迁移学习项目说明文档

实践项目以小组为单位，每组人数<=4。使用所学的算法，解决训练数据分布与测试数据分布不一致情形下的情感分类问题。本项目要求如下：

1. 将基于线性回归的分类算法作用在情感分析数据集（books, dvd, electronics, kitchen）上。具体地，在一个数据集（如books）上训练分类模型，在另一个数据集（如dvd）上测试分类模型的分类精度。一共有4个数据集，可构造4\*3=12个分类任务。为方便称呼，把这种分类算法称为Least Squares Classification (LSC).
2. 将迁移学习算法（除了权重因子以外，损失函数&假设空间都与LSC算法相同）作用在情感分析数据集（books, dvd, electronics, kitchen）上。具体地，在一个数据集（如books）上训练分类模型，在另一个数据集（如dvd）上测试分类模型的分类精度。一共有4个数据集，可构造4\*3=12个分类任务。为方便称呼，把这种分类算法称为Importance Weighted Least Squares Classification (IWLSC).
3. **对比LSC 与IWLSC两种算法在每个分类任务上，测试精度的差异，尝试分析其中的原因（注意“训练数据分布与测试数据分布不一致”这个关键问题）。使用以下表格展示实验结果：**

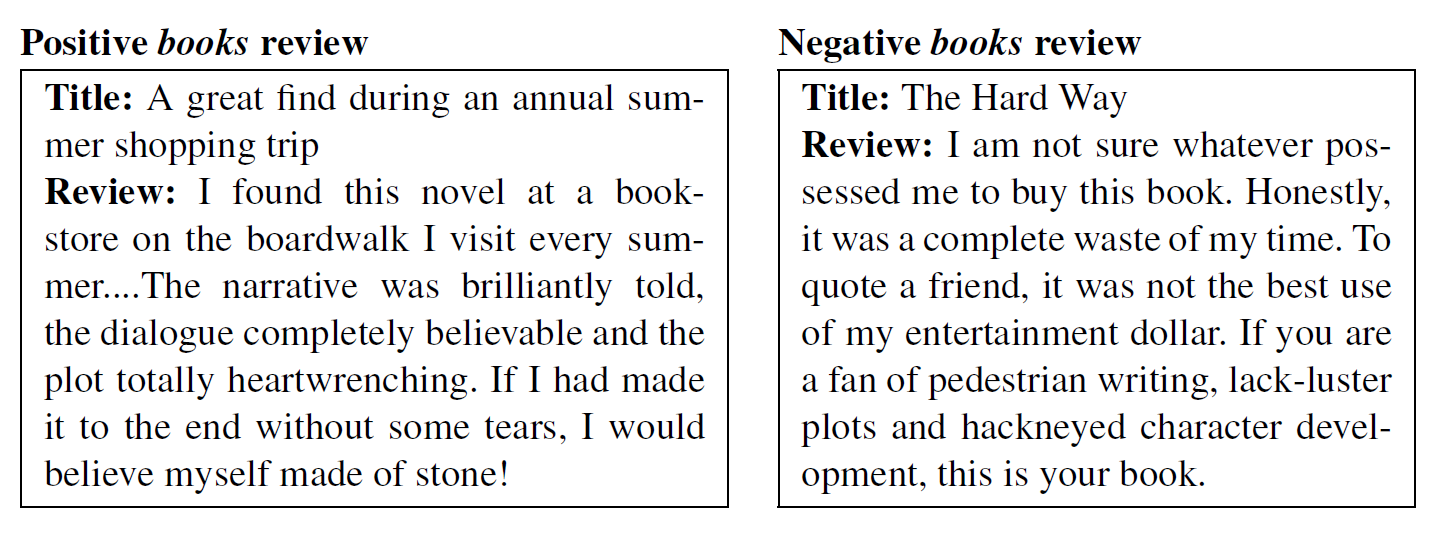
|  |  |  |
| --- | --- | --- |
| 迁移学习任务 (训练数据 to 测试数据) | LSC算法分类精度(%) | IWLSC算法分类精度(%) |
| books to dvd |  |  |
| books to electronics |  |  |
| … |  |  |
| kitchen to electronics |  |  |

1. 将这个过程的代码进行演示（代码给出必要的注释），并给出运行结果。

使用PPT展示与代码解说的方式呈现小组的项目实践结果。

附录：

1. books.svmlight，dvd.svmlight，electronics.svmlight，kitchen.svmlight 每个数据集都包含2000个样本数据。books.svmlight数据集包含关于购买书本的网上商品评论，每一个评论由一个5000维的向量表示，一共有2个类别（分别代表正面评论、负面评论）。部分评论如下图所示。



1. 文件ReadData.py提供了读取数据集的API接口，可供参考。