EMATM0044人工智能介绍课件1

截止日期：5月11日星期三13:00

问题1（40分）

对于非金融科技数据科学课程的学生，从Blackboard下载数据集coursework other.csv。该数据集包括每小时出租自行车的数量（列出租自行车计数），以及其他特征，如一天中的小时、日期、湿度等。你的任务是建立一个模型，根据数据集的其他特征值，预测每小时的出租自行车数量。

对于金融科技数据科学课程的学生，可以从黑板上加载数据集coursework fintech.csv。这个数据集包括苹果公司从1995年1月3日至2021年12月31日的每日股票价格，包括开盘价、最高价、最低价、收盘价和调整后的收盘价。你的任务是建立一个模型来预测调整后的收盘价，给定数据集的其他特征值。

对于所有学生，你应该考虑以下方面。

- 要使用的算法种类（例如：分类/回归/聚类）。
- 用来衡量模型性能的指标
- 用什么样的基线来比较模型（sklearn有一个模块sklearn.dummy 这可能有助于生成一个基线)
- 如何选择你的模型的超参数
- 如何测试你的模型的性能

具体来说，你应该使用scikit-learn的两种算法，并比较它们在数据集上的表现。你还应该将你选择的模型的性能与基线进行比较--即一个简单的模型，更复杂的模型应该能够战胜它。sklearn有一个模块sklearn.dummy，在生成基线时可能很有用。你应该使用

技术来评估模型对未见过的数据进行归纳的能力，并确保你对模型性能的评估是可靠的。

工作表13、14、16和17的材料在这里会有帮助。

你对这个问题的回答应该采取简短报告的形式（最多4页），同时附上评论代码，详细说明你将采取的方法。确保你能解决上面所有的要点，并解释你的决定。例如："我选择使用X算法是因为Y"。因为Z，我使用了M指标"。你应该适当地使用图表来说明你的决定。

代码将不会被标记为优雅，但它应该能正确运行。如果你使用的是jupyter，一个好的提示是确保你已经重新启动了内核，并确保代码在提交之前可以从头开始运行。

第一部分评分标准（40分）

至少应测试2种算法。如果只测试了一种，那么该题的最高分是20分。你可以用2种算法加上基线来获得满分。

（5分）报告的整体表述，包括使用适当的章节、图画、图表或表格来说明你的观点。请不要在报告中加入代码片段。相反，要用文字或方程式来描述你所实现的内容。方程格式要正确。

（3分）选择一个合适的算法类型（分类/回归/聚类）并说明这一选择。第13周的讲座和工作表在这里会有帮助。

（3分） 适当选择性能指标(如：准确度/精确度/平均平方误差等)并说明理由。第13周的讲座和工作表在这里会有帮助。

（4分） 讨论与哪种基线进行比较。 (sklearn有一个模块 sklearn.dummy在生成基线时可能是有用的）。

（15分）使用适当的方法来选择所选算法的超参数。在解释选择哪些超参数时，应辅以以下内容
例如，用表格和图表来说明选择了哪些超参数值以及原因。请至少选择一个使用超参数的模型，以便展示你在这个领域的知识。如果你选择了一个没有超参数的模型，那么请用几句话来解释选择没有超参数的模型的好处是什么。第13周的讲座和工作表在这里会有帮助。

分解

- 3分。证明你了解什么是超参数以及如何选择超参数。

- 5分。看看不同的超参数选择对你的模型性能的影响。
- 5分。使用表格、图示或其他表现形式，展示不同的超参数选择对模型性能的影响。
- 2分。说明你做了哪些超参数选择，以及为什么。

（10分）训练和测试模型的性能，以显示模型是否能够概括到未见过的数据，并确保模型的性能是强大的。第13周的讲座和工作表在这里会有帮助。

- 4分。训练模型和选择超参数的方式，以获得稳健的性能
- 3分。测试你的模型的性能并比较它们的性能
- 3分。确保对你的模型进行测试，以显示它们是否能够对未见过的数据进行归纳总结

短报告的建议结构

短报告应不超过4页。 更短的也行。 你应该使用LATEX、MS Word或类似的文本编辑器来准备报告，并以PDF文件的形式提交。

- 导言。说明问题是什么。说明需要使用哪种算法（分类/回归/聚类），并解释为什么需要使用该种算法。
- 方法。说明你将使用哪些具体的算法。说明你将使用哪些性能指标以及为什么。说明你将用什么基线来衡量你的算法。描述你将如何选择算法的超参数。说明你为每个模型选择了哪些超参数，用表格或图表来说明你的决定。
- 结果。报告你的模型的结果。适当地使用表格或图表来说明你的结果。

问题2：10分

Flickr-Faces-HQ（FFHQ）数据集可在https://github.com/NVlabs/ffhq-dataset，并在Karras等人[2019]的论文附录A中描述。注意：你不需要阅读
整个论文的内容。 **我提供了一个模板**，其中包括Gebru等人[2021]的论文中3.2（组成）、3.3（收集过程）和3.5（用途）部分的数据表问题。请**对模板中的问题**提供答案。

页面指南。该模板有2页长。完成的模板和你的答案应该有3页长--大多数问题需要一两句话来回答。有些问题可能需要更长或更短的答案。

问题2的评分标准

- 第3.2节：构成。5分
- 第3.3节：收集过程。3分
- 第3.5节：用途。2分

黑板上有一个只包含相关问题的模板。
第19周的工作表在这里会有帮助。数据表的例子也可以在论文的附录中看到。

参考文献

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 数据集的数据表。*Communications of the ACM*, 64(12):86-92, 2021.URL https://arxiv.org/abs/1803.09010。

Tero Karras, Samuli Laine, and Timo Aila.基于风格的生成器架构，用于生成性对抗网络。在*IEEE/CVF计算机视觉和模式识别会议上*，第4401-4410页，2019年。url https://arxiv.org/abs/ 1812.04948.