

数智NLP组2022第一阶段考核(3周)

前言

第一周期考核

1 基础工具推荐

2 基础库的介绍

3 主要考核内容

3.1 模型

3.1.1 基础 (15)

难度1

难度2

3.1.2 进阶 (进阶选做) (20)

3.2 数据预处理

3.3 特征工程

3.4 评估模型

3.5 可视化

模型使用 (10)

附加题: (10)

答辩 (100)

文档要求!!! 往详细、清晰、干净的学术论文的感觉去靠 (45)

课内重要!!! 不要挂科!!!

数智NLP组2022第一阶段考核(3周)

前言

十分开心、十分高兴能够遇到大家。初次考核，我先话痨一下，大家放松放松情绪，有什么不满或者意见都可以提出来，我们相逢，我考核大家，并不是为难大家，大家都知道选拔是一个什么概念，这同时是整个社会或者说是一个系统进步必要的流程；难度会根据大家的表现来决定。有句话说，一张好地试卷，就应该让同学考出来的成绩服从正态分布，但是呢，样本量太少，我也难说出得好不好。但是一个很残酷但是亘古不变的事实便是：“优胜略汰”。这里加了双引号，我想表达的是：更优者进。可以知道，你们一选择加入我这个群的瞬间，或者是报名参加的瞬间，甚至说报名的念头开始，这种想要学习的念头就让你们已经很优秀了，都代表着你们比其他同学更加与众不同！但，因为赛场总有输赢现象、世界总有强弱之分，所以或许有时候不得不承认别人更加优秀。就算这次赢了，今后依旧面对着大把困难；就算输了，以后的路仍然需要你去行走，所以，首先端正好自己参与这次考核的姿态，你可以是“我必胜”，也可以是“我学习”。我更倾向于后者，因为，前者不一定为真，后者主动权在你。当然，你说：“我学习” and “我必胜”。Fine！

第一周期考核

1 基础工具推荐

- Pycharm很方便使用，很友好
- jupyter notebook / jupyter lab在数据预处理等工作上十分方便
- Anaconda / miniconda这俩自己选，都好，其集成了数据科学需要的许多包

2 基础库的介绍

- numpy
- pandas
- matplotlib

3 主要考核内容

3.1 模型

3.1.1 基础 (15)

其中代码颜值站5分，模型做的好能拿比较高的分数，两个模型都可以拿到10分，难度高的几率大。

难度1

- 线性回归模型（不用框架）
 - 1、使用最小二乘法【难度较小】
 - 2、使用梯度下降法【难度比上方的大】
 - 或者使用牛顿法

难度2

- softmax回归模型（多元分类）（不用框架）
 - 使用梯度下降法

以上难度二选一

3.1.2 进阶（进阶选做） (20)

提供K-means算法

提供决策树（CART）

3.2 数据预处理

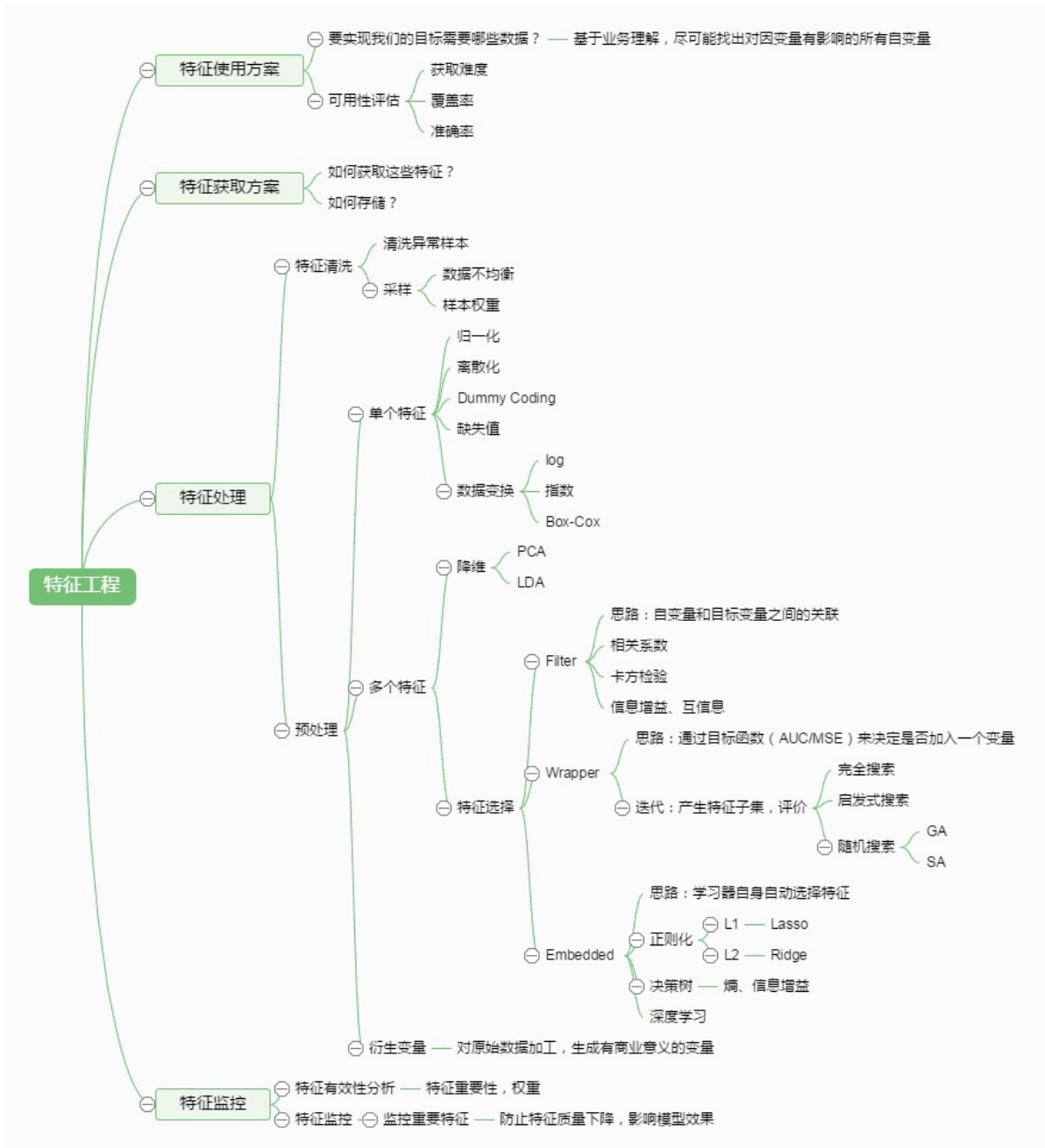
学习以下几种数据处理方法：

- 归一化和标准化
- 独热编码
- 连续值离散化处理
- 缺失值的处理
- 异常值的检测和处理
- 分割数据集

以上的算法需要时使用，处理方面的一般使用numpy和pandas就可以了，没啥困难的，这些或许能让你的模型表现更加出色。

3.3 特征工程

在数据挖掘的过程中十分重要，有空也学习学习；知识都是相互贯通的。



3.4 评估模型

不一样的任务需要的评估模型是不一样的。

- recall
- precise
- accuracy
- f1
- ...

3.5 可视化

类似效果如下。使用matplotlib就可以啦~!

模型使用 (10)

上边的所有基础都搞定了之后，接下来就是检验自己的模型正确与否或者说检验效果了。

此处分数，用kaggle上显示的分数（是0-1的小数）乘10，就是你此处的分数

1. 这里提供了 [线性回归数据集](#) 进行检验你自己的模型是否能用, 此题[提交网址](#)
2. 难度二 提供的 [多分类数据集](#), 此题[提交网址](#)

附加题： (10)

有空可以做，分数不大，时间不足就选择放弃也是可以的。课内重要！！！不要挂科！！！

如果有python基础不好的师弟师妹，上边的内容做不完也不要紧，考核看的是态度以及学习能力；学完基础就可以进行爬虫任务：

- [豆瓣电影top250](#)

◦ 要求：

- 每部电影的排名、名称、导演、豆瓣评分、语言、上映日期 (3)

No.1 豆瓣电影Top250

肖申克的救赎 The Shawshank Redemption (1994)



导演: 弗兰克·德拉邦特

编剧: 弗兰克·德拉邦特 / 斯蒂芬·金

主演: 蒂姆·罗宾斯 / 摩根·弗里曼 / 鲍勃·冈顿 / 威廉姆·赛德勒 / 克兰西·布朗 / 吉尔·贝罗斯 / 马克·罗斯顿 / 詹姆斯·惠特摩 / 杰弗里·德曼 / 拉里·布兰登伯格 / 尼尔·吉恩托利 / 布赖恩·利比 / 大卫·普罗瓦尔 / 约瑟夫·劳格诺 / 祖德·塞克利拉 / 保罗·麦克兰尼 / 芮妮·布莱恩 / 阿方索·弗里曼 / V.J.福斯特 / 弗兰克·梅德拉诺 / 马克·迈尔斯 / 尼尔·萨默斯 / 耐德·巴拉米 / 布赖恩·戴拉特 / 唐·麦克马纳斯

类型: 剧情 / 犯罪

制片国家/地区: 美国

语言: 英语

上映日期: 1994-09-10(多伦多电影节) / 1994-10-14(美国)

片长: 142分钟

又名: 月黑高飞(港) / 刺激1995(台) / 地狱诺言 / 铁窗岁月 / 肖申克的救赎

IMDb: tt0111161

豆瓣评分

9.7  2565874人评价

5星 85.7%

4星 12.9%

3星 1.2%

2星 0.1%

1星 0.1%

好于 99% 剧情片

好于 99% 犯罪片

- 每部电影的评论记得分类（好评、中评、差评），每部电影需要128条评论，注意数量的分配，注意评论的质量（比如有一个评论是：“啧啧啧”，显然，字数可以是一个过滤条件）。（7）

评分 默认排序

☒ 全部 ☐ 好评 98% ☐ 一般 1% ☐ 差评 1%

 犀牛 看过 ★★★★★ 2005-10-28 00:28:07 19088 有用
当年的奥斯卡颁奖礼上，被如日中天的《阿甘正传》掩盖了它的光彩，而随着时间的推移，这部电影在越来越多的人们心中的地位已超越了《阿甘》。每当现实令我疲惫得产生无力感，翻出这张碟，就重获力量。毫无疑问，本片位列男人必看的电影前三名！回顾那一段经典台词：“有的人的羽翼是如此光辉，即使世界上最黑暗的牢狱，也无法长久地将他围困！”

 文泽尔 看过 ★★★★★ 2008-01-14 01:53:08 3506 有用
人的生命不过是从一个洞穴通往另一个世界..然后在那个世界的雨中继续颤抖.i hope

 kingfish 看过 ★★★★★ 2006-03-22 12:38:09 32022 有用
不需要女主角的好电影

 如小果 看过 ★★★★★ 2008-02-27 21:43:23 20682 有用
恐惧让你沦为囚犯，希望你重获自由。——《肖申克的救赎》

 Eve|Classified 看过 ★★★★★ 2008-05-09 23:15:34 9819 有用
“这是一部男人必看的电影。”人人都这么说。但单纯从性别区分，就会让这电影变狭隘。《肖申克的救赎》

答辩（100）

提交文档格式：如果不对，我看心情扣分哦~~不懂的话可以随时问我。

```
Submission_your_name/  
...data/  
.....train.csv  
.....test.csv  
.....submission.csv  
.....etc.  
...code  
...doc  
etc.
```

#

文档要求！！！往详细、清晰、干净的学术论文的感觉去靠（45）

课内重要！！！不要挂科！！！！