数据科学导论 2022秋 第三次作业

第一题 - 机器学习

TA最近学习了机器学习的相关内容，然后就想试试拳脚。正巧抖腿公司派TA去管理公司的邮件系统，于是TA决定在邮件系统里增加垃圾邮件检测这一功能。可TA学艺不精，请你帮帮TA。

1. TA不知道该用哪种学习范式(监督学习、无监督学习、强化学习), 请你先告诉TA这三种学习范式的异同。

2. 说说在**每种学习范式**下，解决垃圾邮件检测(spam detection)的**具体过程**。

3. 你觉得用哪种学习范式最好，为什么?

第二题 - 神经网络

TA觉得深度学习很神奇，但有人告诉TA，深度学习就像炼丹，要训练好深度学习的模型是一件很玄的事情，因为深度学习有很多超参数要调。TA不相信，决定自己动手试试。你也一起来试试吧。

1. 请在Neural Network Playground里**好好玩耍**：了解每个术语、颜色、按钮的意思, 对于不同的任务和数据集，尝试训练你自己的神经网络。

2. 对于最后一个分类任务(双螺旋线)，把Ratio of training to test data设为**70%**, Noise设为**40**，然后在此数据集下，去调整超参数并训练你的模型，使你的模型尽可能快地收敛，同时取得良好的分类效果。提交训练结束时的截图。

3. 最后谈谈你在玩耍过程中有哪些发现，比如怎样选择学习率、不同超参数的调整优先级是怎样的。

（可选题） 机器学习应用

TA最近学习了一些机器学习算法，对于泰坦尼克号上乘客的生存情况分析再次有了新的思路。他好奇轮船上的最终生存与否和乘客本身信息的关系，是不是能使用机器学习方法找出来，你能帮帮TA吗?

请观察titanic文件夹中的train.csv和readme.md，了解乘客都有什么基本信息（与作业1/2的可选题相同），思考怎么抽象成为机器学习问题。

阅读材料 Step-5:-Model-Data 部分（可以略过代码，只看文字、图片和表格）。简述他使用了哪些机器学习模型？怎么确定该机器学习任务的baseline准确率？怎么做模型选择?

作业要求

- **严禁抄袭**!
- 上传**单个pdf文件**至网络学堂。
- 推荐使用 markdown 或 latex 写作业。