IML2024 学期项目：预测饱和蒸气压

DATA11002 机器学习导论（2024）

一、总结

学期项目 Kaggle 挑战赛邀请链接（请勿分享链接）：

- 常规 Kaggle 挑战赛：https://www.kaggle.com/competitions/iml24-term-project
- 高级 Kaggle 挑战赛：https://www.kaggle.com/competitions/iml24-adv

在本次学期项目中，你将基于大气测量数据集训练一个回归模型。为了完成项目，你需要提交以下内容：

- 向 Kaggle 竞赛提交，即将测试集的预测结果提交至课程的 Kaggle 页面。
- 在 Moodle 上提交项目报告的初稿，格式为单个 PDF 文件。
- 学期项目的展示。
- 在 Moodle 上提交最终报告，格式为单个 PDF 文件。

二、关于数据

学期项目基于 GeckoQ 数据集，该数据集包含 31,637 种由α-蒎烯、甲苯和癸烷氧化生成的大气相关分子的原子结构。GeckoQ 数据集旨在补充大气科学中的数据驱动研究。它提供了与气溶胶粒子生长和新粒子形成（NPF）相关的分子数据。与气溶胶粒子生长密切相关的一个关键分子属性是饱和蒸气压（pSat），它衡量分子凝结为液相的能力。具有低 pSat 的低挥发性有机化合物（LVOC）对 NPF 研究尤其重要。GeckoQ 中的所有数据均涉及 LVOC。（更多信息请参见：Besel et al. https://doi.org/10.1038/s41597-023-02366-x）

GeckoQ 为每种分子提供了重要的热力学性质：饱和蒸气压（pSat）、化学势[kJ/mol]、分子在混合物中的自由能[kJ/mol]以及汽化热[kJ/mol]。在这些性质中，学期项目将重点研究对数形式的饱和蒸气压。使用对数刻度而非原始 pSat，是为了使数据量程更易于管理。

你将在项目中可以选择使用两种特征类型：可解释特征（如下详细描述）和分子的拓扑指纹（TopFP）。之前的研究使用 TopFP 描述符作为机器学习模型的输入，以学习不同数据集中原子结构与 pSat 之间的关系。（Wang et al. https://doi.org/10.1073/pnas.1707564114）

以下是训练/测试数据集的列组成。除了 ID 和 log_pSat_Pa 列外，其他列构成分子的可解释特征：

- ID：用于命名文件的唯一分子索引。
- log_pSat_Pa：使用 COSMOtherm 计算的分子的对数饱和蒸气压（Pa）。
- MW：分子的分子量（g/mol）。
- NumOfAtoms：分子中的原子数。
- NumOfC：分子中的碳原子数。
- NumOfO：分子中的氧原子数。
- NumOfN：分子中的氮原子数。
- NumHBondDonors：分子中的氢键供体数，即与氧原子结合的氢原子数。

- Parentspecies：分子的母体物种，可能是“decane”、“toluene”、“apin”中的一种或它们的组合，用下划线连接，表示不明确的来源。在 243 个案例中，由于无法检索母体物种，该值为“None”。
- NumOfConf：由 COSMOconf 发现并成功计算的稳定构象数。
- NumOfConfUsed：用于计算热力学性质的构象数。
- C=C (non-aromatic)：分子中发现的非芳香性 C=C 键数。
- C=C-C=O in non-aromatic ring：分子中非芳香环内发现的“C=C-C=O”结构数。
- hydroxyl (alkyl)：分子中发现的烷基羟基数。
- aldehyde：分子中的醛基数。
- ketone：分子中的酮基数。
- carboxylic acid：分子中的羧酸基数。
- ester：分子中的酯基数。
- ether (alicyclic)：分子中的脂环醚基数。
- nitrate：分子中的脂环硝酸酯基数。
- nitro：分子中的硝基酯基数。
- aromatic hydroxyl：分子中的芳香羟基数。
- carbonylperoxynitrate：分子中的羰基过氧硝酸酯基数。
- peroxide：分子中的过氧化物基数。
- hydroperoxide：分子中的过氧化氢基数。
- carbonylperoxyacid：分子中的羰基过氧酸基数。
- nitroester：分子中的硝酸酯基数。

三、你的任务

你需要与 1-3 名学生组成小组合作。
饱和蒸气压是一个连续变量，因此，你的任务是构建一个基于回归的机器学习模型，该模型使用上述的可解释特征或分子的拓扑指纹。

注意：这是一个非平凡的回归任务，有多种实现方式。最简单的回归模型是线性回归，但由于输入特征与饱和蒸气压（pSat）之间的关系是非线性的，因此线性回归对于该任务来说效率不高。因此，你需要进行全面的数据探索、预处理、特征选择、模型选择、性能评估等，并在学期项目报告中报告和分析你的选择和结果。

本项目的目的并不是（甚至不尝试！）复制文献中的任何方法，也不是制作一个超复杂的、表现最佳的分类器来击败所有其他模型，或试图使用其他数据源等来获得最佳的性能评分。你不应该使用你自己不理解的方法！测试数据预测的准确性本身不是评分标准，尽管糟糕的性能可能表明你的方法存在问题（这可能会影响评分）。

四、在线挑战

我们组织了一场非正式的竞赛（或“挑战”），以使项目更加有趣。我们将使用 R^2分数（https://en.wikipedia.org/wiki/Coefficient_of_determination）作为评估提交结果的指标，所有分数通过将你的测试数据预测与真实标签（我们有，但你没有）进行比较计算得出。R^2分

数本质上是模型预测与真实标签之间相关性的衡量标准。R^2分数越高越好，但如果所选模型不适合这个问题，R^2分数也可能为负值。

由于我们使用 Kaggle 收集你的提交内容，我们将在私有排行榜中使用测试样本的一个子集。对于不熟悉 Kaggle 的同学，私有测试数据行的提交分数将用于确定最终排名。这个"私有排行榜"在竞赛截止日期前仅竞赛主办方可见，截止后将向参赛者公布。

我们为学期项目设置了两个截止日期（请参阅 Moodle 或课程日程了解具体截止日期）。第一份提交内容应包括：

- 你的模型预测结果提交至 Kaggle
- 项目报告的初稿提交至 Moodle

在此截止日期之后，我们将在 Kaggle 上公布私有排行榜分数。

初步报告应描述迄今为止所做的工作。报告不需要精雕细琢或完整，但应包含解决方案中使用的基本思路。团队在提交最终报告前可以修改其方法和报告。然而，请不要简单复制在竞赛中表现良好的团队所使用的方法！

五、最终报告

你需要通过 Moodle 提交最终报告的 PDF 文件（具体截止日期请参见 Moodle）。

最终报告应包括但不限于以下内容：

- 小组成员的姓名。
- 你用于在 Kaggle 提交预测的团队名称。
- 数据分析的各个阶段，包括你如何查看和理解数据（可视化、无监督学习方法等）。
- 所考虑的机器学习方法的描述，以及为该应用选择的方法的优缺点。
- 你为选择良好的特征和模型参数所采取的步骤。
- 结果总结、获得的见解以及回归模型的表现。

最后一部分需包括自评分报告（最多 1 页），使用附带的评分说明（见下文）为自己建议一个分数（整数 0-5）。
为了通过项目，只需使用一种基本算法，按说明进行特征和模型选择（交叉验证可能是个好主意），并准备一份写得很好的报告即可。
撰写报告的实用说明：

你的报告应该像一个自成一体的博客文章或技术报告，没有任务描述，但能够被理解。你应解释你所做的事情以及为什么这么做，使熟悉机器学习的人能够理解并原则上可以基于你的报告重现工作。请重视报告的呈现和可读性（这是评分标准之一）：想象报告的读者是你未来的老板，他们会欣赏清晰简洁的表达。
你不需要提交任何程序代码。因此，你的报告应与代码列表不同！你的报告可以包含代码片段，但应解释读者从代码中得出什么结论。我们可能会查看它们，但不会从代码中去寻找结

果和缺失的细节。换句话说，报告的所有相关部分应在不查看任何代码的情况下可理解。如果需要包括较大块的代码，请将它们放在主报告正文后的附录中。
你的报告可以包含表格或图形。请详细解释这些表格或图形展示了什么，以及读者应该从中得出什么结论。如果有图或表，文本中至少应引用一次。
你可以使用能生成清晰 PDF 输出的合适排版软件（如 LaTeX、Word、R Markdown 等）。没有严格的页数限制，因此可以使用任何可读字体（如 12 号衬线字体）、页边距和适当大小的图形。请注意，Jupyter Notebook 生成的 PDF 格式通常较差。根据我的经验，从其他课程的 16 份类似的最终报告中随机抽样，这些报告得分最高，任务与此相同但没有自评分（这可能会增加一页）。这些最终报告的页数在 7 到 14 页之间，中位数为 12.5 页。
即使你可以为最终报告修改方法和调整算法，但不需要（也可能不应该）做重大更改。目的是完善报告并完成你计划的步骤。

学期项目（最终报告和挑战赛提交）将按整数 0 到 5（1-5 为通过）评分；参见下文评分标准。

最终报告将通过 Ouriginal 抄袭检测系统进行处理。

六、项目评分标准

在课程结束时，你需要为项目的成果（最终报告、演示和挑战赛提交）打一个整数分数，评分范围为 0（不及格）到 5（优秀）。你应在最终报告的最后部分附上评分意见（“评分部分”）。评分部分的长度最多为 1 页。
通常情况下，小组的所有成员将获得相同的课程部分成绩。（如果某些小组成员的贡献存在重大问题，可能会获得不同的分数。如有任何问题，请尽快联系课程工作人员以解决！）课程工作人员将在为你的学期项目打分时参考你的自我评估。

- 成果评分（Grade for the deliverables）

请使用以下评分指南为小组的成果（最终报告、演示和挑战赛提交）打一个整数分数，范围为 0 到 5。请在评分部分的开头清楚说明你给自己打的分数！你的成果在某一方面可能存在不足，但可以通过另一方面的优秀表现来弥补。你应尝试平衡弱点和优势，给出一个能够真实描述小组成果的分数。

关于挑战赛提交的说明：测试数据预测的 R^2分数（及其他性能指标）本身不是评分标准，但较低的 R^2分数可能表明你的方法存在问题，这可能会影响评分。

除了数值评分外，请简要解释（最多 1 页）你评分的原因，使用下述评分标准。评分标准类似于数据科学硕士论文的评估标准。请不要仅仅重复评分标准；说明它们如何适用于你的工作并与之相关。

评分标准

5 分（优秀）：	对主题的处理表现出深入的理解，使用并引用了相关的资料，讨论表现出成熟性。选择并正确应用了合适的机器学习及其他方法。对所用方法进行了充分分析。报告简洁且准确。得出的结论深入且完整。对发现的

	讨论表现出独立、批判和创新的研究与思考能力。报告和演示已达到“可直接发表”的水平。工作具有创造性和独立性，并在规定的时间内完成。成果按照提供的说明完成。
3 分（良好）:	对主题的处理表现出理解。对主题和文献的分析大体上是批判性的。研究材料和方法（包括机器学习方法）适合问题，其使用得到了充分论证。发现的结果大体上清晰地报告。研究问题得到了可行的回答。语言准确，术语定义明确。表达准确，尽管风格可能有所变化。工作主要按照计划时间表进行。成果大体上遵循了给定的说明。
1 分（及格）:	对主题和范围的动机不清晰，对主题和目标的理解也不充分。工作显示出领域知识的重大不足，引用的资料通常很少或质量低下。结果的报告和分析存在重大弱点。结论和讨论不符合科学风格。成果不够精炼。工作未按计划进展。未遵循大部分给定的说明。然而，工作仍然满足最低要求。
0 分 (不及格):	成果未能满足最低要求。

- 小组整体评分

请为你的小组整体打一个 1 到 5 的整数分数，并简要（通常为一段文字）解释你的评分。你可以使用以下评分标准作为指南，尽管不需要对每个标准单独评分。此评分不会直接影响你的课程成绩。如果你是独自完成学期项目，则无需进行此部分。

标准	评分: 5	评分: 3	评分: 1
关于内容的讨论	小组进行分析和批判性的讨论。讨论包括成员自身经验的见解。几乎没有无关的闲聊。	讨论主要围绕项目主题。有一些来自自身经验的例子。与主题无关的讨论有限。	有一些关于项目主题的讨论。有一些个人经验的例子，但它们与工作其他部分分离。存在许多无关或不太相关的讨论。
设定目标及实现目标的努力	小组有一个共同目标，并考虑了每个成员的个人目标。小组确保所有目标都能实现，并在必要时根据工作进展调整目标。	小组有一个共同的目标，部分考虑了个人目标。小组有条理地朝着目标努力，尽管可能无法达到所有目标。	小组没有共同目标。成员各自为战，没有平等分担责任。部分成员没有尽到应有的责任。
参与、责任、互动、氛围	每个人都积极参与讨论和小组工作。所有成员对小组工作负责，同时也尊重他人的想法。责任分配公平。小组氛围鼓励学习和工作，任何冲突都得到解决和吸取教训。	小组成员积极参与会议。责任和工作量分配大致公平。小组氛围良好，尝试解决冲突。	小组难以达成会议时间的共识，且并非所有成员都参与会议。责任和工作分配不均。有些成员做了大部分工作，而其他成员几乎没有贡献。氛围不鼓励学习，冲突未得到解决。
工作成果及对学习的附加价值	小组工作显著促进了成员的学习成果。	小组工作在一定程度上提升了成员的学习质量。	小组工作未对成员的学习带来任何附加价值。