毕业设计

吴沛豪

目录

数据挖掘很难先选定题目再去找数据，所以我的大致想法思路如下：

1 获取数据

数据要有明确结局有特征变量的大样本数据，结局如果是一项连续性的指标最好，这样的话操作性更强，因为我们可以从线性模型开始向外扩展。但如果是分类变量也是可以进行，依据具体情况而定。

我现在的目标数据库是 nhanse 数据库。**美国国家健康与营养调查**（NHANES, National Health and Nutrition Examination Survey）是一项基于人群的横断面调查，旨在收集有关美国家庭人口健康和营养的信息。项目每年调查一个全国代表性的样本，约 5000 人，分为访谈和体检数据两大

部分。访谈部分包括人**口统计学、社会经济学、饮食和健康相关**问题；体检部分包括**基础医疗信息，包括血压，测听检查、口腔健康、握力等等以及大量的实验室检测数据及部分放射科数据**。并且变量足够多，可以为我们引入稀疏性的讨论。

- 暂定使用 2017-2018 年数据
- 直接下载 XPT 文件，用 R 语言进行读取
- 用"nhanseA" 包中的函数下载（推荐，但网络连接会出问题）
- nhanseTranslate 翻译变量

2 最终目标/想要做什么

1. 通过模型比较，选用合适模型，讨论模型的稳定性、准确性、可解释性；
2. 解释模型/预测结果；
3. 特征变量重要性分析，找出主要因素。

3 建模方式

1. R 语言：最近几年流行使用 Tidymodels，使用 workflow 建模，大大减轻工作量。R 语言的作图系统成熟美观。19 年起，R 语言支持调用 tensorflow，深度学习的包日益发展强化。
2. Python：python 在机器学习方面起步早，深受大众喜爱。

4 变量处理

1. 当变量很多的时候，我们应该适当择选变量，比如通过单元回归、lasso 等方法剔除一些变量。

2. 分析之前，有必要进行计算相关性矩阵，做出对应气泡图，排除共线性。
3. 连续性变量应该进行数据变换，log、exp、1/x、sqrt、box-cox 等等。
4. 进入模型之前，最好进行正则化处理，模型计算时别忘了逆正则化。

5 模型建立

1. 多元回归：线性既是优点（可解释性高，探索交互等），又是缺点（原始分布非高斯）。
2. lasso：稀疏性，简化特征
3. GLMs：泊松对数等
4. GAMs：样条回归，非线性
5. 决策树：分类，CART
6. 朴素贝叶斯分类器：分类
7. K-最近邻：分类/回归
8. 随机森林：经典
9. 神经网络：调用 tensorflow+keras，比如 RNN

6 补充

1. 交叉验证确定模型参数
2. bootstrap 进行参数估计
3. 以 MSE、MAE 等描述连续性结局指标的预测性能，以混淆矩阵及其相关的指标描述分类变量结局的预测性能。