2022-2023 秋季《机器学习方法基础》课程第 1 次作业

提交时间 10 月 9 日

1. 现有某生鲜电商平台于 2019/03/27 至 2019/12/27 的日销量序列数据，其中生鲜产品 A 的销量训练数据如文件 Tseq_Sales.csv，训练数据用 0 标记，测试数据用 1 标记（具体数值已经经过脱敏处理），尝试建立不同的销量随时间变动的预测模型（至少三种方法），

   A. 给出模型假设，提出模型表示；

   B. 选择合适的数据对模型涉及的参数进行估计；

   C. 给出估计的检验结果，讨论系数估计的实际含义；

   D. 给出预测测试误差的方法和结果；

   模仿《ISLR2》P34 图 2.11，画出不同模型的拟合序列图、模型评价（如 MSE）随模型柔性（flexibility）变化的分析图像，并据此讨论所选模型中哪一种更适合进行预测建模。

2. 数据 inks5_CLASSdataset.txt 数据中是 5 种墨迹鉴定数据，每种墨迹按照三种主要的墨迹化学成分结构比例测量后，记为 x,y,z，可以根据不同比例结构对墨迹的生产厂家给予鉴定，数据中的 Name 中的编号代表厂家，一共 5 个厂家，分别记为 ink_1,ink_2,ink_3,ink_4,ink_5，每个厂家又分不同包装型号，厂家编号无小数点为瓶装，加小数点为简装，比如 ink_1.1 代表的是简装包装，由于简装包装可能会加速墨汁成分变异，加大了鉴定的难度。而鉴定任务无法获知产品出厂使用的是简装包装还是瓶装包装。本项目的研究目标是，选定一种二分类模型根据数据找到能够对 5 种生产厂家进行区分的功能：

   A.训练数据是每个厂家独立提供的 10 个采样标本，用 piece 标记样品标号。通过训练数据（标识为 Itemtype=TRAIN）根据 x,y,z 给出能将不同墨迹区分的模型，估计模型参数；

   B.比较使用不同的训练数据时（只用），在估计模型的估计结果上有怎样的不同？训练误差有怎样的不同？

   C.B 中选用不同的训练数据建立了的模型在不同的测试数据（Itemtype=TRAIN）上的测试误差如何？

   C.选择 B 中表现较好的模型给出每一种墨迹的训练误差和测试误差计算结果；

   D.给出对每一种墨迹进行鉴别的微训练误差和微测试误差的计算结果；

   E.给出整体实验的微训练误差和微测试误差微训练误差和微测试误差。

   F.对以上实验流程给出讨论。

   G.参考教材《Statistical Analysis in Forensic Science》（已上传至资料区）

   D.10 部分给出的对两组数据绘制直方图等共 4 种图形的代码，从墨迹鉴定数据中选择两种或两种以上墨迹绘制这 4 种图形。

格式要求：

1. 最终请提交程序及报告文档至课堂派，程序以 R-Markdown 或 Python Notebook 形式提交， 要求有必要的注释，并对程序运行结果做出适当解释。
2. 报告文档以 latex 书写，请在附件中下载 tex 模板，按模板格式书写报告并输出 pdf 作为提交的报告文档。