小组作业。应用深度学习 - 菲利普-布莱特琴博士

任务概述

当使用机器学习进行分类时，如果类是 "平衡的"，事情就最容易了--也就是说，当属于每个类的观察值的数量是相同的数量级。不幸的是，情况往往不是这样的。在这项作业中，你将处理一个汽车保险索赔的数据集，并尝试将索赔分为欺诈性（1）和非欺诈性（0）。数据集中有超过10,000项索赔，但只有大约100项是欺诈性的。尽管如此，我们还是想创建一个模型，帮助保险商有针对性地开展调查工作。

学习目标

- 了解如何在不平衡的数据集背景下，根据具体的应用来评估机器学习模型。
- 评估具有不同复杂程度的模型。
- 了解自动编码器的广泛用途，以及如何应用它们来检测异常情况。
- 获得关于如何训练神经网络的直觉。

任务描述

1. 仔细阅读Moodle上第五周材料中的案例 "抓贼"。Shift的欺诈检测算法会标记出那些看起来可疑的索赔，然后由人类索赔处理人员进行调查。该算法成功的一个关键指标是 "命中率"（或转换率）。然而，在命中率和检测率之间存在着权衡。请用300字以内的文字描述这两个指标，它们之间的权衡，以及你将如何以货币形式评估一个具有命中率h和检测率d的模型。

2. 加载数据集 "Insurance_claims.csv"（数据和数据词典也可在Moodle的第5周材料中找到）。根据用例对数据集进行适当的预处理。

3. 首先考虑一个简单的模型：如果事件发生在认购日期的t天内，则索赔是可疑的。对于不同的t值，这个模型的命中率和检测率是多少？ 如何选择一个合理的t值？这个模型的基础是什么逻辑？

4. 接下来考虑一个中间模型：特别是实现一个决策树分类器来识别欺诈案件。这个模型的命中率和检测率是多少？请确保你在一个单独的测试集上评估该模型!

5. 接下来考虑一个复杂的模型：特别是在TensorFlow中实现和调整一个（深度）神经网络。这个模型的命中率和检测率是多少？请确保你在一个单独的测试集上评估这个模型!

6. 因为我们只有有限的数据，我们可能想使用异常检测而不是预测欺诈标签。
   a) 创建一个只包含非欺诈性索赔的训练集，以及包含非欺诈性和欺诈性索赔的验证和测试集。确保将欺诈性索赔均匀地（但随机地）分布在验证和测试集中。
   b) 使用TensorFlow，创建一个自动编码器。因为你的训练集只包含非欺诈性的索赔，所以自动编码器应该学会很好地表示这些索赔。一个好的指标是真实数据和重构数据之间所有特征的平均平方误差。
   c) 在直方图中画出你的验证集的预测和真实数据之间的误差--在一个好的模型中，欺诈性索赔的误差应该比非欺诈性索赔高得多。请确保在直方图中标出不同类型的索赔。
   d) 如果你的直方图中的区别不明显，你可能要对你的模型进行微调。在选择指标时要谨慎。
   e) 一旦你确定了一个最终的模型，就对实际数据和预测数据之间的误差设定一个阈值。也就是说，如果一个观察结果的误差比阈值大，你就预测它是一个欺诈性的索赔。如果误差低于阈值，你就预测它是非欺诈性的。
   f) 使用你的模型和你的阈值，找到命中率和检测率。请确保你在测试集上评估该模型!

7. 从命中率和检测率的角度，讨论你所尝试的不同方法。同时，考虑每种方法背后的透明度。如果没有明确的解释为什么该工具将某一事件标记为可能是欺诈性的，为什么会有问题？请用300字以内的篇幅来回答。

8. 你还能设想出哪些处理不平衡数据的方法来改善Shift对保险欺诈的预测？请用150字以内的文字来回答。

提示

- 在第五部分，不要花太多时间来完善模型。此时最重要的是了解用这样一个不均衡的数据集训练模型所面临的问题，并探索一系列的模型。只要你能打败决策树分类器（意味着你在相同或更高的检测率下获得更高的命中率），你的模型的性能对评估来说并不重要。
- 在你的自动编码器中（第6部分），确保你的网络可以重新创建你的输入。例如，

如果你的输出层使用tanh激活，输出将在-1和-2之间。
1. 记住要适当地扩大你的投入。

- 对于第6部分，从创建一个最小可行产品开始，牢记我们如何定义一个有用的模型。只有当所有的东西都运行后，你才应该回到它，看看如何改进你的模型。你的模型的性能将对评估很重要，但不如有一个完整的答案那么重要。
- 清理数据集也是如此：结合罕见的类别，不要过度设计新的特征，而只从现有的子集开始。然后，一旦一切工作顺利，你就可以考虑加强你的数据集。

要提交的材料

- 一个Jupyter笔记本，可以重新创建你的解决方案。如果你已经尝试并放弃了不同的方法，请将这些方法清除掉，但一定要描述你的过程。
- 你在第5和第6部分开发的训练好的模型为.h5-文件
- 你的书面答案，可以在Jupyter笔记本中，也可以在一个单独的.pdf文件中。请确保在你的笔记本和单独的.pdf文件中创建与当前任务（1-8）相对应的编号部分。
- 一个最多3张幻灯片的PowerPoint幻灯片，比较你所开发的不同方法，讨论你所面临的主要挑战。我们将随机挑选几个小组在下一堂课上展示他们的工作。被选中的小组将在提交当天的晚上8点前被告知。请将您的演示保持在5分钟以内。

评估

您提交的材料将根据四个标准进行评估。

- 适当使用课堂上讨论的概念和框架。
- 建议的答案/解决方案的有效性。
- 建议的答案/解决方案的原创性和创造性。
- 所提交材料的组织和清晰度。