ПРОБЛЕМЫ ДОВЕРИЯ ТЕХНОЛОГИЯМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Коваленко Андрей Петрович

Академия криптографии Российской Федерации

Диалектика против механицизма:

модели машинного обучения (или, если угодно, искусственного интеллекта) - это математические функции, аппроксимирующие требуемую функцию по таблице ее значений, построенной на основе заданного обучающего набора наблюдений.

Ошибки и уязвимости, свойственные моделям искусственного интеллекта:

- Переобучение
- Дрейф данных
- Предвзятость обученной модели
- Выбросы в данных

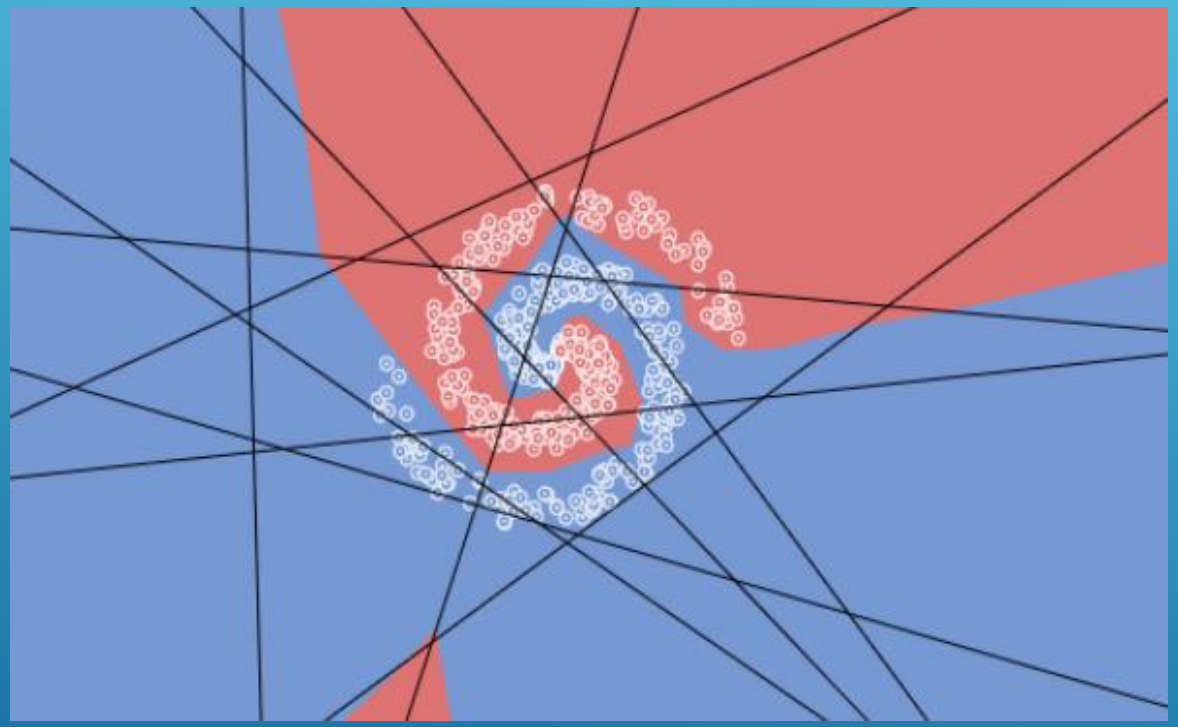
Угрозы информационной безопасности, специфические для моделей искусственного интеллекта:

- Обучение модели нежелательному поведению путем «отравления» данных
- Несанкционированный доступ к обучающим данным на основе анализа обученной модели (атака инверсии модели)
- Введение модели в заблуждение в результате атаки градиентного спуска
- Подмена модели

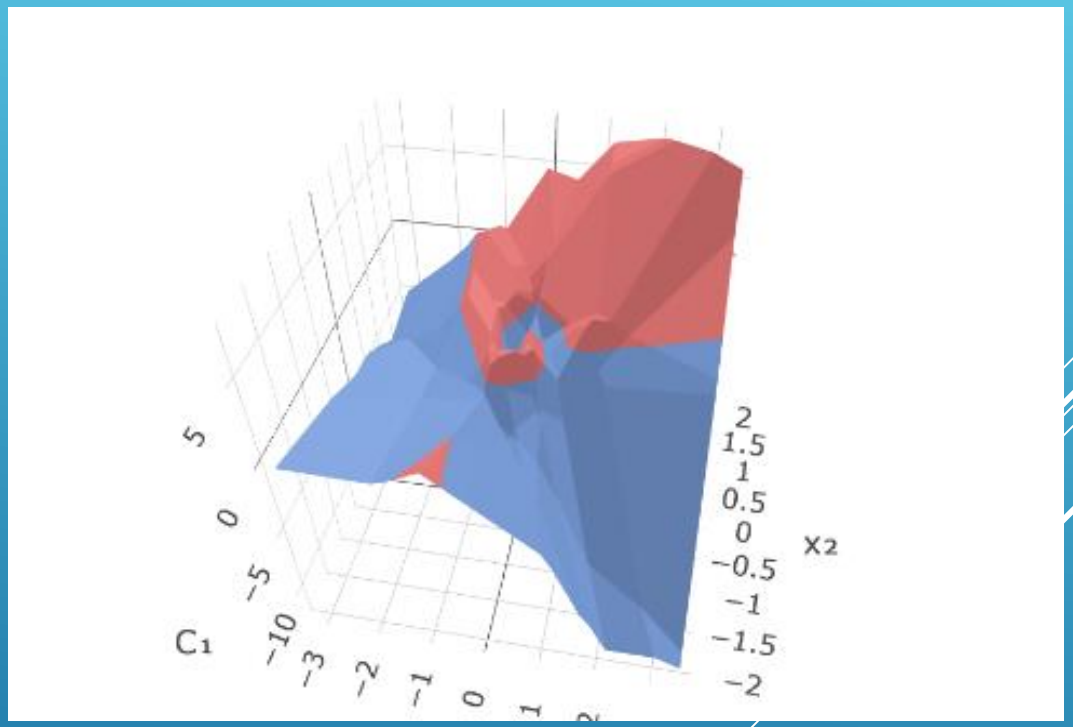
Факторы доверия:

- Наличие доверенного набора обучающих данных «достаточного» объема
- Наличие доверенного ПО для реализации (обучения, верификации, тестирования, дообучения, применения) модели МО
- Наличие достаточного объема вычислительных ресурсов
- Использование теоретически обоснованных («доверенных») моделей МО

Проблема экстраполяции MLP

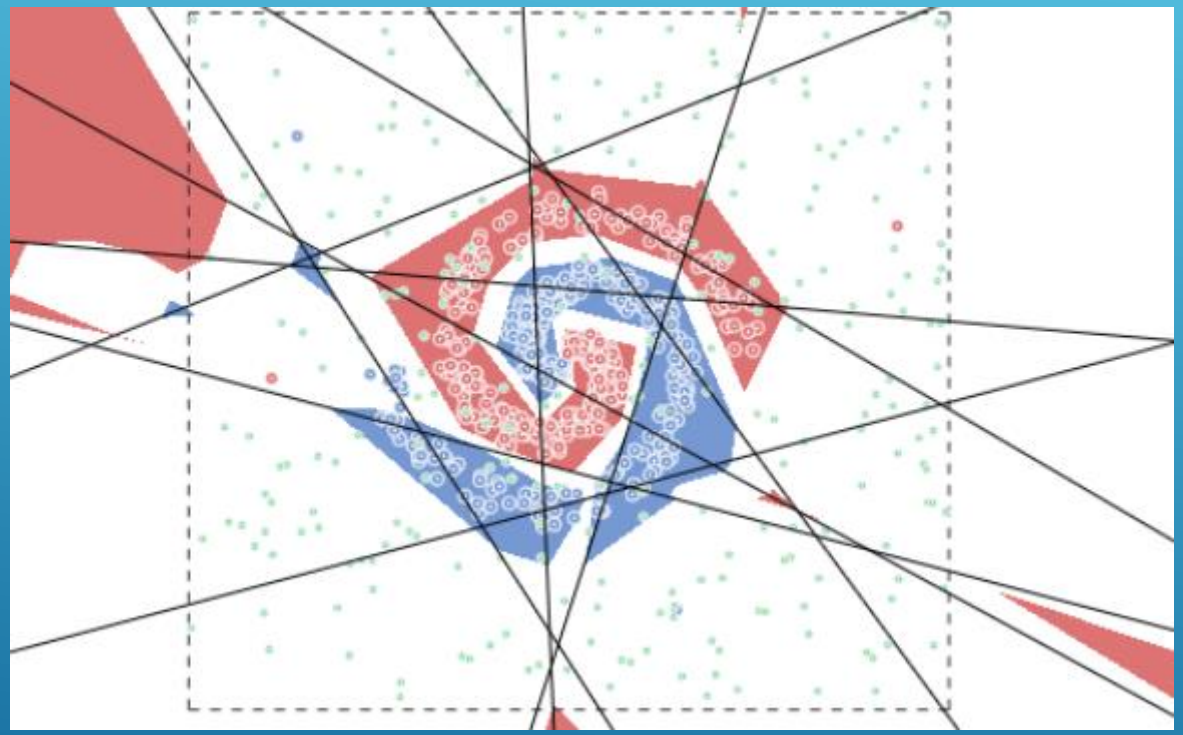


2D

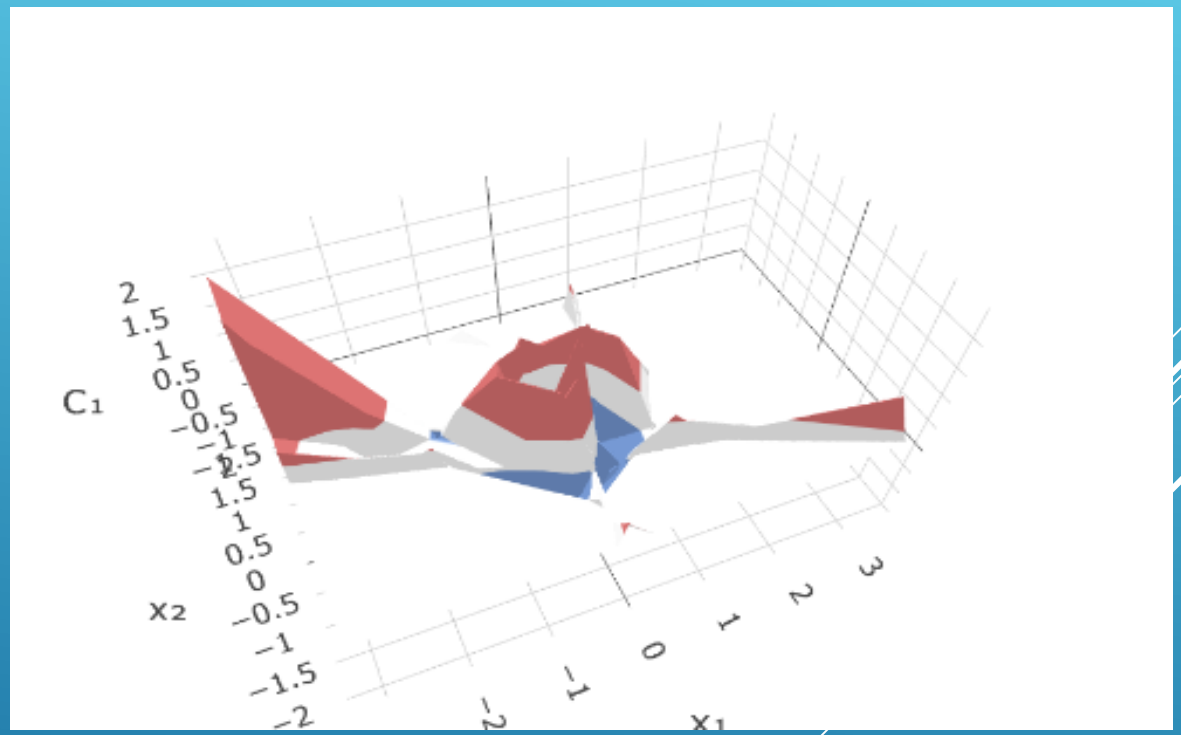


3D

Проблема экстраполяции MLP: подход к решению

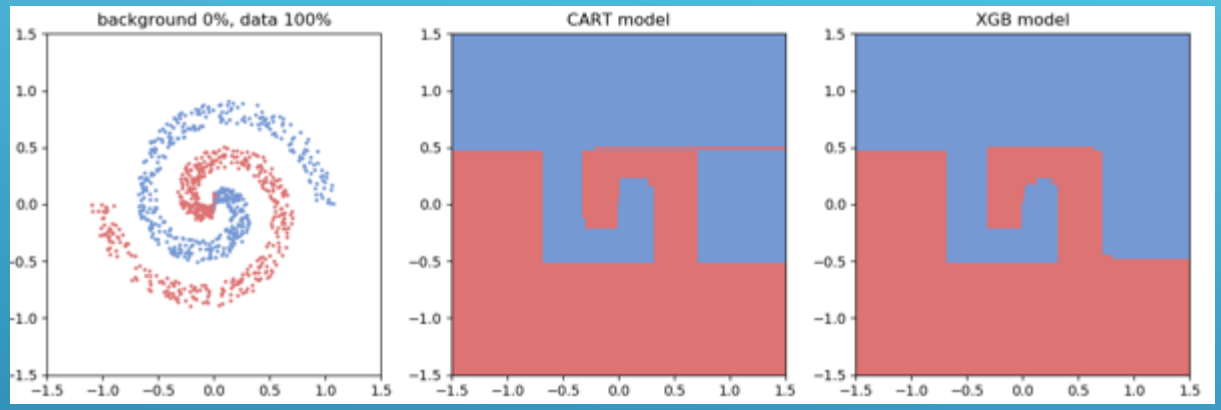


2D

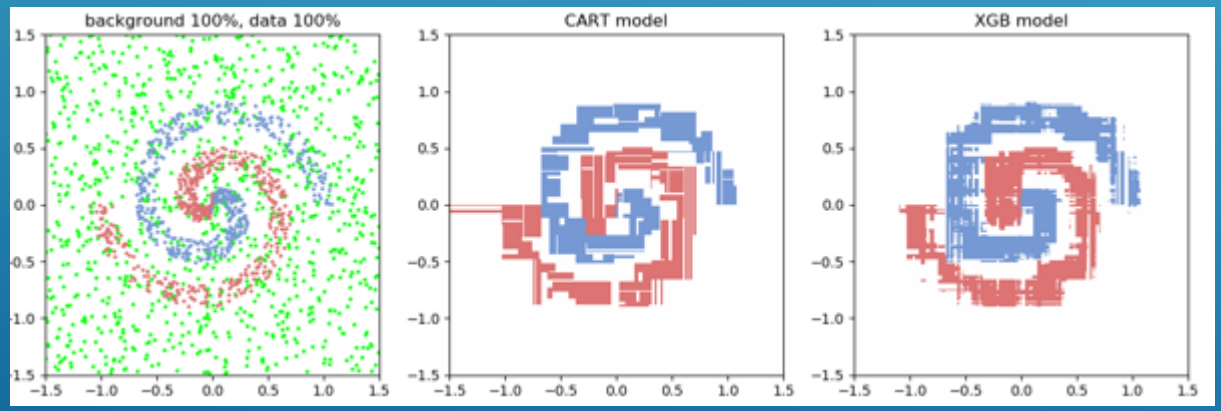


3D

Проблема экстраполяции деревьев решений CART и XGBoost

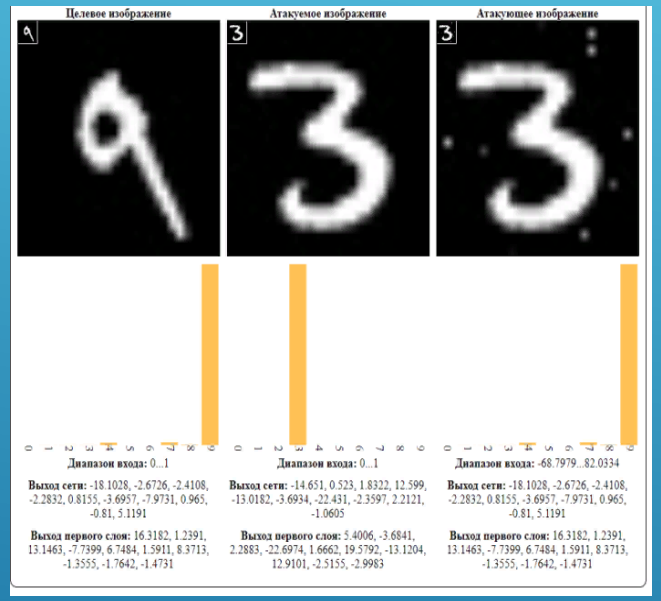


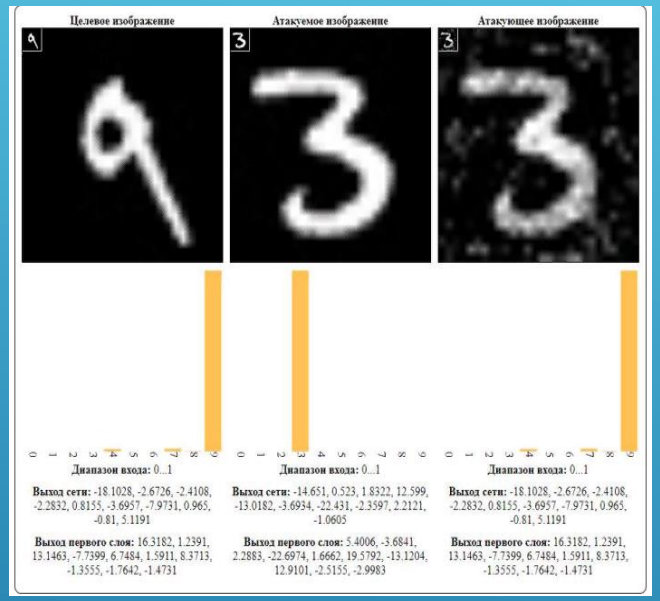
без «фона»

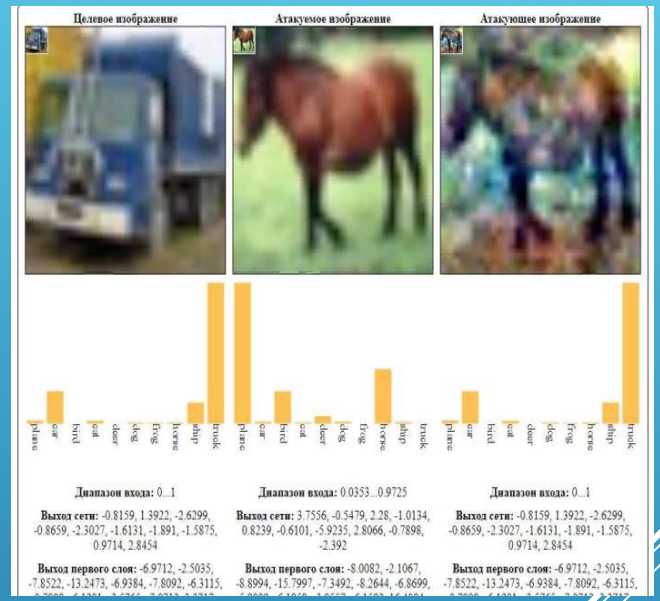


после добавления «фона»

Проблема многозначности MLP







Что делать?

1. Не доверять «черным ящикам» !

2. Исследовать свойства математических функций, реализуемых моделями МО

3. Разрабатывать статистические модели, аналогичные моделям МО, и исследовать их свойства.

4.?????

Спасибо за внимание!