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

- sotfmax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

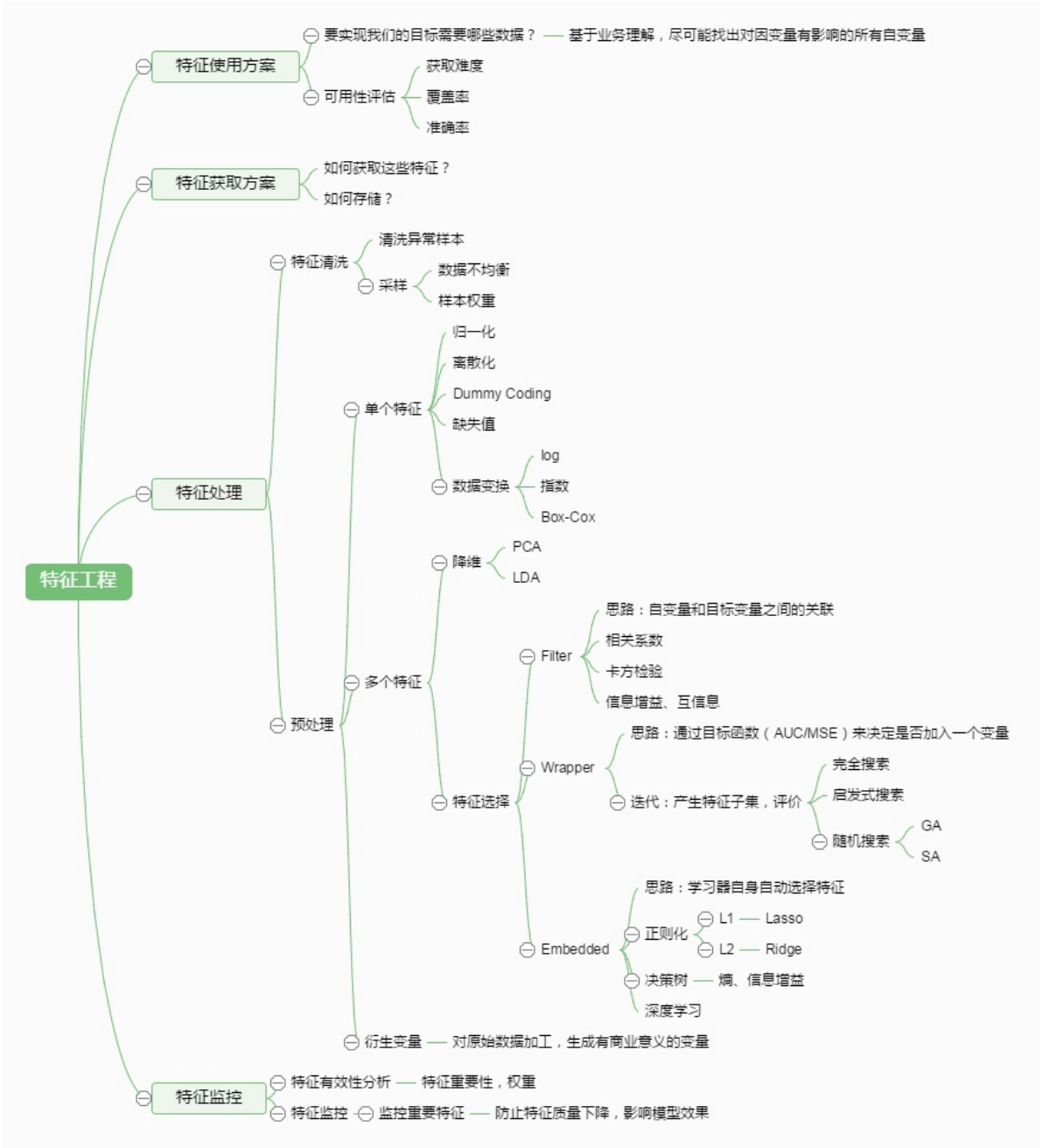
学习以下几种数据处理方法：

- 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）

        - 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）

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！！！