基于内容的推荐算法

姓名：黄宇辰　　学号：10185102253

---

1. 内容分析

1.1 样本集

对于每个用户，数据集中包含信息有： 用户ID，history(已点击过的新闻)，impression(待判断的标签)。

对于每条新闻，数据集中包含信息有：新闻ID，Category(主要类别)，SubCategory(次级类别)，Title(标题)，Abstract(摘要)。

1.2 特征向量矩阵构建

1.2.1 类别信息结构化

对于每条新闻，将不同主要类别给定一个评分，部分评分定义如下：

```
{'autos': 10,
 'entertainment': 20,
 'finance': 30,
 'foodanddrink': 40,
 'games': 50}
```

次级类别评分定义，类似主要类别评分，部分如下：

```
{'ads-latingrammys': 10,
 'ads-lung-health': 20,
 'advice': 30,
 'animals': 40}
```

1.2.2 标题、摘要信息结构化

标题与摘要信息为文本信息，这里使用nltk与gensim训练词向量模型，之后得到句向量模型生成特征信息。

对于文本使用正则表达式去除标点，并将字母全部转化为小写字母，再对文本进行单词切分。之后统计所有单词，生成单词表，再训练得到词向量模型。这里，个人觉得title与abstract信息得到句向量后对于结果影响较小，所以只训练得到20维的词向量。**句向量由词向量加权平均得到。**

某条新闻特征向量如下：

```
array([[ 1.40000000e+02,  1.00000000e+03,  1.12987089e+00,
        -7.24367082e-01,  1.60059357e+00,  1.20050967e+00,
         2.49791220e-02, -1.23277020e+00, -5.00418723e-01,
         7.83931732e-01, -2.00756812e+00, -2.02456787e-01,
```

```
         7.97964573e-01,  9.00134742e-01,  6.51192009e-01,
         4.64457005e-01, -5.49746454e-01,  1.16794896e+00,
        -9.48290765e-01,  1.21693611e+00,  1.22371280e+00,
        -1.51993608e+00,  0.00000000e+00,  0.00000000e+00,
         0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
         0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
         0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
         0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
         0.00000000e+00,  0.00000000e+00,  0.00000000e+00,
         0.00000000e+00,  0.00000000e+00,  0.0000000e+00]])
```

2. 模型选取与构建

2.1 模型选取

模型选取逻辑回归模型。因为基于内容的新闻推荐主要任务是判断一个用户会不会点击一条新闻，所以我们可以当作任务是一个二分类的分类任务。在此次分类任务中，由于对于每个用户来说，训练集的样本数量均在几十到几百条新闻左右，数据量较小，所以直接调用了sklearn的逻辑回归类。

2.2 正负样本选取

2.2.1 正样本选取

在正样本选取的问题上，十分简单，我们只需要将每个用户的history中的新闻特征向量提取出来，标签置为1即可。

2.2.2 负样本选取

在负样本选取的问题上，较为困难。因为对于一个用户来说，我们在数据集里面只可以看见history中，他点击进去的新闻有哪些，选取哪些新闻作为负样本是模型好坏的关键问题。

在训练集中，用户样本与测试集中用户样本只有少量重叠，所以明显不可以将训练集中的用户模型直接在测试集上使用。

这里，我统计出了所有新闻的被点击量，选取所有新闻中点击量最高的十条新闻作为参考标准。**在构建一个用户逻辑回归模型选取负样本时候，选取点击量最高的十条新闻中未被点击的新闻作为负样本。**

3. 逻辑回归准确率

3.1 训练集上结果

在逻辑回归模型构建时，最大迭代次数设置为1000。

在训练集上查看结果时，将阈值设置为0.9。即对于某一用户来说，传入一个新闻特征向量，若判定该用户会点击该新闻的概率在0.9以上，则判定该用户会点击该新闻。

在调整阈值时，我将阈值从0.5调高到0.9的过程中，发现设定阈值越高，在训练集上的准确率越高，由此可见，在基于内容的推荐算法上，使用逻辑回归模型，有一定不错的效果。准确率提高过程大致如下：0.5——28%、0.6——36%、0.8——45%、0.9——54%(阈值——准确率)。

3.2 结果分析

在内容上使用逻辑回归算法进行新闻的推荐。

由于逻辑回归是判别式模型，不关心数据如何生成，只关注数据之间的差别，通过差别对给定数据进行分类，我认为逻辑回归较为适合基于内容的推荐算法。

逻辑回归适合离散化后的特征。在此次基于内容的逻辑回归推荐算法中，我们对新闻特征向量进行了离散化处理，这样的优势有如下几条：

1. 离散后稀疏向量内积乘法运算速度更快，计算结果也方便存储，容易扩展；
2. 离散后的特征对异常值更具鲁棒性，如 age>30 为 1 否则为 0，对于年龄为 200 的也不会对模型造成很大的干扰；
3. LR 属于广义线性模型，表达能力有限，经过离散化后，每个变量有单独的权重，这相当于引入了非线性，能够提升模型的表达能力，加大拟合；
4. 离散后特征可以进行特征交叉，提升表达能力，由 M+N 个变量编程 M*N 个变量，进一步引入非线形，提升了表达能力；
5. 特征离散后模型更稳定，如用户年龄区间，不会因为用户年龄长了一岁就变化。

总的来说，特征离散化以后起到了加快计算，简化模型和增加泛化能力的作用。

4. 代码文件与结果文件

代码文件：

使用逻辑回归算法的基于内容的推荐算法.ipynb

结果文件：

test_res.tsv