机器学习实验 - 2023 春

实验六：聚类和提示学习

实验内容：聚类和提示学习

作业提交截止时间：2023/06/11 23:59:59

环境要求：

Python，numpy 支持多维度的数组和矩阵运算，pandas 数据处理和分析工具，Matplotlib 图形化工具，huggingface 库。

任务一：聚类

本任务中你将自己实现 kmeans 聚类算法，将数据完成聚类，针对数据选择合适的簇数，并且自己实现聚类算法过程，尝试不同的初始中心点，观察聚类结果，将结果可视化并分析。

输入： 样本集$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
聚类簇数k.

过程：

从D中随机选择k个样本作为初始均值向量$\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k\}$
repeat
  令$C_i = \emptyset \ (1 \leq i \leq k)$
  for $j = 1, \ldots, m$ **do**
    计算样本$\boldsymbol{x}_j$与各均值向量$\boldsymbol{\mu}_i \ (1 \leq i \leq k)$的距离：$d_{ji} = ||\boldsymbol{x}_j - \boldsymbol{\mu}_i||_2$;
    根据距离最近的均值向量确定$\boldsymbol{x}_j$的簇标记：$\lambda_j = \arg\min_{i \in \{1,2,\ldots,k\}} d_{ji}$;
    将样本$\boldsymbol{x}_j$划入相应的簇：$C_{\lambda_j} = C_{\lambda_j} \bigcup \{\boldsymbol{x}_j\}$;
  end for
  for $i = 1, \ldots, k$ **do**
    计算新均值向量：$\boldsymbol{\mu}'_i = \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}$;
    if $\boldsymbol{\mu}'_i \neq \boldsymbol{\mu}_i$ **then**
      将当前均值向量$\boldsymbol{\mu}_i$更新为$\boldsymbol{\mu}'_i$
    else
      保持当前均值向量不变
    end if
  end for
until 当前均值向量均未更新
return 簇划分结果

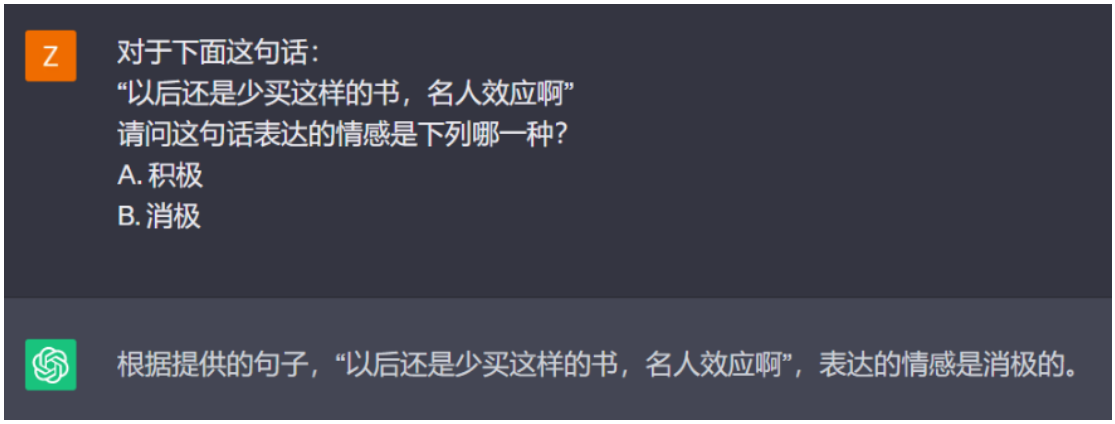
输出： 簇划分$\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

文件 ex6data.csv 包含我们的线性可分类问题的数据集。x，y 代表纵横坐标。

任务二：提示学习（Prompt Learning）

本任务中你讲尝试使用提示学习完成文本情感分析。

提示学习通过构造提示，引导预训练语言模型完成文本分析、生成等相关任务。以下是使用提示学习进行文本情感分析的一个例子。



对于原始输入“以后还是少买这样的书，名人效应啊”，我们首先构造了提示

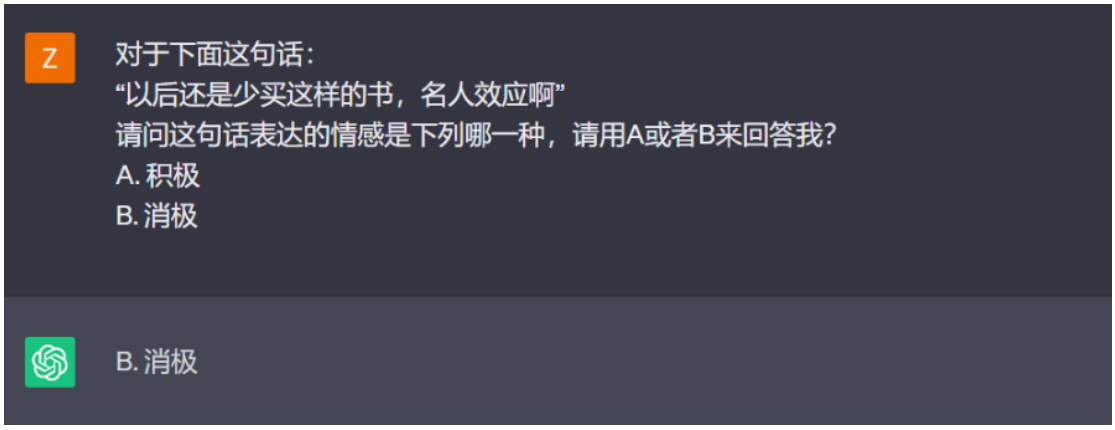
对于下面这句话：

“……”

请问这句话表达的情感是下列哪一种？

随后，我们将用提示扩展以后的输入作为预训练语言模型的输入，引导预训练模型按照我们的要求完成任务。

然而，预训练语言模型的输出可能需要进一步处理。例如，在上述例子中，我们想要预训练语言模型直接输出该句子的情感分析结果（积极/消极），然而预训练模型的输出中还包含了其他信息。针对这一情况，我们可以通过改进提示模版或者编写后处理(post processing)代码，完成从模型输出到预测值的映射。在下图中，我们通过改进提示模版，引导模型生成了更为直接的答案。



本任务中，你将以预训练语言模型 ChatGLM 为基础，对实验五中的评论数据

集进行情感分类，考虑到模型推理时间，本次实验只使用 20%的数据作为测试，并与实验五中 LSTM 的结果进行对比。

文件 comments.csv 包含我们的情感分析的数据集。

模型加载、调用相关代码可参考 https://github.com/THUDM/ChatGLM-6B

更多提示学习示例请参考： https://learn.deeplearning.ai/chatgpt-prompt-eng/lesson/5/inferring

作业提交内容：

1. 作业代码截图
2. 实验结果及分析
3. 请将以上结果保存在实验报告（pdf 或者 word 格式）中，命名为 学号+姓名+第几次实验，邮件发送到 facanhe@163.com