使用随机森林估计和推断异质性治疗效果

译者 Tinkle

2025 年 3 月 12 日

摘要

尽管被广泛的采用，机器学习模型大多是黑匣子。然而，了解预测背后的原因对于评估信任非常重要，如果一个人计划根据预测采取行动，或者在选择是否部署新模型时，这一点至关重要。这种理解还提供了对模型的见解，可用于将不可信的模型或预测转换为可信的模型或预测。

在这项工作中，我们提出了 LIME，这是一种新颖的解释技术，通过围绕预测在本地学习可解释模型，以可解释和忠实的方式解释任何分类器的预测。我们还提出了一种解释模型的方法，通过以非冗余的方式呈现具有代表性的单个预测及其解释，将任务构建为子模优化问题。我们通过解释文本（例如随机森林）和图像分类（例如神经网络）的不同模型来展示这些方法的灵活性。我们通过模拟和人类受试者的新颖实验来展示解释在各种需要信任的场景中的效用：决定是否应该相信预测、在模型之间进行选择、改进不可信的分类器以及确定为什么不应该信任分类器。

1 引言

机器学习是许多最新科学和技术进步的核心。不幸的是，人类的重要作用是该领域经常被忽视的一个方面。无论人类是直接使用机器学习分类器作为工具，还是在其他产品中部署模型，一个至关重要的问题仍然存在：如果用户不信任模型或预测，他们就不会使用它。区分信任的两种不同（但相关）定义很重要：（1）信任预测，即用户是否足够信任单个预测以根据它采取一些行动，以及（2）信任模型，即用户是否相信模型在部署后会以合理的方式运行。

两者都收到人类对模型行为的理解程度，而不是将其视为黑匣子。当模型用于决策时，确定对单个预测的信任度是一个重要问题。例如，当使用机器学习进行医疗诊断 [6] 或恐怖主义侦查时，不能盲目相信地进行预测，因为后果可能是灾难性的。