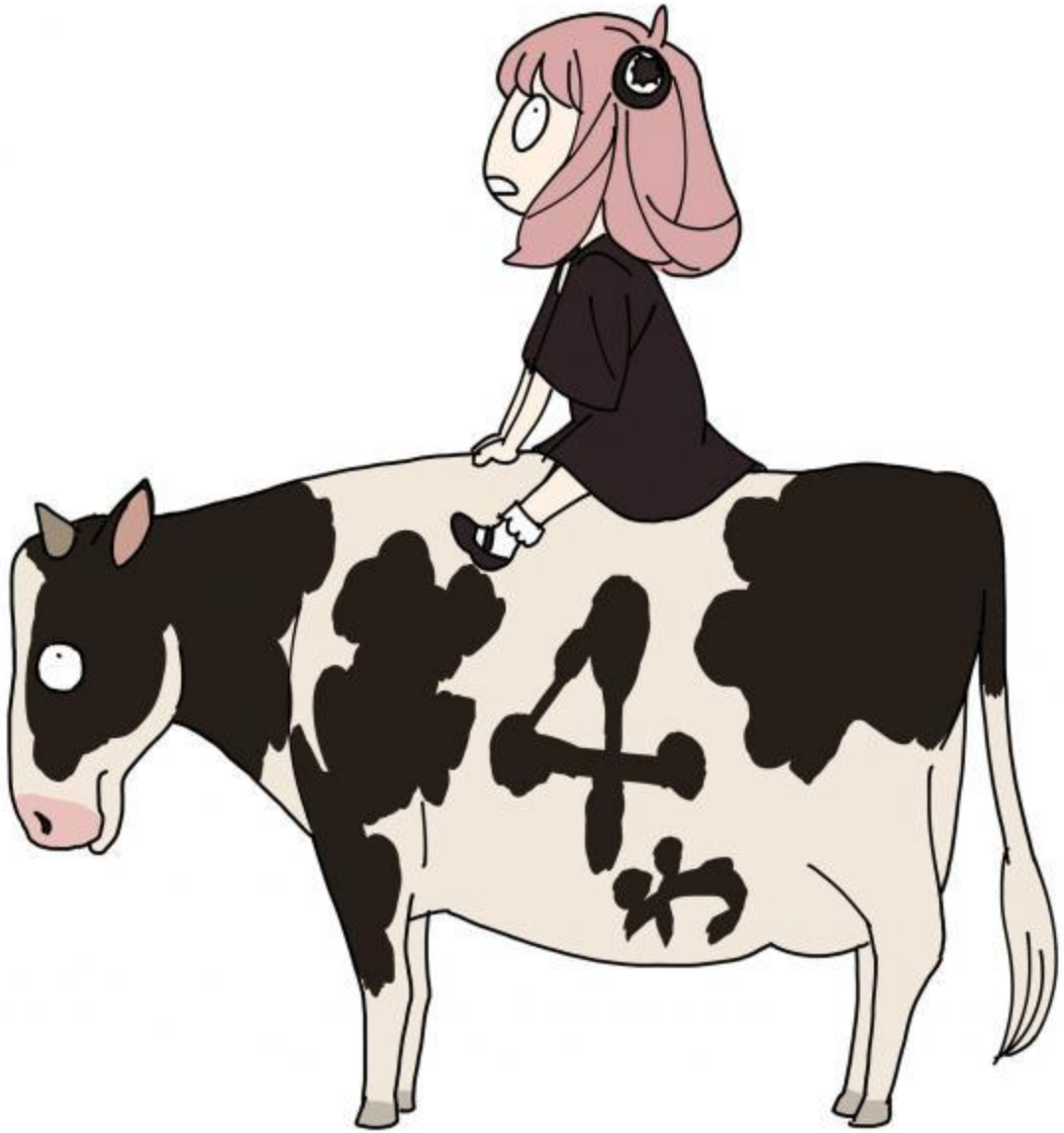
Tabnet系列讲解

【模型介绍及使用介绍】

目录

- 引言
- 相关研究
- TabNet介绍
- 实验



引言

- 表格数据的深度学习仍未得到充分探索，集成决策树(DTs)的变体仍然主导着大多数应用。
- 其中GBDT方法仍是主流，在kaggle上大放异彩[2]。

GBDT的优势

- 对于具有近似超平面边界的决策流形具有代表性的有效性，这些边界在表格数据中很常见；
- 其基本形式具有高度的可解释性(例如，通过跟踪决策节点)，它们的集成形式有流行的事后可解释性方法，这是许多现实世界应用中的一个重要问题；
- 训练起来很快(归功于大佬们的深入的性能优化，如：LightGBM)；

神经网络的优势

- 深度学习在大数据集上的预测性能高（相比于传统机器学习）；
- 有效编码多种数据类型，如图像、文本和表格数据；
- 减轻特征工程的需要，这是目前基于树的表格数据学习方法的一个关键方面；
- 可以从流数据学习，也许最重要的是(iv)端到端模型允许表示学习；

相关研究

- 曾经提出的许多的nn架构不适合款表格数据:例如，堆叠的卷积层或多层感知器(MLPs)被过度参数化——缺乏适当的归纳偏差往往导致它们无法找到表格决策流形的最优解决方案；
- 但，也有不少工作在“模仿tree的行为”中努力，考虑怎么用DNN实现类似于tree的功能；

如：

软(神经)DTs使用可微的决策函数，实现可微决策树[3]；

软分集函数，通过低效地枚举所有可能的决策来模拟dnn中的DTs[4]；

...（虽然研究了很多，但实际效果并不理想）

TabNet介绍-表格数据DNN架构

- TabNet不需要任何预处理就可以输入原始的表格数据，并使用基于梯度下降的优化进行训练，从而灵活地集成到端到端学习中；
- 在每个决策步骤中，TabNet使用顺序注意来选择要从哪些特征进行推理，这使得可解释性和更好的学习成为可能，因为学习能力用于最显著的特征；（看原文图片）
- TabNet采用单一的深度学习架构进行特征选择和推理；

- 使得：
- 【性能】TabNet在不同领域的分类和回归问题上优于或与其他表格学习模型持平；
- 【可解释性】TabNet支持两种可解释性:局部可解释性和全局可解释性。

TabNet介绍-pytorch_tabnet

- pytorch_tabnet[5]：是基于pytorch的TabNet实现；（见文档[6]）
- 可实现：TabNet的二分类、多分类和回归问题。

实验设置

- 数据来源：kaggle 心脏病个人关键指标数据
- 数据量大小：‘no’：292422 ‘yes’：27373
- 优化器：Adam
- 评价指标：balanced accuracy 以及 auc

实验结果

- valid auc = 0.83 （TabNet）　　valid auc = 0.84 （XGBoost）
- test auc = 0.85 （TabNet）　　test auc = 0.84 （XGBoost）

- valid balanced accuracy = 0.77 （TabNet）
- test balanced accuracy = 0.84 （TabNet）

- valid balanced accuracy = 0.54 （XGBoost）
- test balanced accuracy = 0.54 （XGBoost）
- （balanced accuracy不知是否不一致）

随机欠采样-tabnet

best_valid_balanced_accuracy = 0.52593

- Best weights from best epoch are automatically used!
- BEST VALID SCORE FOR heart_2020_cleaned : 0.5259322288168171
- BEST VALID AUC SCORE FOR heart_2020_cleaned : 0.5254561521944999
- FINAL TEST AUC SCORE FOR heart_2020_cleaned : 0.4920871019261001

SMOTE过采样-tabnet

- valid balanced_accuracy = 0.79419
- test balanced_accuracy = 0.79219
- valid auc = 0.87370
- test auc = 0.87503

smote-xgboost

- ******* AUC *******
- 0.9715797518318 15(valid)
- 0.9698806883788007(test)
- ******* balanced_acc *******
- 0.9132421823559185(valid)
- 0.9112478971519495(test)

SMOTE_ENN-TabNet

- valid balanced_accuracy = 0.85645
- test balanced_accuracy = 0.85156

- valid auc = 0.93028
- test auc = 0.92721

SMOTE_ENN-XGBoost

- valid balanced_accuracy = 0.93531
- test balanced_accuracy = 0.93841

- valid auc = 0.98351
- test auc = 0.98441

接下来工作

- 分别用欠采样、混合采样处理，看结果；
- 了解TabNet原理；
- 检查该数据集是否能继续下一步实验；

参考文献

[1] Arik S Ö, Pfister T. Tabnet: Attentive interpretable tabular learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(8): 6679-6687.

[2] https://medium.com/machine-learning-insights/winning-approach-ml-competition-2022-b89ec512b1bb

[3] Kontschieder P, Fiterau M, Criminisi A, et al. Deep neural decision forests[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1467-1475.

[4] Yang Y, Morillo I G, Hospedales T M. Deep neural decision trees[J]. arXiv preprint arXiv:1806.06988, 2018.

[5] https://github.com/dreamquark-ai/tabnet

[6] https://dreamquark-ai.github.io/tabnet/index.html