COMP3702 人工智能（2022 年第 2 学期）
作业 3:HexBot 强化学习

关键信息：

·截止时间:10 月 27 日星期四下午 4 点

- 该作业将评估您在开发解决强化学习的算法方面的技能问题。
- 作业 3 占您最终成绩的 20%。
- 该作业由两部分组成:(1) 编程和(2) 报告。
- 这是一项个人作业。
- 代码和报告都将通过 Gradescope (https://www.gradescope.com/) 提交。
- 您的程序（第 1 部分,60/100)将使用 Gradescope 代码自动评分器进行评分,使用与https://gitlab.com/3702-2022/a3-support 提供的支持代码中的测试用例类似的测试用例。
- 您的报告（第 2 部分,40/100)应符合提供的模板,采用 .pdf 格式并根据格式 a3-COMP3702-[SID].pdf 命名,其中 SID 是您的学生证。报告将由教学团队评分。

HexBot人工智能环境

您的任务是开发一种用于自动控制 HexBot 的强化学习算法,HexBot 是一种在六边形环境中运行的多功能机器人。对于此版本的任务,HexBot 必须导航到指定目标之一,同时尽可能降低罚分。不需要考虑小部件。为了帮助您完成这项任务,我们为 HexBot 机器人环境提供了一个模拟器和可视化,您将与之交互以开发您的解决方案。

对于 A3,HexGrid 环境具有不确定的动作结果,其概率未知,每个测试用例都是随机的。每个可用操作的成本以及与障碍物和危险碰撞的惩罚也是未知的,并且对于每个测试用例都是随机的。从 A2 对本文档的添加和修改以洋红色文本显示。但是,A2 环境描述的某些方面也已被删除。建议您阅读整个部分,而不是浏览粉红色的变化。

六边形网格环境由六
边形网格表示。十六进制网格的每个单元格都由（行、列)坐标索引。十六进制网格的索引从上到下,从左到右。也就是说,左上角的坐标为 (0, 0),右下角的坐标为(nrows − 1, ncols − 1)。偶数列（从零开始)位于行的上半部分,奇数列位于行的下半部分。图 1 显示了一个示例。

如果六角网格中的两个单元共享一条边,则认为它们是相邻的。对于每个非边界单元格,有 6 个相邻单元格。

机器人
HexBot 机器人占据六角网格中的一个单元格。在可视化中,机器人由标有字符“R”的单元格表示。标有“*”的单元格一侧代表机器人的正面。机器人的状态由其（行、列)坐标及其方向（即其前侧指向的方向)定义。

机器人有 4 个可用的标称动作：

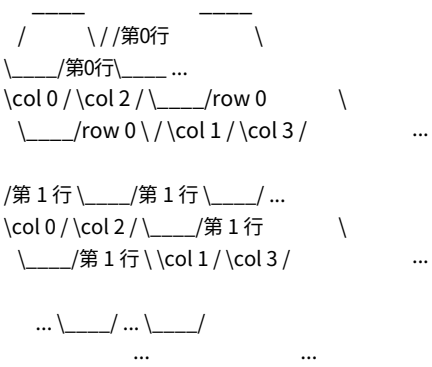

图 1:显示行和列索引顺序的示例六边形网格

- 前进→ 向机器人前方方向移动到相邻单元格（保持不变）
  方向）
- 反向 → 以与机器人前方相反的方向移动到相邻的单元格（保持
  相同的方向）
- 向左旋转 → 向左旋转（相对于机器人正面,即逆时针方向)60 度（停留在
  同一个单元格）
- 向右旋转 → 向右旋转（即顺时针)60 度（留在同一个单元格中）

机器人每次选择一个动作,都有一个固定的概率（根据每个动作的种子随机设置）
测试用例)让机器人在执行选定的标称动作之前以顺时针或逆时针方向“漂移”60度（每个漂移方向的概率不同）。漂移发生的概率

取决于选择的标称动作,有些动作更可能导致漂移。顺时针漂移
和 CCW 是互斥事件。

此外,机器人有一个固定的概率（也根据每个测试用例的种子随机设置）
到“双重移动”,即执行两次名义上的选定动作。双重移动发生的概率
取决于选择的操作。双重运动可能与漂移（CW 或 CCW)同时发生。

每次行动后获得的奖励是每次行动获得的奖励中最小/最负的
标称动作和任何附加（漂移/双重移动)动作。

障碍

六角网格中的一些单元格是障碍物。在可视化中,这些单元格填充有字符“X”。任何
导致机器人或 Widget 的任何部分进入障碍物单元的动作导致碰撞,导致
代理接收一个负的障碍物碰撞惩罚作为奖励。这个奖励代替了移动成本
否则代理人会招致的。六边形网格的外边界表现相同
方式作为障碍。

此外,环境现在包含一种额外的障碍物类型,称为“危险”。危害表现在
与障碍物的方式相同,但是当发生碰撞时,会收到不同的（更大的)惩罚作为奖励。
因此,避免与危险碰撞比避免与障碍物碰撞更重要。
危险用“！！！”表示在可视化中。

目标

十六进制网格包含许多“目标”单元格。在可视化中,这些单元格标有“tgt”。为了
要考虑解决的 HexBot 环境,其中一个目标单元格必须被 HexBot 占用。
环境可能包含多个目标。

HexBot作为强化学习问题

在本作业中,您将编写一个程序组件来玩 HexBot,目标是找到
使用各种强化学习算法的问题的高质量解决方案。本次作业将

测试你为实际问题定义强化学习算法的技能,以及对关键算法特征和参数的理解。

提供给您什么

我们将仅提供 Python 中的支持代码,形式为:

1.一个代表HexBot游戏地图的类和一些辅助函数
2. 一种解析器方法来获取输入文件（测试用例)并将其转换为 HexBot 映射
3. 一名测试员
4. 测试和评估您的解决方案的测试用例
5. 一个脚本,让您以交互方式玩 HexBot
6.解决方案文件模板

支持代码可在以下位置找到： https://gitlab.com/3702-2022/a3-support。有关更多详细信息,请参阅 README.md。代码的自动评分将通过 Gradescope 完成,因此您可以测试您的提交并根据此反馈继续改进它 - 强烈建议您使用此反馈。

你的分配任务

你的任务是开发两种强化学习算法来计算代理的路径（一系列动作）,并编写一份关于你的算法性能的报告。您将根据提交的计划（第 1 部分,60%)和报告（第 2 部分,40%)对您进行评分。这些百分比将被缩放到该评估项目的 20% 课程权重。

提供的支持代码提供了一个生成的 HexBot 环境,您的任务是提交实现以下两种强化学习算法的代码:

1. Q-学习
2. 非典型肺炎

个别测试用例指定将应用哪种策略（Q-learning/SARSA）。请注意,有 3 个十六进制网格,每个算法都在单独的测试中应用（使 6 个测试总计超过 3 个网格的 2 个算法）。

一旦您实施并测试了上述算法,您将完成“第 2 部分 - 报告”部分中列出的问题并将报告提交给 Gradescope。

下面给出了编程和报告部分所需的更多详细信息。

第 1 部分 编程任务

您的程序将使用 Gradescope 自动评分器进行评分,使用类似于https://gitlab.com/3702-2022/a3-support 提供的支持代码中的测试用例。

与测试用例和自动评分器的交互

我们现在为您提供一些详细信息,解释您的代码将如何与测试用例和自动评分器交互（特别感谢尼克柯林斯再次努力使这项工作无缝地工作！）。

首先,请注意环境类的作业 3 版本（在 environment.py 中)与以前的作业不同,因为转换和奖励函数现在是随机的且未知的

给代理。动作结果概率（双移动、顺时针漂移和逆时针漂移）和成本/惩罚（动作成本、碰撞惩罚和危险惩罚）基于文件名的种子在某个固定范围内随机化，并且都存储在私有变量。您的代理不知道这些值，因此必须与环境交互以确定最佳策略。

使用提供的 solution.py 模板文件实现您的解决方案。您需要填写以下方法存根：

- __在里面__（）
- q_learn_train()
- q_learn_select_action（状态）
- sarsa_train()
- sarsa_select_action（状态）

如果需要，您可以添加到 init 方法中，如果您愿意，可以添加其他辅助方法和类（在 solution.py 或您创建的文件中）。为确保自动评分器正确处理您的代码，您应该避免在上述方法的实现中使用任何 try-except 块（因为这会干扰我们的超时处理）。自动评分器不允许您上传自己的 environment.py 副本。

有关详细信息，请参阅 solution.py 中的文档。

编程组件的评分标准（总分：60/100）

对于标记，我们将使用 6 个测试用例来评估您的解决方案。每个测试用例都使用指定为求解器类型的算法。每个测试用例得分为 10.0 分，分为以下四个类别：

- 代理成功达到目标
- 总培训奖励
- 总评价奖励
- 时间流逝

每个测试用例的 10 分平均分配给四个类别（即每个测试用例的每个类别分配 2.5 分）。

- 每个测试用例都有总训练奖励、总评估奖励和经过时间的目标。
- 当您的计划在每个类别中匹配或超过目标时，将获得最高分
- 部分分数可用于总训练奖励、总评估奖励和经过时间。
- 测试用例的总分是每个类别分数的加权总和
- 总代码标记是每个测试用例的标记总和

第 2 部分 报告

该报告测试您对强化学习的理解以及您在代码中使用的方法，并贡献了 40/100 的分配分数。

问题 1. Q-learning 与马尔可夫决策过程的价值迭代算法密切相关。

a) （5 分）描述 Q 学习和价值迭代之间的两个关键相似之处。

b) (5 分)给出 Q-learning 和 Value Iteration 之间的一个关键区别。

对于问题 2、3 和 4,考虑用于 Q-learning 的测试用例 ex3.txt 和等效的 SARSA 测试用例 ex4.txt。

问题2。

a) (5 分) 解释离策略和在策略强化学习算法之间的区别。
解释这种差异在 Q-learning 和 SARSA 算法中的体现。

b) (5 分)off-policy 和on-policy 算法之间的差异如何影响 Q-learning 和 SARSA 解决测试用例 ex3.txt 和 ex4.txt 的方式?如果您无法解决这些测试用例,则根据测试用例文件中描述的设置,参考您认为会发生的情况来回答就足够了。

对于问题 3 和 4,您被要求绘制每集的解决方案质量,由您的学习代理收到的 50 步移动平均奖励给出。在时间步 t,50 步移动平均奖励是您的学习代理在 [t - 50, t] 集（包括集重新开始)中获得的平均奖励。如果 Q 值意味着质量较差的策略,则该值将很低。如果 Q 值对应于高价值策略,则 50 步移动平均奖励会更高。我们在这里使用移动平均线,因为奖励只是偶尔收到,并且在转换和探索策略中存在随机性来源。

问题 3。

a) (5 分)绘制（全部在一个图上)通过 Q-learning 在 testcase ex3.txt 中学习到的策略的质量与三个不同的 learning_rate 固定值（在讲义中称为 α 和在许多文本和在线教程中）,由 50 步移动平均奖励给出（即对于这个问题,不要随时间调整 α,而是在整个学习过程中保持相同的值）。您的绘图应显示解决方案质量,直到性能稳定的情节计数。根据所使用的学习率,这可能需要大量的情节（例如> 50,000）。请注意,策略质量可能仍然存在噪声,但算法的性能将停止增加,其平均质量将趋于平稳。您的绘图应包括轴标签和图例。

b) (5 分)讨论改变 learning_rate 的效果。您应该参考 Q 学习算法来支持您的讨论。如果您能够在 Q3a 中生成绘图,您也可以在讨论中参考这一点。

问题 4。

a) (5 分)分别在测试用例 ex3.txt 和 ex4.txt 中,根据 50 步移动平均奖励给出的 Q 学习和 SARSA 下的集数（在单个图上)绘制学习策略的质量。您的绘图应显示解决方案质量,直到两个算法的性能都稳定的情节计数。您的绘图应包括轴标签和图例。

b) (5 分) 参考你的情节,比较两种算法的学习轨迹,以及它们最终的解决方案质量。讨论 Q-learning 和 SARSA 的解决方案质量在他们学习解决测试用例时的变化方式,无论是在他们学习的过程中还是在他们稳定下来之后。

学术不端行为

大学将学术不端行为定义为“一系列旨在让学生比同龄人获得不公平和不劳而获的优势的不道德行为”。 UQ 非常重视学术不端行为,任何可疑案件都将通过大学的标准政策进行调查(https://ppl.app.uq.edu.au/content/3.60.04-student-integrity-and-misconduct)。如果你被判有罪,你可能会被大学开除,没有任何奖励。

学生有责任确保您了解什么是学术不端行为,并确保您不违反规则。如果您不清楚需要什么,请询问。

学生还有责任采取合理的预防措施,以防止他人未经授权访问他/她的作品,无论以何种格式存储,无论是在评估之前还是之后。

在 COMP3702 作业的编码部分,您可以使用可公开访问的资源并提供教程解决方案,但您必须通过执行以下操作来引用或注明其来源:

- 您从公共来源获取的所有代码块都必须在您的
  代码。
- 还请在您的solution.py 文档字符串中包含一个引用列表,表明您使用的代码。

但是,在任何情况下,您都不得向任何其他学生展示您的代码或与任何其他学生共享您的代码。您不得将代码发布到公共论坛（包括 Ed Discussion)或将代码保存在可公开访问的存储库中（检查您的安全设置）。您不得查看或复制任何其他学生的代码。

所有提交的文件（代码和报告)都将接受电子抄袭检测,并对涉嫌抄袭或串通的学生提起不当行为诉讼。即使注释、变量名、格式等被修改,电子抄袭检测也可以检测代码结构中的相似性。如果您串通开发您的代码或回答您的报告问题,您将被抓住。

欲了解更多信息,请查阅以下大学网页:

- 关于学术诚信和不当行为的信息:
  - https://my.uq.edu.au/information-and-services/manage-my-program/student-integrity-and 行为/学术诚信和学生行为– http://ppl.app.uq.edu.au/
    content/3.60.04-student-integrity-and-misconduct
- 学生服务信息:
  - https://www.uq.edu.au/student-services/

迟交

学生不应该在最后一刻才准备作业,并且必须计划他们的工作量以满足广告或通知的截止日期。有效地管理时间是您的责任。

自动评分器最多可能需要一个小时才能对您的提交进行评分。您有责任确保足够早且足够频繁地上传代码,以便能够在截止日期之前解决自动评分器可能发现的任何问题。在截止日期之前提交非功能性代码,并且没有足够的时间来更新代码以响应自动评分器的反馈,这不是延迟提交而不受处罚的正当理由。

逾期处罚:如果在原始截止日期之后提交评估项目而没有获得批准的延期,则将适用逾期处罚,详见 COMP3702 电子课程简介 (ECP)。逾期罚款应为

- 评估项目的最大可能分数的 10% 将被扣除每个日历日（或其中的一部分）,最多七 (7) 天。 7天后,将不会获得任何分数

物品。一天被认为是从评估项目到期时间开始的 24 小时块。负号
不会被授予。

在特殊情况下,您可以提交延期申请。您可以在此处找到关于可接受的延期理由的指南 https://my.uq.edu.au/information-and-services/manage-my program/exams-and-assessment/applying-extension所有延期请求必须是在UQ上提交

在提交截止日期前至少 48 小时申请延期评估表格。