第二次作业报告

辛极　　2014012288

余海林　2014012280

https://github.com/Ji-Xin/UniversityDataMining

提交的zip文件只包含代码和报告，完整的数据请参考此github网页。

- 数据获取和预处理
  - 数据获取
    - 取自https://catalog.data.gov/dataset/college-scorecard的关于美国大学的各项属性表格，更新日期是March 8, 2017，其中包括7703所美国大学，以及其对应的各种属性共1743维。
    - 2017年USNews美国最佳大学排名--综合大学排名，来自https://wenku.baidu.com/view/a77a4134cdbff121dd36a32d7375a417866fc1bc.html，对美国前231名的大学做了排名。
  - 大学名字匹配`1-0-parse_name.py`。

    因为两个来源的数据中大学名字常不匹配（例如University of California-Berkeley和University of California--Berkeley，或者University of Alabama和The University of Alabama），所以要进行基于经验的消歧。例如：

    ```
    def rep(i, pat1, pat2, head, tail):
        temp = head + rank[i, 1].replace(pat1, pat2) + tail
        if temp in name_list:
            rank[i, 1] = temp

    for i in range(len(rank)):
        rep(i, "St. ", "St ", "", "")
        rep(i, "St. ", "Saint ", "", "")
        rep(i, "--", "-", "", "")
        rep(i, "--", " at ", "", "")
        rep(i, "--", "-", "The ", "")
        rep(i, "--", " at ", "The ", "")
        rep(i, "--", " in ", "", "")
        rep(i, "--", " in ", "The ", "")
        rep(i, "--", " ", "", "")
        rep(i, "--", " ", "The ", "")
        rep(i, "", "", "The ", "")
        rep(i, "&", "and", "", "")
        rep(i, "", "", "", "-Main Campus")
    ```

    这是从`1-0-parse_name.py`中截取的代码片段，目的是把大学名称中的`pat1`换成`pat2`，在加上`head`和`tail`。对于一些无法确定的学校，整个tuple被删去。最后一共有210个学校。

    保存在`data`目录下的`1-new_rank.npy`和`1-new_stat.npy`。
  - 属性选择`2-preprocess.py`。

如果某个属性在这210个学校当中，有不少于150个都是数（而不是字符串或者NULL），那么就把这个属性保留下来。一共1743个属性，由此选出499个属性。这样做，一方面是为了避开全都是字符串或者NULL的无用属性，另一方面是为了不让少量的缺失浪费整个有用属性。缺失的数据由全局平均值补齐。

保存在`data`目录下的`2-attr.npy`和`2-final_stat.npy`中。

- 训练和测试
  - 采用线性SVM的rank功能`3-svm.sh`。

    实现的源代码网址：https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html。

    每个大学对应一个属性向量x，SVM初始化一个参数向量w，两者内积即是该大学的得分。以拿到的排名为依据，优化的目标是排名高的大学得分更高，以此训练SVM。

    此后在测试集上测试效果。测试集包括6间学校，是从一开始名字无法匹配的学校里手动抽出来并且匹配名字的，抽的时候基本上是每隔四五十的名次抽一间。
  - 分析参数中每一维的内容`4-analyse.py`。

    具体每一维的含义在https://github.com/Ji-Xin/UniversityDataMining/blob/master/raw_data/useless/FullDataDocumentation.pdf 中。
- 背景介绍和未来展望

  详情请看ppt文件。