机器学习纳米学位

##毕业项目 Joe 优达学城

2050 年 12 月 31 日

I. 问题的定义

(大概 1-2 页)

项目概述

在这个部分，你需要用浅显简洁的语句描述这个项目的一个总体的概念。有几个问题是需要考虑的：

- *需要解决的问题涉及哪个领域？做这个项目的出发点？有哪些相关的数据集或输入数据？*
- *问题的背景信息能够让完全没接触过这个问题的人充分了解这个问题吗？*

问题陈述

在这个部分，你需要清楚地为你将要解决的问题下定义，这应该包括你解决问题将要使用的策略（任务的大纲）。你同时要详尽地讨论你期望的结果是怎样的。有几个问题是需要考虑的：

- *你是否清楚地定义了这个问题。站在读者的角度，他们能否明白你将要解决的问题是什么。*
- *你是否详尽地阐述了你将会如何解决这个问题？*
- *你期望什么样的结果，读者能明白你期望的这个结果吗？*

评价指标

在这里，你需要说明你将要用于评价自己的模型和结果的**指标**和**计算方法**。它们需要契合你所选问题的特点及其所在的领域，同时，你要保证他们的合理性。需要考虑的问题：

- *你是否清晰地定义了你所使用的指标和计算方法？*
- *你是否论述了这些指标和计算方法的合理性？*

II. 分析

(大概 2-4 页)

数据的探索

在这一部分，你需要探索你将要使用的数据。数据可以是若干个数据集，或者输入数据/文件，甚至可以是一个设定环境。你需要详尽地描述数据的类型。如果可以的话，你需要展示数据的一些统计量和基本信息（例如输入的特征（features)，输入里与定义相关的特性，或者环境的描述）。你还要说明数据中的任何需要被关注的异常或有趣的性质（例如需要做变换的特征，离群值等等）。你需要考虑：

- *如果你使用了数据集，你要详尽地讨论了你所使用数据集的某些特征，并且为阅读者呈现一个直观的样本*
- *如果你使用了数据集，你要计算并描述了它们的统计量，并对其中与你问题相关的地方进行讨论*
- *如果你**没有**使用数据集，你需要对你所使用的输入空间（input space)或输入数据进行讨论？*
- *数据集或输入中存在的异常，缺陷或其他特性是否得到了处理？(例如分类变数，缺失数据，离群值等）*

探索性可视化

在这一部分，你需要对数据的特征或特性进行概括性或提取性的可视化。这个可视化的过程应该要适应你所使用的数据。就你为何使用这个形式的可视化，以及这个可视化过程为什么是有意义的，进行一定的讨论。你需要考虑的问题：

- *你是否对数据中与问题有关的特性进行了可视化？*
- *你对可视化结果进行详尽的分析和讨论了吗？*
- *绘图的坐标轴，标题，基准面是不是清晰定义了？*

算法和技术

在这一部分，你需要讨论你解决问题时用到的算法和技术。你需要根据问题的特性和所属领域来论述使用这些方法的合理性。你需要考虑：

- *你所使用的算法，包括用到的变量/参数都清晰地说明了吗？*
- *你是否已经详尽地描述并讨论了使用这些技术的合理性？*
- *你是否清晰地描述了这些算法和技术具体会如何处理这些数据？*

基准模型

在这一部分，你需要提供一个可以用于衡量解决方案性能的基准结果/阈值。这个基准模型要能够和你的解决方案的性能进行比较。你也应该讨论你为什么使用这个基准模型。一些需要考虑的问题：

- *你是否提供了作为基准的结果或数值，它们能够衡量模型的性能吗？*
- *该基准是如何得到的（是靠数据还是假设）？*

III. 方法

(大概 3-5 页）

数据预处理

在这一部分， 你需要清晰记录你所有必要的数据预处理步骤。在前一个部分所描述的数据的异常或特性在这一部分需要被更正和处理。需要考虑的问题有：

- *如果你选择的算法需要进行特征选取或特征变换，你对此进行记录和描述了吗？*
- ***数据的探索**这一部分中提及的异常和特性是否被更正了，对此进行记录和描述了吗？*
- *如果你认为不需要进行预处理，你解释个中原因了吗？*

执行过程

在这一部分， 你需要描述你所建立的模型在给定数据上执行过程。模型的执行过程，以及过程中遇到的困难的描述应该清晰明了地记录和描述。需要考虑的问题：

- *你所用到的算法和技术执行的方式是否清晰记录了？*
- *在运用上面所提及的技术及指标的执行过程中是否遇到了困难，是否需要作出改动来得到想要的结果？*
- *是否有需要记录解释的代码片段(例如复杂的函数）？*

完善

在这一部分，你需要描述你对原有的算法和技术完善的过程。例如调整模型的参数以达到更好的结果的过程应该有所记录。你需要记录最初和最终的模型，以及过程中有代表性意义的结果。你需要考虑的问题：

- *初始结果是否清晰记录了？*
- *完善的过程是否清晰记录了，其中使用了什么技术？*
- *完善过程中的结果以及最终结果是否清晰记录了？*

IV. 结果

（大概 2-3 页）

模型的评价与验证

在这一部分，你需要对你得出的最终模型的各种技术质量进行详尽的评价。最终模型是怎么得出来的，为什么它会被选为最佳需要清晰地描述。你也需要对模型和结果可靠性作出验证分析，譬如对输入数据或环境的一些操控是否会对结果产生影响（敏感性分析 sensitivity analysis）。一些需要考虑的问题：

- *最终的模型是否合理，跟期待的结果是否一致？最后的各种参数是否合理？*
- *模型是否对于这个问题是否足够稳健可靠？训练数据或输入的一些微小的改变是否会极大影响结果？（鲁棒性）*
- *这个模型得出的结果是否可信？*

合理性分析

在这个部分，你需要利用一些统计分析，把你的最终模型得到的结果与你的前面设定的基准模型进行对比。你也分析你的最终模型和结果是否确确实实解决了你在这个项目里设定的问题。你需要考虑：

- *最终结果对比你的基准模型表现得更好还是有所逊色？*
- *你是否详尽地分析和讨论了最终结果？*
- *最终结果是不是确确实实解决了问题？*

V. 项目结论

（大概 1-2 页）

结果可视化

在这一部分，你需要用可视化的方式展示项目中需要强调的重要技术特性。至于什么形式，你可以自由把握，但需要表达出一个关于这个项目重要的结论和特点，并对此作出讨论。一些需要考虑的：

- *你是否对一个与问题，数据集，输入数据，或结果相关的，重要的技术特性进行了可视化？*
- *可视化结果是否详尽的分析讨论了？*
- *绘图的坐标轴，标题，基准面是不是清晰定义了？*

对项目的思考

在这一部分，你需要从头到尾总结一下整个问题的解决方案，讨论其中你认为有趣或困难的地方。从整体来反思一下整个项目，确保自己对整个流程是明确掌握的。需要考虑：

- *你是否详尽总结了项目的整个流程？*
- *项目里有哪些比较有意思的地方？*
- *项目里有哪些比较困难的地方？*
- *最终模型和结果是否符合你对这个问题的期望？它可以在通用的场景下解决这些类型的问题吗？*

需要作出的改进

在这一部分，你需要讨论你可以怎么样去完善你执行流程中的某一方面。例如考虑一下你的操作的方法是否可以进一步推广，泛化，有没有需要作出变更的地方。你并不需要确实作出这些改进，不过你应能够讨论这些改进可能对结果的影响，并与现有结果进行比较。一些需要考虑的问题：

- *是否可以有算法和技术层面的进一步的完善？*
- *是否有一些你了解到，但是你还没能够实践的算法和技术？*
- *如果将你最终模型作为新的基准，你认为还能有更好的解决方案吗？*

** 在提交之前， 问一下自己... **

- 你所写的项目报告结构对比于这个模板而言足够清晰了没有？
- 每一个部分（尤其**分析**和**方法**）是否清晰，简洁，明了？有没有存在歧义的术语和用语需要进一步说明的？
- 你的目标读者是不是能够明白你的分析，方法和结果？
- 报告里面是否有语法错误或拼写错误？
- 报告里提到的一些外部资料及来源是不是都正确引述或引用了？
- 代码可读性是否良好？必要的注释是否加上了？
- 代码是否可以顺利运行并重现跟报告相似的结果？