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

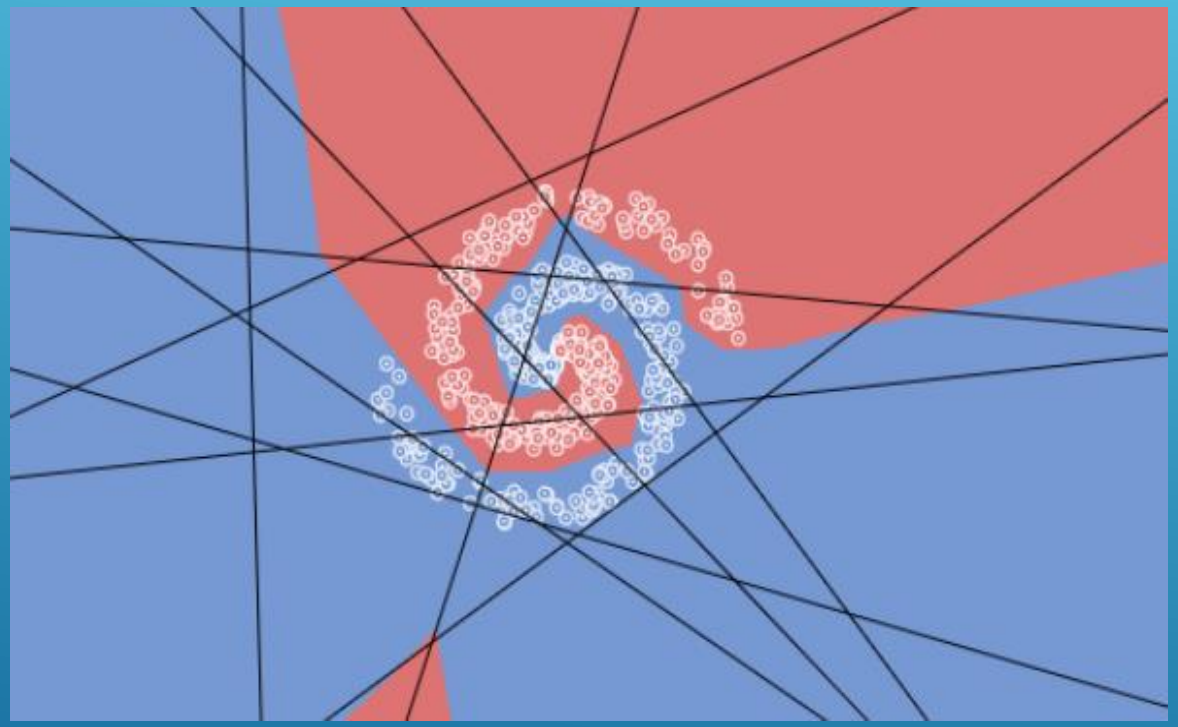
Угрозы информационной безопасности, специфические для моделей искусственного интеллекта:

- 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

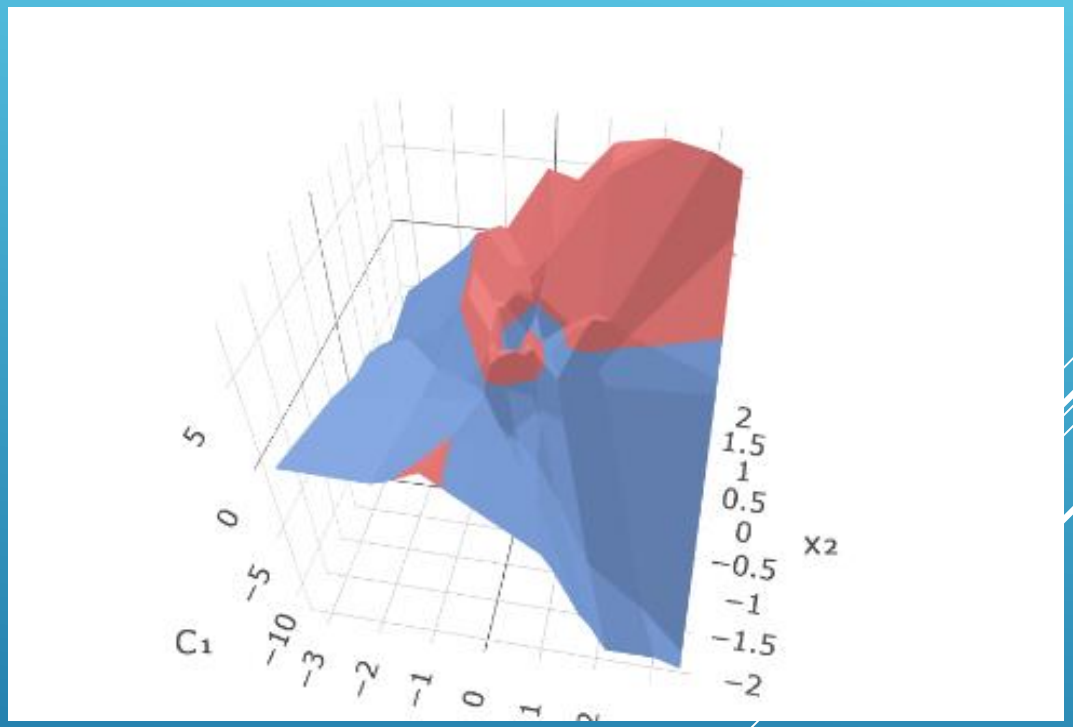
Факторы доверия:

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

Проблема экстраполяции MLP



2D

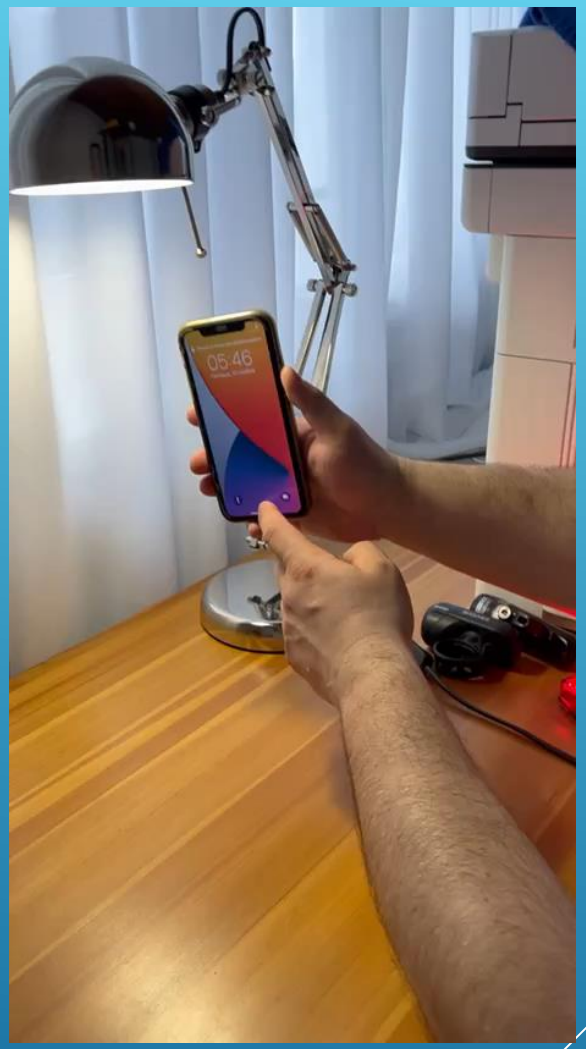


3D

Атака на Face ID

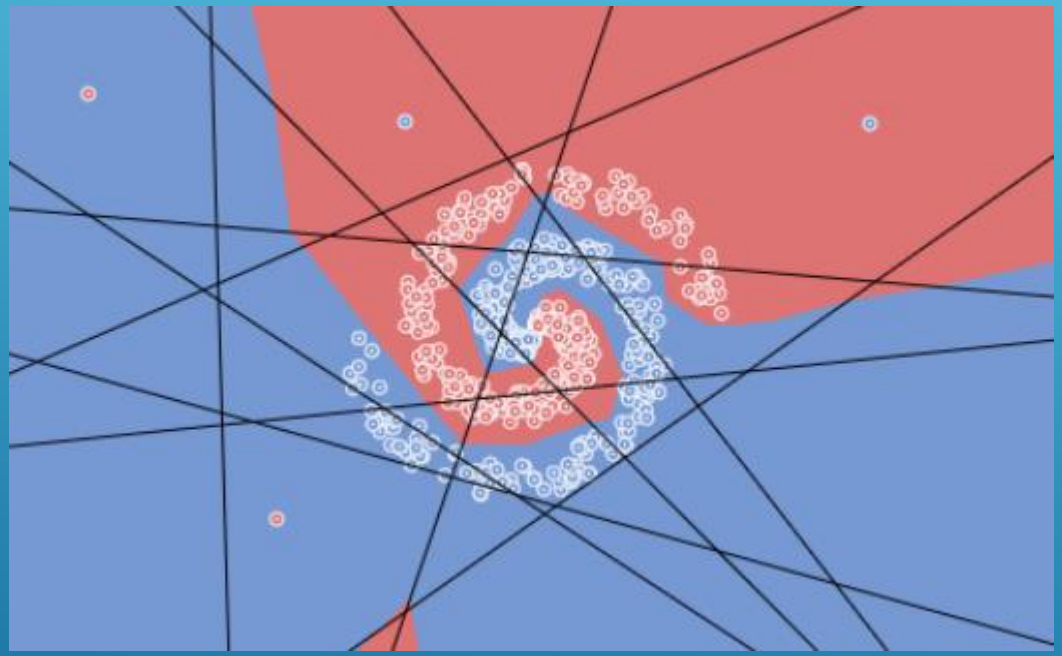
Регистрация неживого объекта

Распознавание неживого объекта

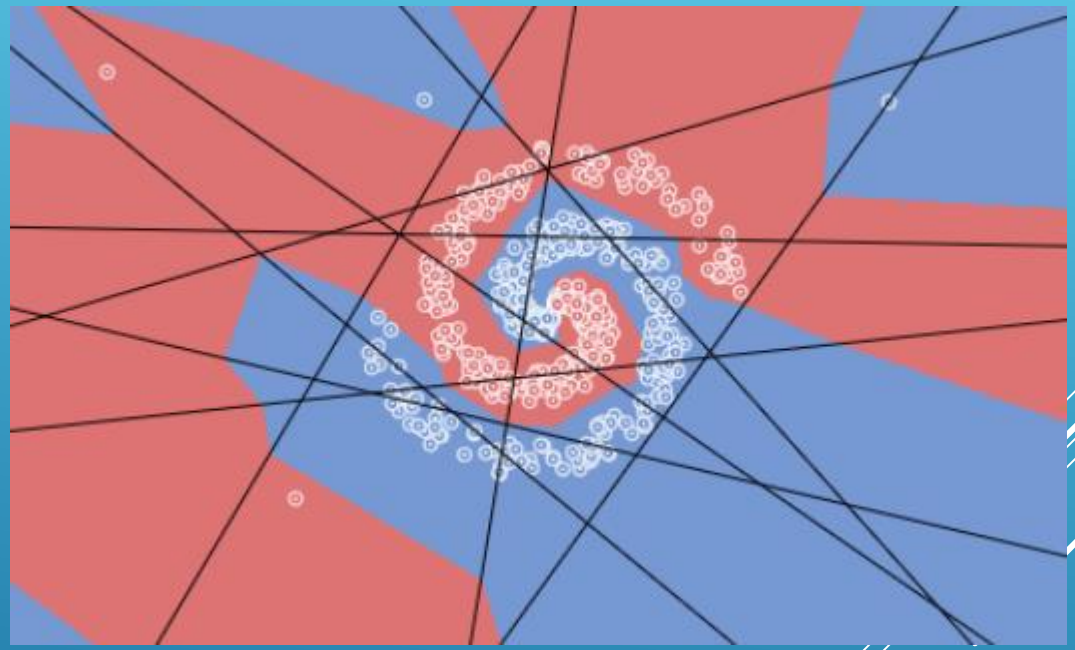


По материалам НП «Русское биометрическое общество»

Проблема экстраполяции MLP (дообучение, «отравление»)

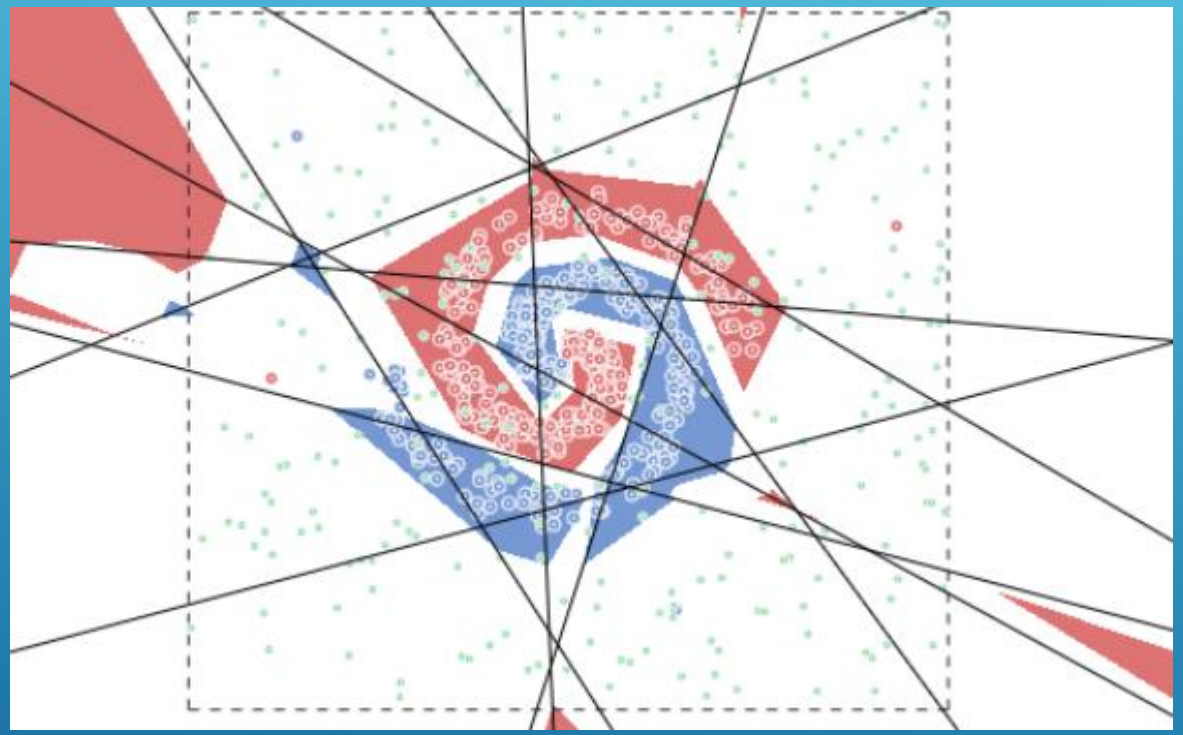


добавление обучающих примеров

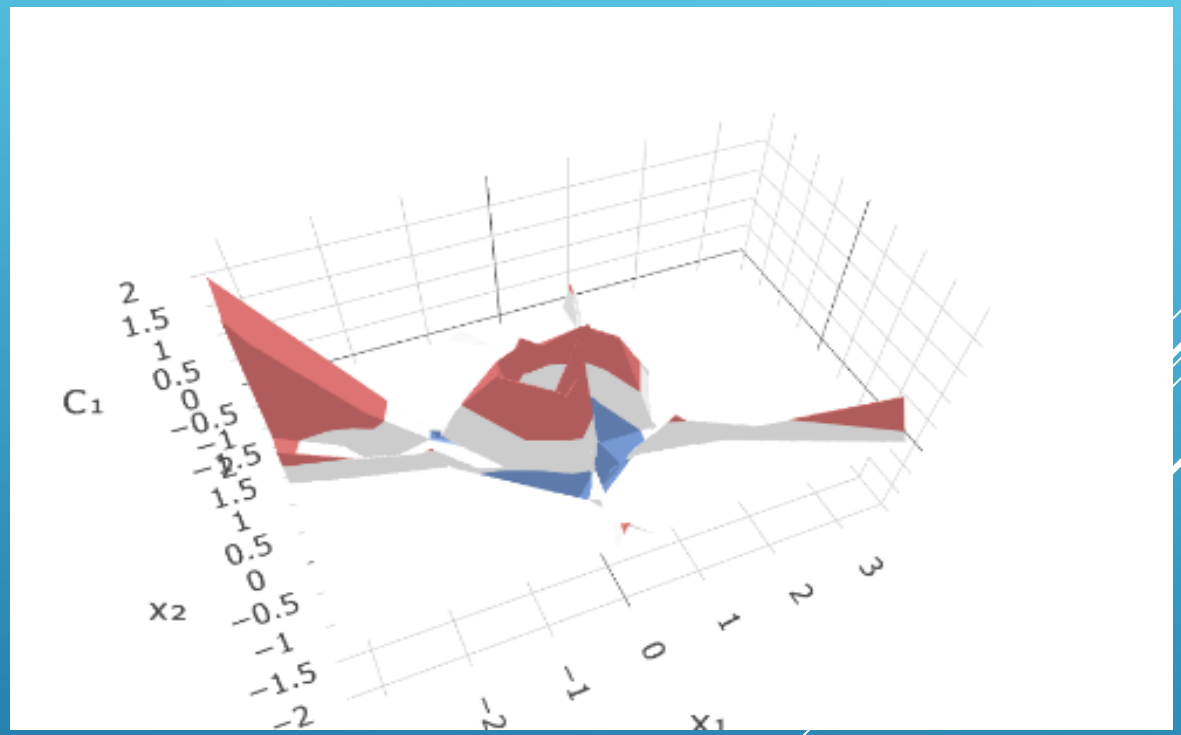


после дообучения

Проблема экстраполяции MLP: подход к решению

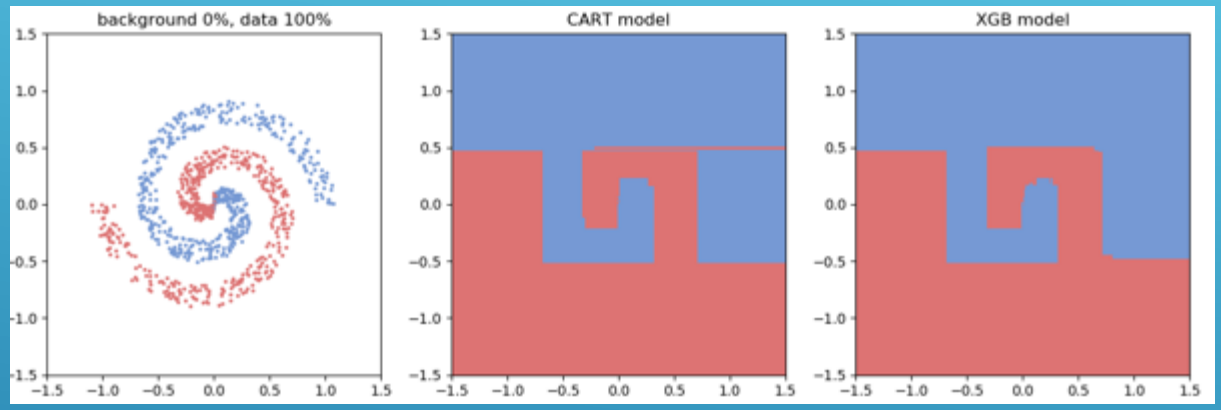


2D

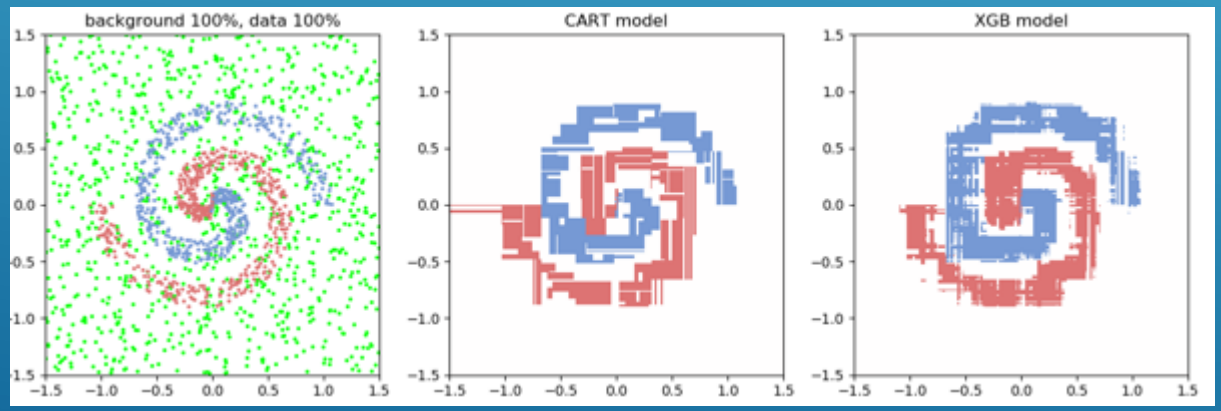


3D

Проблема экстраполяции деревьев решений CART и XGBoost

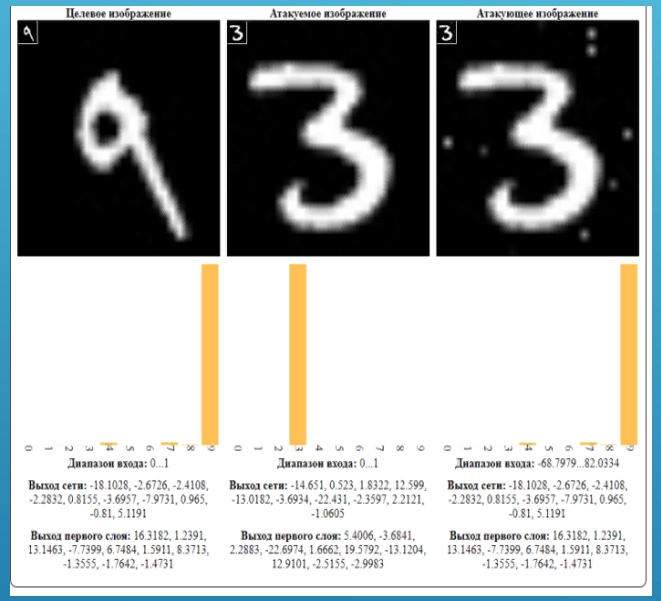


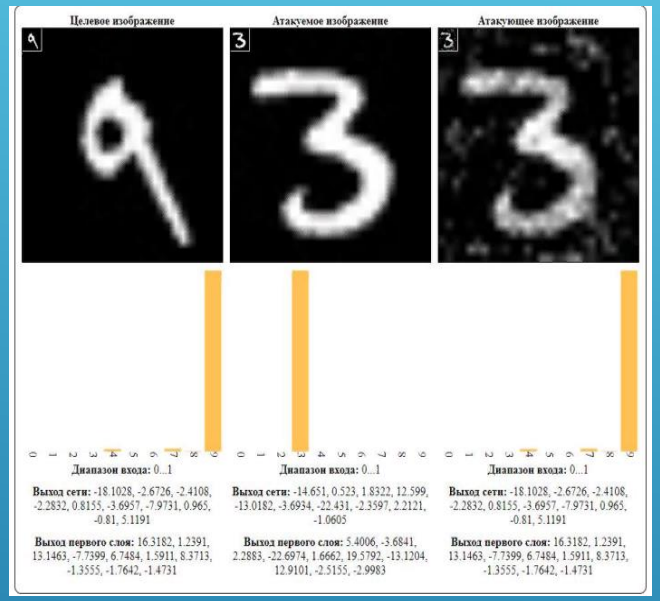
без «фона»

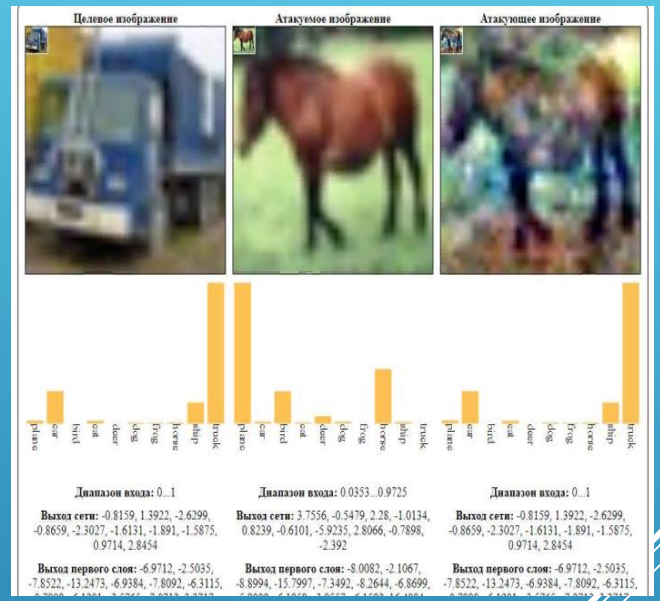


после добавления «фона»

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. Ставить и решать конкретные прикладные (специальные) задачи.

Стандарты оценки доверия системам машинного обучения

März 2024

DIN SPEC 92005

DIN

ICS 35.240.01

Künstliche Intelligenz –
Quantifizierung von Unsicherheiten im Maschinellen Lernen;
Text Englisch

Artificial Intelligence –
Uncertainty quantification in machine learning; Text in English

Intelligence Artificielle –
Quantification de l'incertitude dans l'apprentissage automatique; Texte en anglais

ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 70462.1— 2022/ ISO/IEC TR 24029-1—2021

Информационные технологии

ИНТЕЛЛЕКТ ИСКУССТВЕННЫЙ

Оценка робастности нейронных сетей

Часть 1

Обзор

(ISO/IEC TR 24029-1:2021, Artificial Intelligence (AI) —
Assessment of the robustness of neural networks — Part 1: Overview, IDT)

1 из 32

Издание официальное

Спасибо за внимание!