### 第三十二次2020年7月26日-8月2日

#### 1 某些测试用例运行报错的问题

某些测试用例报错问题， BCP在处理Exactly\_one约束产生的两文字同时为真的冲突时，可能会改变Comp中的clause，可能报错（clause为空导致MAX\_LIT启发式内存访问错误），没有报错也会导致结果出错。

一种方法是修正对Comp的错误修改，另一种方法是对这种两文字同时为真的冲突进行特殊处理。

#### 2 XAI - Explainable Artificial Intelligence

XAI的意义：In order to trust deployed AI systems, we must not only improve their robustness, but also develop ways to make their reasoning intelligible. Intelligible will help us spot AI that makes mistakes due to distributional drift or incomplete representations of goals and features. Intelligibility will also facilitate control by humans in increasingly common collaborative human/AI teams. Furthermore, intelligibility will help humans learn from AI. Finally, there are legal reasons to want intelligible AI, including the European GDPR and a growing need to assign liability when AI errs.

XAI的几个问题：

1.Assess robustness

• How easy it is to fool an ML model?

2.Learn interpretable models

• Decision trees; decision sets; decision lists; etc.

3.Explain black-box models

• By using some accepted definition of explanation

4.How about heuristic approaches?

• No formal guarantees provided

研究步骤：

第01部分：接触正式推理工具

第02部分：学习可解释的模型

第03部分：评估ML模型的鲁棒性

第04部分：ML模型的严格说明

第05部分：解释和总结

规范用词：

Rigorous严格 Recognition承认 Brittle脆

Adversarial对抗的 Unsettling令人不安的 Benign良性 Malignant恶性 Intelligible可理解的 Interpretable可解释的 Facilitate促进 Collaborative合作的 Liability责任

Draft草案 Drift偏移 Redundant多余的