本划分到离它最近的聚类中心上;
3. 对每个聚类，利用所有属于该类的样本，重新计算它的聚类中心:$\{g_1,\ldots,g_k\}$;
4. 重复以上的2-3步，直到某个终止条件。

数据集



- MNIST 数据集
  - 发布方：National Institute of Standards and Technology
  - 手写数字图像的大型数据库
  - 训练集大小：60,000；测试集大小：10,000；
  - 图像进行了归一化，确保数字居中，大小为28x28像素
  - 数据集地址：https://paperswithcode.com/dataset/mnist
- 仅使用训练数据进行K-means聚类和分析
- 数据集可以通过PyTorch、TensorFlow、Keras等下载.

```
from torchvision import datasets
train_data = datasets.MNIST(root = "./data/", train=True, download=True)
```

算法实现

- 自行编写代码，实现K-means算法

问题：

- 如何确定聚类**数量**?
- 如何**初始化**聚类中心?
- 如何用**特征向量**表示图片?
- 如何衡量样本**距离**?
- 如何设置**终止**条件?

模型评价

- 模型表现（定量）：准确率 (Accuracy)
  - 汇报训练集上的准确率
- 模型表现（定性）：可视化聚类结果

- 如何推断每个类的标签?
  - e.g., 多数投票
- 如何可视化高维数据?
  - e.g., t-SNE
  - 大数据量：采样表示
- 评估聚类结果的其他指标?

实验分析

- 实验中遇到或者探究了什么问题?
- 不同实现细节对结果的影响?
  - e.g., K的取值, 初始化聚类中心位置, 图片表示方式, 样本距离衡量
- 可视化结果的分析?
  - e.g., 反映困难样本
- 其他聚类方法的使用?

评分要求

- 实现一个K-means算法 (50%)
- 用准确率分数衡量模型表现 (20%)
- 可视化聚类结果 (20%)
- 其他实验分析 (10%)

- 注意: 评分并非基于模型性能(准确率), 而是基于如何实现算法、如何评估模型以及如何分析实验结果

作业提交格式

包含以下内容的压缩文件:

- 源代码:
  - 包含必要的注释
  - 确保助教能够理解并运行代码复现结果 (注意设置随机种子, 如聚类初始化)
- README
  - 文本文件(utf8 编码)：姓名、学号、联系方式、代码运行指引
- 实验报告:
  - PDF文件：实验设计、结果、分析、讨论
  - 请不要直接复制代码
- 请不要上传数据集

截止时间及其他

- 截止时间: 2024.04.21 (周日) 23:59:00
  - 上传压缩文件到网络学堂, 文件名为 姓名_学号
  - 迟交作业会相应扣分:
    - 迟交 ≤ 一周: 0.8;　迟交 > 一周: 0.6
    - 如有特殊情况，请提前联系助教
  - 严禁抄袭代码和报告
    - 学校提供了作业查重系统，被确认为抄袭的作业会严重扣分
- 如有任何问题，请联系助教:
  - 李佳玉，王亦凡，何祉瑜
  - {jy-li20, yf-wang21, hezy22}@mails.tsinghua.edu.cn