COMP9417项目

2022年6月14日

宗旨

本作业的学习目标。

- 一个自我选择的任务，以扩展课程材料的各个方面
- 涉及机器学习问题的实际方面，即
  - 实施或修改算法和/或
  - 对数据集进行算法的实验评估
- 在激励、记录和总结特定任务的工作中发挥书面交流技能

提交

这项作业的交卷有两部分。

- 包含对数据集做一些有趣的事情的程序代码的文件和/或在数据集上运行程序的结果。
  - 文件的压缩档案
  - 任何编程语言都可以使用
  - 必须合并成一个tar或zip档案。
- 关于你所做事情的报告。
  - 必须是一个PDF格式的单一文件。
  - 必须包括**所有**团队成员的姓名和ZID

注意：团队中**只有一个**人提交作业的两个部分。

注意：**请确保提交一个包含你的报告的单一PDF文件，以及一个包含其他文件的单一tar/zip文件。不要把你的PDF文件**

提交

和你的**tar**文件结合起来。

标记

总分：可得30分

- **第一部分：**[15分]
  - 8分：解决了题目中描述的基本问题
  - 7分：额外的功能，或1个人解决了大部分或全部的问题。> 1个人的问题*[由评分者决定]*
- **第二部分：**[10分]
  - 6分：描述问题和你的解决方案
  - 4分：对结果有良好的表述和交流
- **成绩。**[5分] 由评分员决定--基本上是根据小组人数、题目难度、分析深度等因素，根据作品的印象如何来评分。

第一部分

分数将通过以下方式获得。

- 通过将问题分解为若干子部分，证明有良好的设计或规划。
- 严格的结果收集
- 在工作过程中使用评论和笔记来记录所做的决定和理由
- 激发对项目的选择和你的方法（例如，为什么这个项目很有趣，以前做过吗？）

第一部分

分数将通过以下方式丢失。

- 程序无法编译或运行
- 缺少的结果文件
- 没有关于所提交文件内容的明确信息（例如在README中）。
- 抄袭的证据（包括提交的文件与网上现有的实施方案非常相似）。这包括回收为其他课程所做的工作，例如，COMP9444。

第二部分

分数将通过以下方式获得。

- 彻底测试一个想法的证据
- 使用表格、图表等对关键结果进行良好的表述和总结。
- 简单、清晰和相关的解释
- 格式良好、条理清晰、经过拼写检查和语法检查的文件

分数将通过以下方式丢失。

- 适当的长度（目标是3+2.5x的长度，其中*x*是小组成员的数量。额外的数字、表格等可以放在一个*合理*长度的附录中）。

  这不是一个硬性规定

  只要长度合理，较长的报告也是可以的（即，内容对报告至关重要，使用你的最佳判断）。
- 赘言、漫无边际或胡言乱语，以不必要的方式填补空间
- 错误或不一致的表述，如
  - 对算法或其属性的描述不正确
  - 为任务选择不好的算法
  - 评价中的错误，如没有使用独立的测试集或交叉验证（如果需要）。
  - 没有基于你的实验结果或参考资料的声明或结论
  - 不正确或不适当地使用统计测试

第二部分 剽窃的证据

集团配置

每个团队必须配置有1-5名目前在读的学生

- 团队可以由不同辅导班的学生组成，小组可以由PG和UG学生组成。
- 较大的团队要做更多的事情（成绩等级会受此影响）。
- 团队应提交每个成员完成的工作总结。如果缺少，我们将假定所有成员的贡献相同。
- 如果需要，你可以使用Moodle小组项目（寻找小组）论坛来寻找小组成员。
- **将你的小组加入Moodle上的 "小组项目-成员选择 "对象。这样做的截止日期是2022年7月1日星期五。你只有在得到其他组员的允许后才能加入一个小组!**

成员的贡献

- 所有小组成员对提交的任何工作都应做出同等贡献。
- 如果小组认为一个或多个学生没有做出足够的贡献，我们将采取以下措施
  重新分配相应的分数。
- 一些好的建议。在整个项目中保持你的贡献记录。保存与其他小组成员的所有沟通记录（电子邮件/聊天）等。在小组出现争议时，我们会要求所有小组成员提供有关贡献的证据。如果不能提供证据，意味着所有小组成员将获得相同的成绩。

报告结构

对项目报告的格式给出一套非常严格的准则是困难的，因为不同的项目是非常不同的。

然而，需要记住的一些事情是。

- **长度。**保持简明。在代码中包含一个README文件，这样你就不必在报告中加入这类信息。
- **介绍。**你必须解释你所解决的问题，为解决这个问题所采取的基本方法，为什么选择这个方法，以及这个方法在机器学习方面的任何重要方面。
- **实施。**如果你的工作主要是实施，请重点说明。否则，请简要描述你做了什么。

报告结构

- **实验。**所有的方法都必须在一些数据上进行测试，所以应该包括这些结果。此外，如果这是一个主要的焦点，你需要解释所做的工作和所取得的成就，例如关于学习任务的设置、评价的选择等等。详细的统计分析可能超出了项目的范围，所以不要包括这些，除非你已经非常熟悉这种事情。
- **参考文献。**应该有所使用的算法或其他方面的工作。
- **附录。**如果你有大量的实验结果，应该使用。然而，可以考虑绘制图表或使用其他可视化手段，如直方图，以简明地总结大量的结果。

最后期限

2022年8月1日 星期一 23:59:59

主题。主题0--提出你自己的建议

本课题的目的是提出一个机器学习问题，提供数据集，并实施一种方法来解决这个问题。这通常来自于你以前有一些经验的工作或研究领域。

- 它必须涉及一些机器学习的实施的实际工作
- 你必须向课程管理员发送一封电子邮件（使用班级账户），说明你的计划（几段就够了），在你开始之前，需要通过电子邮件回复批准。
- 它不能涉及到重复投资，即成为另一门课程的项目的一部分，或者对于研究型研究生来说，它必须包括一个声明，大意是它不是为论文计划的主要工作的一部分（尽管它可以是相关的）。

主题。主题0--提出你自己的建议

- 如果你选择做课题0，提出项目的截止日期是7月11日。

专题。主题1--机器学习论文

本课题的目标是选择一篇期刊或会议论文，总结其结论，并在新的或模拟的数据集上实现所提出的算法。

- 好的论文来源是。神经IPS，ICML，JMLR，JAIR，ICLR，或ArXiv。
- 你也可以选择一系列的论文，并比较解决同一问题的各种方法。
- 在你开始做这个项目之前，也要给课程管理员发邮件。如果你选择做课题1，提出项目的截止日期是7月9日。

主题。主题2 - 竞争与挑战 - Kaggle

- Kaggle竞赛在这里举办。你只能参加标有**特色**或**研究**或**分析**的比赛。你可以从正在进行的比赛或已完成的比赛中选择一个来进行工作。
- 仔细评估你需要多少时间来了解竞争要求、熟悉数据和运行你计划使用的算法。
- 对于现场比赛，您可以在提交时将您的作品在排行榜上的排名包括在内。但是请注意，你的成绩不会完全由你在排行榜上的排名决定。当然，能在比赛中取得好成绩是很好的，但我们主要是根据你的方法和最终报告来给你评分。
- 你不需要管理员批准这个题目。你**必须**在报告的第一页包括一个竞赛的链接。如果不这样做，将被立即扣掉2分。

专题。其他考虑因素

- 不要选择需要大量数据处理的项目，或 "创造 "一个数据集，因为我们在这个课程中主要对机器学习感兴趣，而不是数据清理。当然，大多数任务都需要进行*一些*预处理。
- 较大的小组预计会取得更多的成绩，在分配成绩和额外功能的分数时，将考虑到小组的规模。
- 选择一个你感兴趣的主题，但在涉及时间要求和项目难度时要务实。
- 在选择比赛/数据集/模型时，请使用常识。如果你选择了一个非常简单的任务，就不要指望获得好成绩。
- 在使用先进的机器学习技术之前，总是使用简单的基线，如

专题。其他考虑因素

决策树或逻辑回归。

例子。项目报告

- 每个项目都是不同的。
- 然而，如果你遵循上面的准则，你的小组应该能够做出一份好的报告。
- 我们提供了两份最近的报告作为例子。
- 第一个是去年的Kaggle比赛。
- 第二个是几年前关于强化学习的应用（这个话题已经不再有了，但它可以为你自己的原创话题提供一些想法）。
- **这些都可以在课程的Moodle页面的 "项目实例 "文件夹下的项目对象中找到。**

例子。主题1--机器学习论文

你不必选择下面的任何建议，它们只是为了给你一个想法。它们的范围从相当的理论到实际。

- 全面审视评价。
- 提议采用一种新的技术，称为阶段性回归。
- 神经网络近似的数学分析。
- 一个用于张量处理的新Python库。
- 一个用于检测离群点的重要任务的Python库。

如果你想做这个题目，请与你的小组仔细讨论，在你做出选择之前，也请在上面的题目1幻灯片中搜索资料来源，以获得更多

例子。主题1--机器学习论文

的选择。

例子。主题2 - 竞争与挑战

Kaggle是一个机器学习问题和数据集的首选网站。这些都为获得应用机器学习的基本技能提供了一个很好的机会。你不需要选择下面的任何建议，它们只是给你一个想法。

注意：Kaggle上的一些数据集很大--你可以为你的项目抽出一个数据子集，只是要确保在你的小组报告中详细说明如何进行抽样。

一些典型的预测任务。

- 欺诈检测。
- 恶意软件的预测。
- 每小时降雨量。

例子。主题2--竞争与挑战（续）

一些图像分析任务。

- 从卫星图像中对云组织模式进行分类。
- 全球小麦检测。
- 木薯叶病分类。
- 识别露脊鲸。

一些自然语言处理（NLP）任务。

- 帮助结束代词决议中的性别偏见。
- 英语文本规范化的挑战。

如果你想做这个题目，请和你的小组仔细讨论，也可以在选择之前搜索**Kaggle**，了解更多的选择。