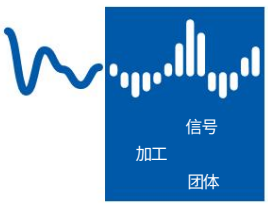
数据科学 1 - 最终项目

TECHNISCHE UNIVERSITÄT DARMSTADT

2022 年 4 月 25 日更新

1 简介

这个最终项目的目标是在整个待解决的实际问题中展示您在数据科学方面获得的所有知识。这包括适当的问题制定和动机，确定合适的算法并展示您的结果。这也是项目研讨会数据科学 II 的完美准备。



博士。 C. Debes
M.Sc Pertami Kunz

2 个任务

- 选择您感兴趣的数据集。
- 在两者之间进行选择：

1.单独的假设检验和回归/分类方法：
分别确定您要测试的假设：您要使用某些/所有其他变量预测的响应变量是什么。

2.回归/分类模型的假设检验：决定你想要预测一些/所有其他变量的响应变量，测试多个模型并决定哪个模型在统计上是最好的（通过假设检验，而不是直接比较性能数字，作为数字可能是偶然出现的）。

3 报告清单

您的最终报告应包括以下项目（带有 X 的项目）。它们不需要总是按照这个给定的顺序排列。每个项目不需要位于单独的部分/小节中。
重要的一点是让你的报告简洁流畅。

- 介绍

X 目标。
介绍报告的目标。以非技术人员可以理解的方式对其进行表述。描述由此产生的商业/社会或其他利益。

X 数据集简介。

X 问题陈述。

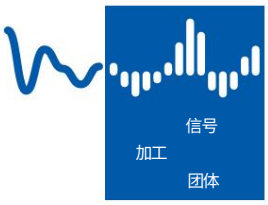
X 一些预处理。
条目是否需要清理、过滤以适合您感兴趣的问题？

- 探索性数据分析 X 一些有用的图

表和/或表格。
导致您的问题和/或假设的图表和表格。

X 一些特征提取/工程 有什么要估算的吗？归一化？可以衍生出新的特征吗？

- 假设检验



博士。 C. Debes
M.Sc Pertami Kunz

X 陈述你的假设

X 描述你的测试统计,为什么选择它。

X 其余的假设检验步骤,直到得出结论。
您需要哪些值来得出结论:置信区间/显着性水平/临界值/p 值?您不需要提及所有这些值,例如,如果您从程序中获得 p 值,那么将其与您选择的显着性水平进行比较就足够了。

· 一个/一些回归或分类模型。

X 您是如何拆分数据的?

X 模型的数学表达式。

X 为什么要探索这个/这些模型?

X 你是如何决定超参数的(如果有的话)?

X 性能(在训练集和测试集上)

X 结论

X 参考资料 重要:
引用数据集。如果您在报告中提到了一个 python 库,那么请同时引用它的参考。

4 有用的链接

以下是一些可供选择数据集的链接(您可以从其他来源中自由选择):

1. UCI 机器学习库
2. Kaggle 数据集

例子:

1.亚马逊图书EDA和假设检验

2.泰坦尼克号的统计分析和机器学习工作流程3.更多假设检验示例

4.回归示例

5.分类示例

机器学习算法假设检验参考:

1.比较监督分类学习的近似统计检验算法

2.多个数据集上分类器的统计比较

3.泛化误差的推断