Лекция 1

Введение в машинное обучение

и другие хайповые слова

Машинное обучение

Андрей Фильченков / Сергей Муравьев

03.09.2020

План лекции

- Организационные вопросы
- Как устроено машинное обучение
- Обучение с учителем
- Проблема переобучения

- В презентации используются материалы курса «Машинное обучение» К.В. Воронцова
- Слайды доступны: **shorturl.at/ltVZ3**

План лекции

- **Организационные вопросы**
- Как устроено машинное обучение
- Обучение с учителем
- Проблема переобучения

Лаборатория машинного обучения

- Часть Центра Компьютерных Технологий
- Области исследований:
  - автоматическое машинное обучение
  - обработка и генерация изображений
  - профилирование пользователей и анализ социальных сетей
  - выбор признаков
  - маршрутизация
  - фундаментальные исследования
  - применение (медицина, анализ кода, финансы, производство)
  - …

План курса

- Обучение с учителем (4 лекции)
- Глубокое обучение (4 лекции)
- Обучение без учителя и обучение с подкреплением (5 лекций)

Как получить оценку?

- Практика — сдача лабораторных работ
- Теория —сдача теоретического минимума и экзамена
  Теоретический минимум теоретически можно закрыть контрольными
- Бонусные баллы

Формирование оценки (группы М3*3*)

$$Grade = \min(\sqrt{T \cdot P} + B, 100),$$

где

- $T \in [0; 120]$ — баллы за теоретическую часть
- $P \in [0; 120]$ — баллы за практическую часть
- $B \in [0; 40]$ — бонусные баллы

Теоретические баллы (группы М3*3*)

$$T = E + D,$$

- $E \in [0; 60]$ — экзамен
- $D \in [0; 60]$ — теоретический минимум

Практические баллы (группы М3*3*)

$$P = CF + D$$

- $CF \in [0; 60]$ — задачи на CodeForces, которые не требуется защищать. В конце осуществляется проверка на списывание.
- $D \in [0; 60]$ — защищаемые задачи:

$$D = \sum_i L_i ,$$

$$L_i = K \cdot (0.6 + 0.4 \cdot T),$$

$$T = 1 / (1 + w),$$

K — балл за задачу, получаемый студентом по результатам сдачи, T — временная поправка, w — количество недель после дедлайна

Формирование оценки (группы М330*)

$$Grade = \min(T + P + B, 100),$$

где

- $T \in [0; 50]$ — баллы за теоретическую часть
- $P \in [0; 60]$ — баллы за практическую часть
- $B \in [0; 40]$ — бонусные баллы

Теоретические баллы (группы М330*)

$$T = E + D,$$

- $E \in [0; 20]$ — экзамен
- $D \in [0; 30]$ — теоретический минимум

Практические баллы (группы М330*)

$$P = CF + D$$

- $CF \in [0; 30]$ — задачи на CodeForces, которые не требуется защищать. В конце осуществляется проверка на списывание.
- $D \in [0; 30]$ — защищаемые задачи:

$$D = \sum_i L_i ,$$

$$L_i = K \cdot (0.6 + 0.4 \cdot T),$$

$$T = 1 / (1 + w),$$

K — балл за задачу, получаемый студентом по результатам сдачи, T — временная поправка, w — количество недель после дедлайна

Бонусные баллы

Можно получить за:

- написание конспектов
- реализацию бэйслайнов из статей
- улучшение курса

Нельзя получить за:

- посещение лекций
- подкуп преподавателей
- дружеские отношения с преподавателями

Источники

1. Géron A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2017

2. Courville A., Goodfellow I., Bengio Y. Deep Learning, 2016

3. Николенко С.И., Кадурин А.А., Архангельская Е.В. Глубокое обучение. Погружение в мир нейронных сетей, 2017.

4. Flach P. Machine Learning: The Art and Science of Algorithms that Make Sense of Data, 2012

5. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: Data Mining, Inference, and Prediction, 2009

Онлайн курсы

Англоязычные курсы:

- A. Ng “Machine Learning” (https://clck.ru/QfQTP)
- G. Hinton “Neural Networks for Machine Learning” (https://clck.ru/QfQcE)
- D. Dye “Mathematics for Machine Learning specialization — Imperial College London” (https://www.coursera.org/specializations/mathematics-machine-learning)

Курсы на русском:

- К.В. Воронцов “Машинное обучение”. (https://www.youtube.com/playlist?list=PLJOzdkh8T5krxc4HsHbB8g8f0hu7973fK)
- С. Суворов, А. Янина, А. Сильвестров, Н. Капырин «Нейронные сети и обработка текста» (https://stepik.org/course/54098/promo)
- Е. Соколов, Специализация «Продвинутое машинное обучение» (https://ru.coursera.org/specializations/aml)

План лекции

- Организационные вопросы
- **Как устроено машинное обучение**
- Обучение с учителем
- Проблема переобучения

Определение машинного обучения

Машинное обучение это процесс, дающий компьютерам способность обучаться новому, не будучи непосредственно запрограммированными делать это.

A.L. Samuel Some Studies in Machine Learning Using the Game of Checkers // IBM Journal. July 1959. P. 210–229.

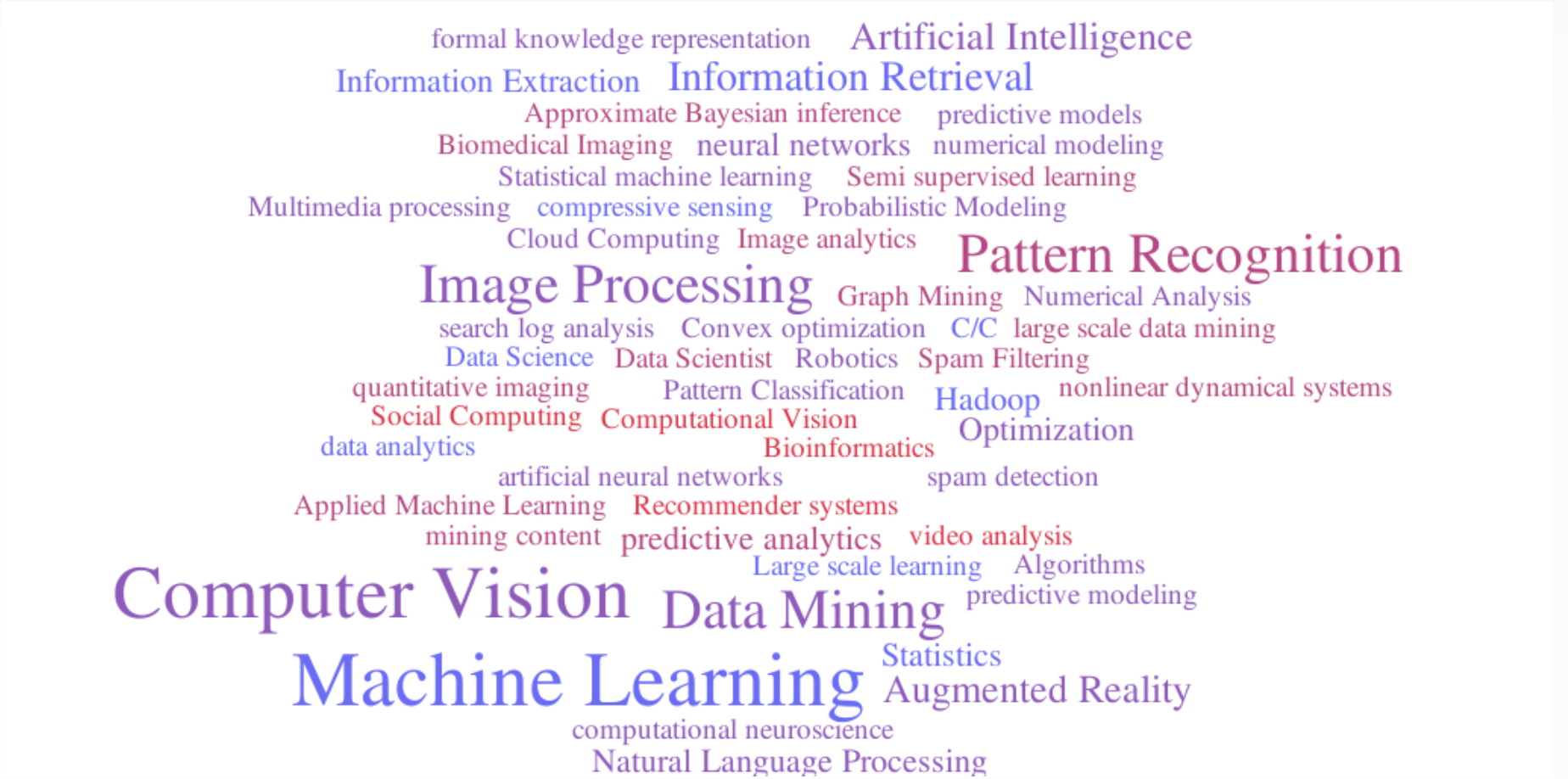
Программа **обучается** с опытом E решению некоторой задачи T по метрике качества P, если качество ее решения T, измеренное согласно P, растет вместе с ростом опыта E.

T.M. Mitchell Machine Learning. McGraw-Hill, 1997.

Подходы в машинном обучении

Применение машинного обучения

Связанные понятия



Связанные области

- Распознавание образов (pattern recognition)
- Машинное зрение (computer vision)
- Обработка естественного языка (natural language processing)
- Большие данные (Big Data)
- Информационный поиск (information retrieval)
- (Интеллектуальный) анализ данных / наука о данных (data mining, data science)
- …

Машинное обучение vs Data Mining

Формально, DM является одним из шагов в **извлечении знаний из баз данных (knowledge discovery in databases**) и включает в себя:

1. Сбор данных
2. Выделение признаков
3. Применение алгоритмов машинного обучения

Фактически, синонимично data analysis.

Машинное обучение vs Data Analysis

Ранее известно как «бизнес-аналитика»

1. Эксплораторный анализ данных (exploratory DA)
2. Конфирмационный анализ данных (confirmatory DA)
3. Предсказательный анализ данных
4. Визуализация данных

Машинное обучение vs Data Science

1. Сбор данных
2. Интеграция данных (data integration)
3. Хранение данных (data warehousing)
4. Анализ данных
5. Высокопроизводительные вычисления (high-performance computing)

Интеллектуальность и знания

- Искусственный интеллект (artificial intelligence)
- Интеллектуальные системы (intelligent systems)
- Математическое моделирование (mathematical modeling)

Искусственный интеллект

- Сейчас обычно говорят ИИ, когда имеют в виду машинное обучение
- Машинное обучение относится к искусственному интеллекту
- **Искусственный общий интеллект** (Artificial General Intelligence), ранее известный как **сильный искусственный интеллект** — более узкое понятие, связанное с достижением или превосходством человеческих когнитивных способностей

Интеллектуальные системы

- Искусственный интеллект
- Интеллектуальные системы

  Экспертные системы vs **системы на основе машинного обучения**
- Математическое моделирование

Математическое моделирование

- Искусственный интеллект
- Интеллектуальные системы
- **Математическое моделирование**

Знания и данные

Знания ≠ данные

Знания это закономерности в некоторой области (принципы, ограничения, отношения, правила, законы), получаемые в ходе профессиональной деятельности, которые позволяют формулировать и решать проблемы в этой области.

На чем покоится машинное обучение

- Теория вероятности и математическая статистика
- Теория оптимизации
- Вычислительные методы
- Линейная алгебра
- Дискретная математика
- Теория сложности алгоритмов
- …

Задачи машинного обучения

- обучение с учителем (supervised learning)
- обучение без учителя (unsupervised learning)
- частичное обучение (semi-supervised learning)
- обучение с подкреплением (reinforcement learning)
- активное обучение (active learning)
- онлайн-обучение (online learning)
- структурные предсказания (structured prediction)
- выбор и валидация модели (model selection and validation)

Обучение с учителем

Дано множество примеров с ответом, необходимо понять, как сопоставлять с ответами все возможные примеры:

- классификация;
- регрессия;
- ранжирование;
- прогнозирование.

Обучение без учителя

Дано множество примеров без ответов. Необходимо определить некоторые закономерности в данных:

- кластеризация;
- поиск ассоциативных правил;
- рекомендательные системы*;
- уменьшение размерности**.

План лекции

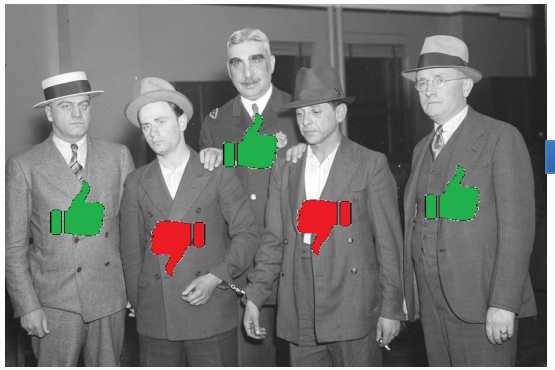
- Организационные вопросы
- Как устроено машинное обучение
- **Обучение с учителем**
- Проблема переобучения

Обучение с учителем

Обучение с учителем (supervised learning, обучение по прецедентам) является наиболее частой задачей машинного обучения

Предсказание

- Наука занимается осуществлением предсказаний
- В принципе, все машинное обучение занимается осуществлением предсказаний



Задача обучения с учителем

X — **множество объектов**;

Y — **множество меток (ответов)**;

$y : X \to Y$ неизвестная **целевая функция (зависимость)**.

$\mathcal{D} = \{(x_i, y_i)\}$ — размеченный **набор данных**,
где $\{x_1, \ldots, x_{|\mathcal{D}|}\} \subset X$ — объекты, а $y_i = y(x_i)$ — известные метки (значения целевой функции).

Задача:
найти $a : X \to Y$ **решающую (классифицирующую) функцию**, приближающую y на X.

В качестве a мы будем рассматривать **алгоритмы**.
В чем отличие между функцией и алгоритмом?

Основные вопросы

1. Что представляют собой объекты?
2. Что представляют собой метки?
3. Что представляет собой множество алгоритмов, из которого выбирается a?
4. Как оценить, насколько хорошо a приближает y?

Что представляют собой объекты?

$f_j : X \to D_j, j = 1, \ldots, n$ — **признаки (features, attributes) объектов**.

Типы признаков:

- **бинарный**: $D_j = \{0, 1\}$ (гендер в XVIII веке);
- **Категориальный (номинальный)**: D_j конечно (цвет);
- **порядковый (ординальный)**: D_j конечно и упорядочено (сорт муки);
- **численный (количественный)**: $D_j = \mathbb{R}$ (длина).

Табличные данные

$\big(f_1(x), \ldots, f_n(x)\big)$ — признаковое описание объекта x. Объект отождествляется с его признаковым описанием.

Данные часто представляются в табличном виде (матрица «объекты — признаки») :

$$F = \big\|f_j(x_i)\big\|_{|\mathcal{D}|\times n} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_{|\mathcal{D}|}) & \ldots & f_n(x_{|\mathcal{D}|}) \end{pmatrix}.$$

Что представляют собой ответы?

Для классификации:

- $Y = \{-1, +1\}$ — бинарная классификация (родился ли человек в СССР);
- $Y = \{1, \dots, M\}$, M непересекающихся классов (в какой стране человек родился);
- $Y = \{0, 1\}^M$, M пересекающихся классов (гражданином каких стран человек является).

Для ранжирования:

- Y — конечно (частично) упорядоченное множество (ранжирование стран по предпочтительности посещения).

Для регрессии:

- $Y = \mathbb{R}$ или $Y = \mathbb{R}^m$ (с какой вероятностью человек посетит Ирак / каждую из стран-членов ООН).

Что представляет собой множество алгоритмов, из которых a выбирается?

Предсказательная модель — параметрическое семейство отображений

$$A = \{M(x, \theta) | \theta \in \Theta\},$$

где $M: X \times \Theta \to Y$ некоторая функция, а Θ — множество возможных значений параметра θ.

Пример: для **линейной модели** вектор параметров таков: $\theta = (\theta_1, \dots, \theta_n)$, $\Theta = \mathbb{R}^n$.

Какую проблему машинного обучения мы решаем, если параметрическое семейство задано следующим образом:

$$M(x, \theta) = \sum_{j=1}^{n} \theta_j f_j(x)\,?$$

Метод обучения

Метод обучения — это отображение

$$\mu: (X \times Y)^{\dim} \to A,$$

возвращающее алгоритм $a \in A$ для заданного набора данных $\mathcal{D} \in (X \times Y)^{\dim}$,
где $(X \times Y)^{\dim} = \bigcup_{i \in \dim N \subseteq \mathbb{N}} (X \times Y)^i$

Метод обучения обычно состоит из двух частей:
1. Валидационный метод μ^{val}
2. Модель обучения μ^A

Две стадии работы с алгоритмами

1. **Обучение**:

Применяя метод обучения μ к набору данных $\mathcal{D}$, получаем на выходе обученный алгоритм

$$a = \mu(\mathcal{D}).$$

2. **Применение**:

Применяя a к новому объекту x, получаем на выходе предсказание его метки

$$a(x).$$

Как оценить, насколько хорошо a приближает y?

Функция потерь (loss function) $\mathcal{L}(a, x)$ — величина ошибки алгоритма a на объекте x

- для классификации

$$\mathcal{L}(a, x) = [a(x) \neq y(x)]$$

- для регрессии:

$$\mathcal{L}(a, x) = d\big(a(x) - y(x)\big),$$

чаще всего, квадратичная функция потерь:

$$d(x) = x^2, \mathcal{L}(a, x) = \big(a(x) - y(x)\big)^2.$$

Эмпирический риск — мера оценки качества a на $\mathcal{D}$:

$$\mathcal{L}(a, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathcal{L}(a, x).$$

Минимизация эмпирического риска

Метод минимизации эмпирического риска

$$\mu_{\text{EmpRM}}(\mathcal{D}) = \operatorname{argmin}_{a \in A} \mathcal{L}(a, \mathcal{D}).$$

Однако уменьшение ошибки на тренировочном множестве может привести к уменьшению обобщающей способности алгоритма.

План лекции

- Организационные вопросы
- Как устроено машинное обучение
- Обучение с учителем
- **Проблема переобучения**

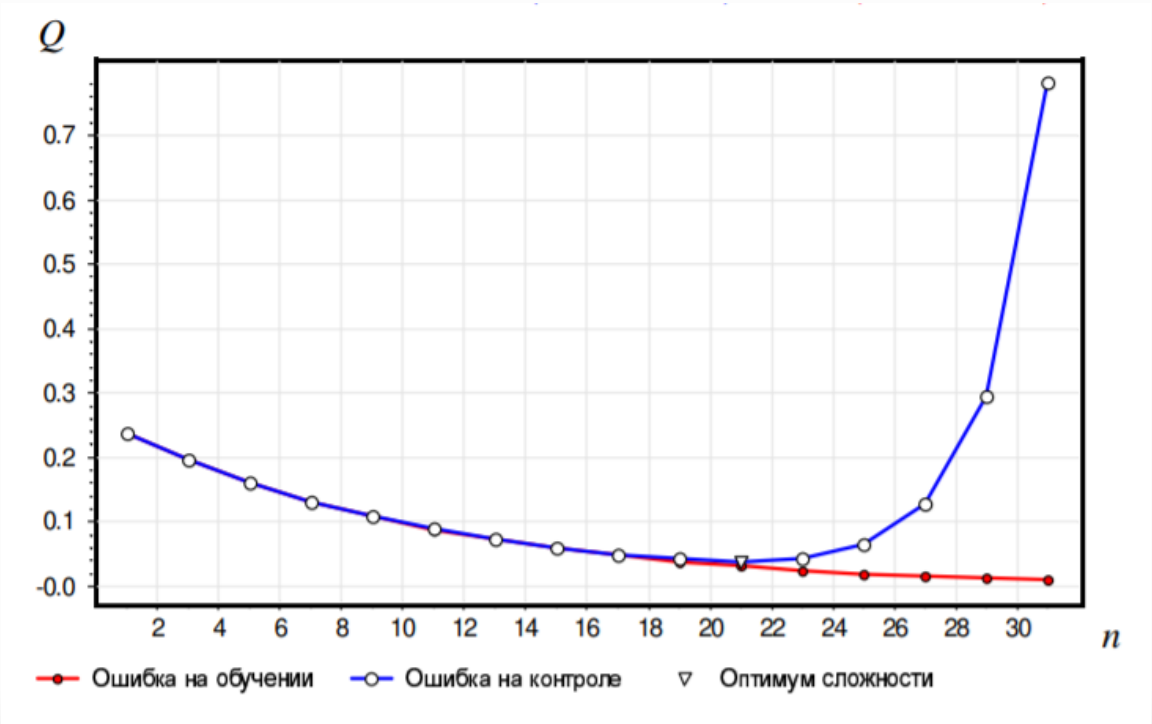
Проблема переобучения

Проблема переобучения — начиная с определенного уровня сложности предсказательной модели, чем лучше алгоритм показывает себя на тренировочном наборе данных $\mathcal{D}$, тем хуже он работает на реальных объектах.

Пример переобучения

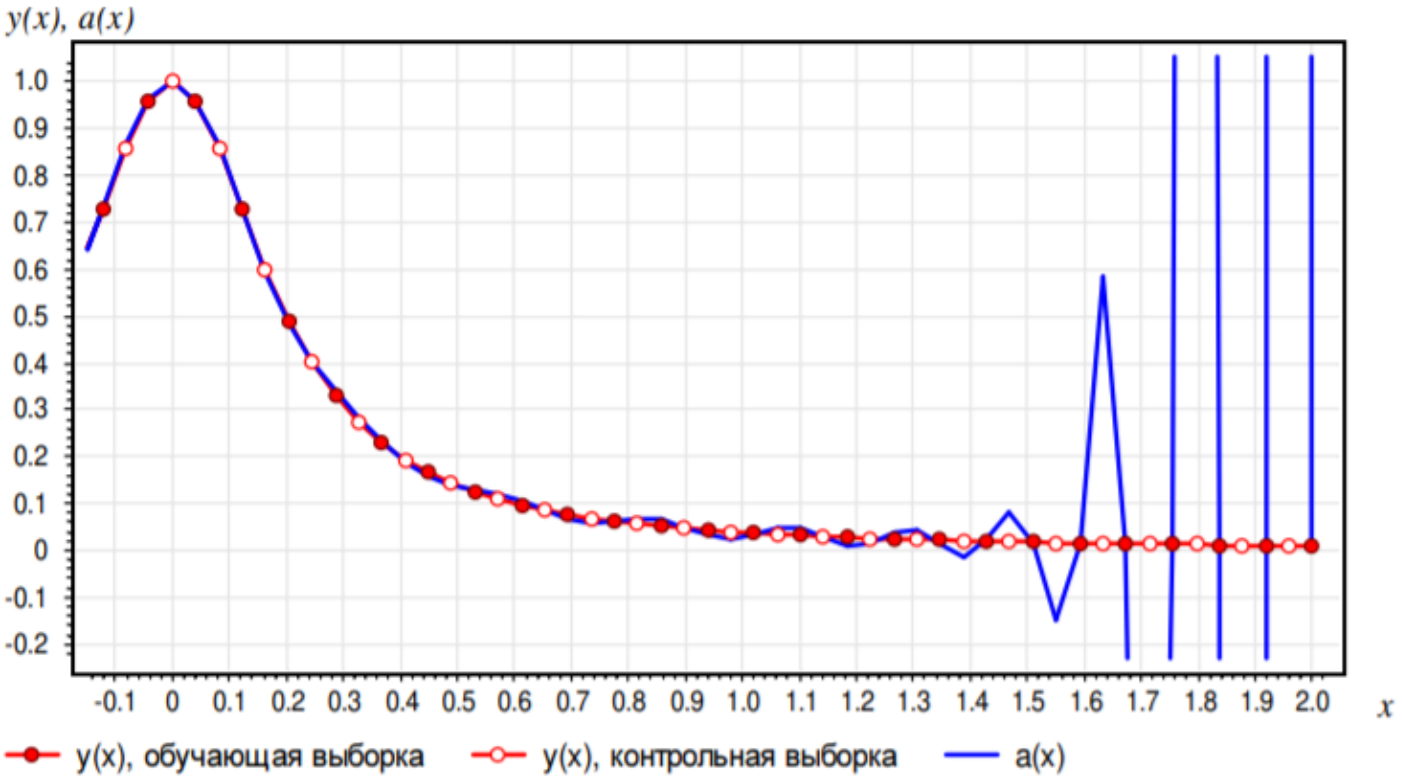
Зависимость $y(x) = \frac{1}{1+25x^2}$ на отрезке $x \in [-2, 2]$.

Будем искать приближение среди многочленов степени n — соответствует сложности модели.



Переобученный алгоритм

$y(x) = \dfrac{1}{1 + 25x^2}$; $a(x)$ — многочлен степени $n = 38$



В следующей серии

- Как бороться с переобучением
- Как отличать уток от не уток
- Как измерять качество модели