ch14人工智能安全

算法局限性 P112

- 数据难以获取 P112
- 数据不完整或偏斜 P114
- 成本局限性 P116
- 算法偏见 P119
- 伦理局限 P122

算法安全 P61

鲁棒性安全 P61

- 人工智能算法的优化原理 P62 ~ 63
- 人工智能算法的可解释性 P64 ~ 69
- 1. 建模前可解释性方法：数据可视化，寻找ProtoTypes 和 Criticisms（典例与特例）——Prototype 是指能够表示大部分数据的数据实例，Criticism 是指不能由Prototype很好表示出的数据实例
- 2. 建立本身具备可解释性的模型：一些具有良好可解释性的模型包括决策树模型、线性模型以及贝叶斯实例模型等
- 3. 使用可解释性方法对模型进行解释：敏感性分析和隐层分析，基于可视化方法的模型解释（结构可视化、训练可视化）
- 人工智能算法的鲁棒性评估 P70

分类维度 P72 ~ 78

- 白盒攻击：攻击者能够获知机器学习所使用的算法，以及算法所使用的参数。攻击者在产生对抗性攻击数据的过程中能够与机器学习的系统有所交互。也就说，攻击者可以将样本送入模型中以获取梯度等信息，然后依据这些信息对输入进行调整
- 黑盒攻击：攻击者对攻击的模型的内部结构，训练参数，防御方法（如果加入了防御手段的话）等等一无所知，只能通过输出输出与模型进行交互。攻击者会使用one-pixel-attack暴力攻击或使用迁移样本进行攻击。这里的迁移样本指的是，对抗样本往往是可迁移特性，即针对机器学习模型A构造的对抗性图像，也会有很大的比例能欺骗机器学习模型B。
- 目标攻击：生成的对抗样本被DNNs误分类为某个指定类别。目标攻击一般发生在多分类问题中。
- 无目标攻击：生成的对抗样本识别结果和原标注无关，即只要攻击成功就好，对抗样本最终属于哪一类不做限制。因为无目标攻击有更多的选择和更大的输出范围，所以比目标攻击更易实现。
- 基于梯度攻击：考虑模型对样本的梯度，根据梯度的方向和大小等对样本进行调整，使损失函数增大
- 基于优化攻击：将输入样本视为可变量，并通过优化算法来最小化或最大化某个特定的目标函数，以找到最优的输入样本，使得模型在该样本上产生误导性的输出结果

- 安全角度审视机器学习系统 P79
- 投毒攻击 P82 ~ 83
- P84 与对抗攻击区别
- P88 ~ 89 投毒攻击的防御
- 对抗样本攻击 P103

人工智能安全应用 P23

- 网络信息安全应用：主要包括网络安全防护应用、信息内容安全审查应用、数据安全管理应用等
- 社会公共安全应用：主要包括智能安防应用、金融风控应用等

框架安全 P32

- Pytorch 相比 TensorFlow 的优势
- **Keras** P36 ~ 37
- Caffe P38~39
- 框架安全漏洞
  - 1. TensorFlow P41 ~ 43：推理/训练中的拒绝服务攻击，分段错误，在数据流图中插入恶意操作后，不影响模型的正常功能，也就是说模型的使用者从黑盒角度是没有感知的
  - Pytorch、Keras 和 Caffe P44 ~ 45：内存泄漏，权重丢失隐患，SQL注入漏洞
- **环境接触带来的漏洞**
  - 1. 第三方基础库漏洞 P47 ~ 50：Numpy 拒绝服务，OpenCV 堆溢出
  - 2. 可移植软件容器漏洞 P51 ~ 52：Kubeflow 挖矿，部署恶意容器