现实世界应用示例

癌症预测

现实世界应用示例:癌症预测

- 模型经过训练后可以根据病历来预测"病人患有癌症的概率"
- 特征包括病人年龄、性别、之前的病史、医院名称、生命体征、 检验结果
- 模型在处理预留检验数据方面表现出色
- 但模型在针对新病人进行预测时表现却很糟糕, 这是为什么呢?

出现上述问题的原因是这样的:像我们所说的,模型中包含的一个特征是医院名称。

其中有些医院名称类似于"贝斯以色列癌症中心"。其实这类名称能清楚的表明相应患者是否真的患有癌症。即使医院名称不包含"癌症"一词,很多医院的各种不同专业也能表明患者的患癌情况。有些医院专门治疗癌症,有些则不是。即使不使用"贝斯以色列癌症中心",把它改成隐藏医院名称的整数,依然会有很多患者的患癌情况与该医院密切相关,因为它是一家专门治疗癌症的医院。但是,对于尚未分配到医院的新患者,我们无从得知该信息。其实想模型提供医院名称会带来一种微妙的欺骗性。在这个示例中,所收集的数据以微妙的方式向模型显示了医生的诊断结论,而在模型尝试取代医生去做判断时却无法获得此类信息。

我们将这种情况成为"标签泄露",即一些训练标签泄露到特征中,使模型带有欺骗性。

这是一种典型的失败案例,请务必避免。

此例子参考文献: https://dl.acm.org/citation.cfm?id=2020496

18世纪文学

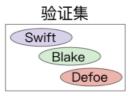
现实世界应用示例: 18 世纪文学

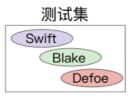
- 某位 18 世纪文学教授想要仅根据作者使用的"心灵隐喻"来预测作者的政治派别。
- 研究小组建立了一个大型的有标签数据集(其中逐句纳入了许多作者的作品),并将其拆分成了训练集/验证集/测试集。
- 训练后的模型在根据测试数据进行预测时的表现几近完美,但研究人员却怀疑结果的准确性。可能出了什么问题?

现实世界应用示例: 18 世纪文学

• 数据拆分方式 A: 研究人员将每位作者的一些样本放在训练集中,一些放在验证集中,另一些 放在测试集中。

训练集 Swift Blake Defoe





现实世界应用示例: 18 世纪文学

• 数据拆分方式 B: 研究人员将每位作者的所有样本都放在单个集中。







现实世界应用示例: 18 世纪文学

- 数据拆分方式 A: 研究人员将每位作者的一些样本放在训练集中,一些放在验证集中,另一些 放在测试集中。
- 数据拆分方式 B: 研究人员将每位作者的所有样本都放在单个集中。
- 结果:根据数据拆分方式 A 训练的模型比根据数据拆分方式 B 训练的模型的准确率要高得多。

现实世界应用示例: 18 世纪文学

结论: 仔细考虑如何拆分样本。

了解数据代表的含义。

在进行这次试验的过程中,我们发现根据测试得出较高的准确率要难得多,而且仅仅根据隐喻数据来预测政治派别也难得多。

这其实挺有意思的。

一方面, 我们可以就此写两篇文章, 每篇文章各阐述一种观点。

另一方面,我们学到了重要的一课,那就是,在随机化处理训练数据和测试数据以进行这种拆分时,考虑采用 哪种方式还是至关重要的。

机器学习研究人员很容易就会认为: 我们只是将数据随机化处理一下就行了。

但实际上,我们需要先了解数据所代表的意义,然后我们才能进行有效的拆分。

此例子参考文献为:http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.103.21

现实世界应用准则

一些有效的机器学习准则

- 确保第一个模型简单易用
- 着重确保数据管道的正确性
- 使用简单且可观察的指标进行训练和评估
- 拥有并监控您的输入特征
- 将您的模型配置视为代码: 进行审核并记录在案
- 记下所有实验的结果, 尤其是"失败"的结果

更多准则: https://developers.google.cn/machine-learning/rules-of-ml/