Введение в машинное обучение

Воронцов Константин Вячеславович
МФТИ • ФИЦ ИУ РАН • ШАД Яндекс • Forecsys • Aithea

Школа глубокого обучения

кружок для старшеклассников

4 ноября 2017 • МФТИ

Машинное обучение — новый двигатель прогресса

«Четвёртая технологическая революция строится на вездесущем и мобильном Интернете, искусственном интеллекте и машинном обучении» (2016)

Клаус Мартин Шваб, президент Всемирного экономического форума

Мир наконец поверил в искусственный интеллект?...
Машинное обучение изменит мир? Или уже меняет?

Бум искусственного интеллекта и нейронных сетей

1997 IBM Deep Blue обыграл чемпиона мира по шахматам

2005 Беспилотный автомобиль: DARPA Grand Challenge

2006 Google Translate – статистический машинный перевод

2011 40 лет DARPA CALO привели к созданию Apple Siri

2011 IBM Watson победил в ТВ-игре «Jeopardy!»

2011–2015 ImageNet: 25% → 3.5% ошибок против 5% у людей

2012 Google X Lab: распознавание видеокадров с котами

2014 Facebook DeepFace распознаёт лица с точностью 97%

2015 Фонд OpenAI в $1 млрд. Илона Маска и Сэма Альтмана

2016 DeepMind, OpenAI: динамическое обучение играм Atari

2016 Google DeepMind обыграл чемпиона мира по игре го

2017 OpenAI обыграл чемпиона мира по компьютерной игре Dota 2

Задача статистического (машинного) обучения с учителем

Задача восстановления зависимости $y(x)$
по точкам *обучающей выборки* (x_i, y_i), $i = 1, \ldots, \ell$.

Дано: векторы объектов $x_i = (x_i^1, \ldots, x_i^n)$, ответы $y_i = y(x_i)$:

$$\begin{pmatrix} x_1^1 & \ldots & x_1^n \\ \ldots & \ldots & \ldots \\ x_\ell^1 & \ldots & x_\ell^n \end{pmatrix} \xrightarrow{y} \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}$$

Найти: функцию $a(x)$, способную давать ответы на *тестовых объектах* $\tilde{x}_i = (\tilde{x}_i^1, \ldots, \tilde{x}_i^n)$, $i = 1, \ldots, k$:

$$\begin{pmatrix} \tilde{x}_1^1 & \ldots & \tilde{x}_1^n \\ \ldots & \ldots & \ldots \\ \tilde{x}_k^1 & \ldots & \tilde{x}_k^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \ldots \\ a(\tilde{x}_k) \end{pmatrix}$$

Типы признаков и типы задач

Типы признаков, $x_i^j \in D_j$, в зависимости от множества D_j:

- $D_j = \{0, 1\}$ — *бинарный* признак;
- $|D_j| < \infty$ — *номинальный* признак;
- D_j упорядочено — *порядковый* признак;
- $D_j = \mathbb{R}$ — *количественный* признак.

Типы задач, $y_i \in Y$, в зависимости от множества Y:

- $Y = \{0, 1\}$ или $Y = \{-1, +1\}$ — *классификация* на 2 класса;
- $Y = \{1, \dots, M\}$ — на M непересекающихся классов;
- $Y = \{0, 1\}^M$ — на M классов, которые могут пересекаться;
- $Y = \mathbb{R}$ — задача восстановления регрессии;
- Y упорядочено — задача ранжирования (learning to rank).

Задачи медицинской диагностики

Объект — пациент в определённый момент времени.

Классы — диагноз или способ лечения или исход заболевания.

Примеры признаков:

- **бинарные:** пол, головная боль, слабость, тошнота, и т. д.
- **порядковые:** тяжесть состояния, желтушность, и т. д.
- **количественные:** возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т. д.

Особенности задачи:

- обычно много «пропусков» в данных;
- как правило, недостаточный объём данных;
- нужен интерпретируемый алгоритм классификации;
- нужна оценка вероятности (риска | успеха | исхода).

Задачи распознавания месторождений

Объект — геологический район (рудное поле).

Классы — есть или нет полезное ископаемое.

Примеры признаков:

- **бинарные:** присутствие крупных зон смятия и рассланцевания, и т. д.
- **порядковые:** минеральное разнообразие; мнения экспертов о наличии полезного ископаемого, и т. д.
- **количественные:** содержания сурьмы, присутствие в рудах антимонита, и т. д.

Особенности задачи:

- проблема «малых данных» — для редких типов месторождений объектов много меньше, чем признаков.

Задачи биометрической идентификации личности

Идентификация по отпечаткам пальцев

Идентификация по радужной оболочке глаза

Особенности задач:

- нетривиальная предобработка для извлечения признаков;
- высочайшие требования к точности.

Задача ранжирования поисковой выдачи

Объект — пара $\langle$короткий запрос, документ$\rangle$.

Классы — асессорские оценки релевантности.

Примеры признаков:

- **количественные:**
  частота слов запроса в документе,
  число ссылок на документ,
  число кликов на документ: всего, по данному запросу,
- **номинальные:**
  ID пользователя, ID региона, язык запроса.

Особенности задачи:

- оптимизируется не число ошибок, а качество ранжирования;
- сверхбольшие выборки;
- проблема конструирования признаков по сырым данным.

Задача ранжирования в рекомендательных системах

Объект — пара $\langle$клиент, товар$\rangle$
(товары — книги, фильмы, музыка).

Предсказать: вероятность покупки или рейтинг товара.

Примеры признаков:

- **количественные:**
  рейтинг схожих товаров для данного клиента;
  рейтинг данного товара для схожих клиентов;
  вектор интересов клиента;
  вектор интересов товара;

Особенности задачи:

- сверхбольшие разреженные данные;
- интересы скрыты, их надо сначала выявить.

Задача прогнозирования стоимости недвижимости

Объект — квартира в Москве.

Примеры признаков:

- **бинарные:** наличие балкона, лифта, мусоропровода, охраны, и т. д.
- **номинальные:** район города, тип дома (кирпичный/панельный/блочный/монолит), и т. д.
- **количественные:** число комнат, жилая площадь, расстояние до центра, до метро, возраст дома, и т. д.

Особенности задачи:

- выборка неоднородна, стоимость меняется со временем;
- разнотипные признаки;
- для линейной модели нужны преобразования признаков.

Задача прогнозирования объёмов продаж

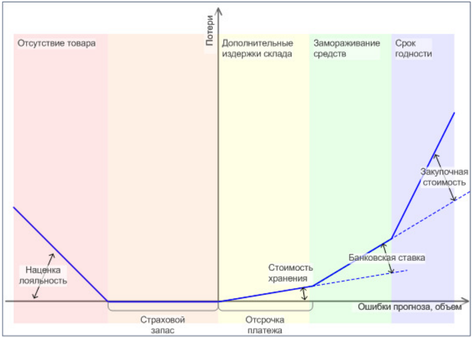
Объект — тройка ⟨товар, магазин, день⟩.

Примеры признаков:

- **бинарные:** выходной день, праздник, промоакция, и т. д.
- **количественные:** объёмы продаж в предшествующие дни.

Особенности задачи:

- функция потерь не квадратична и даже не симметрична;
- разреженные данные.



Метод наименьших квадратов (Гаусс, 1795)

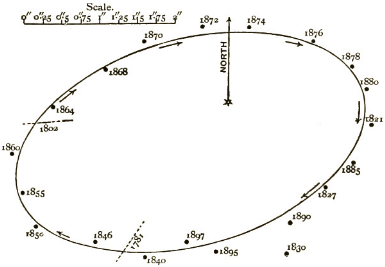
Линейная модель регрессии:

$$a(x, w) = \sum_{j=1}^{n} w_j x^j, \qquad w \in \mathbb{R}^n.$$

Метод наименьших квадратов:

$$Q(w) = \sum_{i=1}^{\ell} \bigl(a(x_i, w) - y_i\bigr)^2 \to \min_{w}.$$

Карл Фридрих Гаусс (1777–1855)

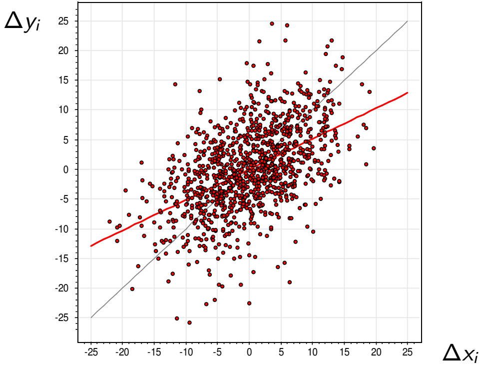


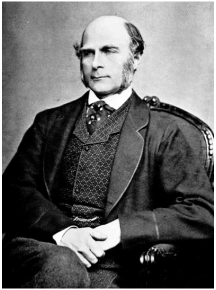
«Our principle, which we have made use of since 1795, has lately been published by Legendre...»

C.F.Gauss. Theory of the motion of the heavenly bodies moving about the Sun in conic sections. 1809.

Откуда пошло название «регрессия» (Гальтон, 1886)

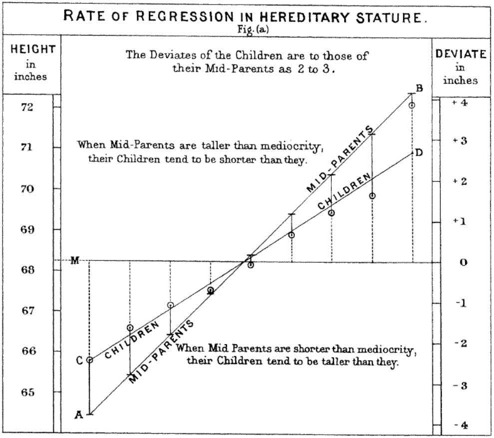
Исследование наследственности роста.
отклонение роста от среднего в популяции:
Δx_i — отклонение роста отца
Δy_i — отклонение роста взрослого сына

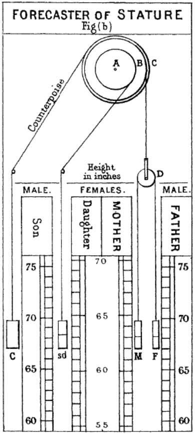




Фрэнсис Гальтон
(1822–1911)

Скрытый смысл: «регрессия» — сначала данные, потом модель





Galton F. Regression towards mediocrity in hereditary stature. 1886.

Общие подходы к решению оптимизационных задач

Аналитический подход (напр. метод наименьших квадратов):
Если w — точка минимума *гладкой* функции $Q(w)$, то

$$\frac{\partial Q(w)}{\partial w_j} = 0, \quad j = 1, \dots, n.$$

Это система n уравнений с n неизвестными.

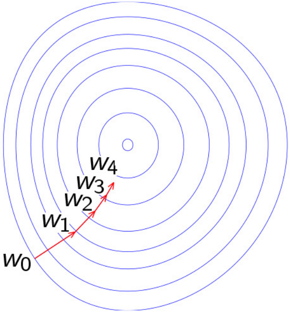
Численный метод — градиентный спуск:

начальное приближение w^0, $t := 0$;
повторять

$$w_j^{t+1} := w_j^t - h^t \cdot \frac{\partial Q(w^t)}{\partial w_j}, \quad j = 1, \dots, n;$$

$$t := t + 1;$$

пока процесс не сойдётся;



Задача проведения прямой через заданные точки

Дано: $x_i, y_i \in \mathbb{R}, \quad i = 1, \dots, \ell$

Найти: параметры $w = (\alpha, \beta)$ линейной модели $y = \alpha x + \beta$

Критерий: $Q(\alpha, \beta) = \sum\limits_{i=1}^{\ell} (\alpha x_i + \beta - y_i)^2 \ \to\ \min$

Аналитический метод решения:

$$\frac{\partial Q}{\partial \alpha} = 0 \quad \Rightarrow \quad \alpha \sum_{i=1}^{\ell} x_i^2 + \beta \sum_{i=1}^{\ell} x_i - \sum_{i=1}^{\ell} x_i y_i = 0$$

$$\frac{\partial Q}{\partial \beta} = 0 \quad \Rightarrow \quad \alpha \sum_{i=1}^{\ell} x_i + \beta \sum_{i=1}^{\ell} 1 - \sum_{i=1}^{\ell} y_i = 0$$

Это система линейных уравнений 2×2:

$$\begin{cases} \alpha S_{xx} + \beta S_x = S_{xy} \\ \alpha S_x + \beta S_1 = S_y \end{cases} \quad \Rightarrow \quad \begin{cases} \alpha = \frac{S_{xy} S_1 - S_x S_y}{S_{xx} S_1 - S_x S_x} \\ \beta = \frac{S_{xx} S_y - S_{xy} S_x}{S_{xx} S_1 - S_x S_x} \end{cases}$$

Метод стохастического градиента (SG, Stochastic Gradient)

Линейная модель регрессии:

$$a(x, w) = \langle w, x \rangle = \sum_{j=1}^{n} w_j x^j, \qquad w \in \mathbb{R}^n.$$

Метод наименьших квадратов:

$$Q(w) = \sum_{i=1}^{\ell} \bigl(a(x_i, w) - y_i\bigr)^2 \to \min_w .$$

Один шаг *градиентного спуска*:

$$w_j^{t+1} := w_j^t - h^t \sum_{i=1}^{\ell} \bigl(a(x_i, w^t) - y_i\bigr)\, x_i^j .$$

Идея ускорения сходимости: брать (x_i, y_i) по одному в случайном порядке и сразу обновлять вектор весов,

$$w_j^{t+1} := w_j^t - h^t \bigl(a(x_i, w^t) - y_i\bigr)\, x_i^j .$$

Алгоритм SG (Stochastic Gradient)

Вход: выборка $x_i = (x_i^1, \ldots, x_i^n)$, y_i, $i = 1, \ldots, \ell$;
Выход: веса $w_1, \ldots, w_n$;

инициализировать веса w_j, $j = 1, \ldots, n$;
повторять
  выбрать случайный объект (x_i, y_i) из обучающей выборки;
  выбрать величину градиентного шага h;
  выполнить градиентный шаг:
  $w_j := w_j - h\big(a(x_i, w^t) - y_i\big)\, x_i^j$ для всех $j = 1, \ldots, n$;
пока процесс не сойдётся куда-нибудь;

Преимущества и недостатки:

⊕ можно брать не только линейные модели
⊕ можно брать не только квадратичную функцию потерь
⊕ хорошо работает на больших выборках
⊖ возможно застревание в локальных экстремумах

Эвристики

- Выбор начального приближения, например, так:

$$w_j^0 := \frac{\langle y, x^j \rangle}{\langle x^j, x^j \rangle} \quad \text{(из одномерной линейной регрессии)}$$

  $x^j = (x_i^j)_{i=1}^{\ell}$ — вектор значений j-го признака,
  $y = (y_i)_{i=1}^{\ell}$ — вектор ответов.

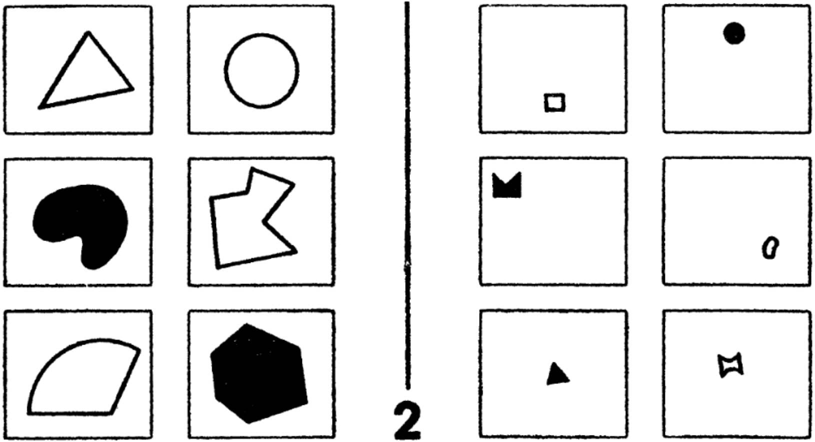
- Выбор темпа обучения (градиентного шага) h^t:
  сходимость гарантируется для выпуклых $Q(w)$ при

$$h^t \to 0, \quad \sum_{t=1}^{\infty} h^t = \infty, \quad \sum_{t=1}^{\infty} (h^t)^2 < \infty,$$

  в частности можно положить $h^t = \frac{1}{t}$;

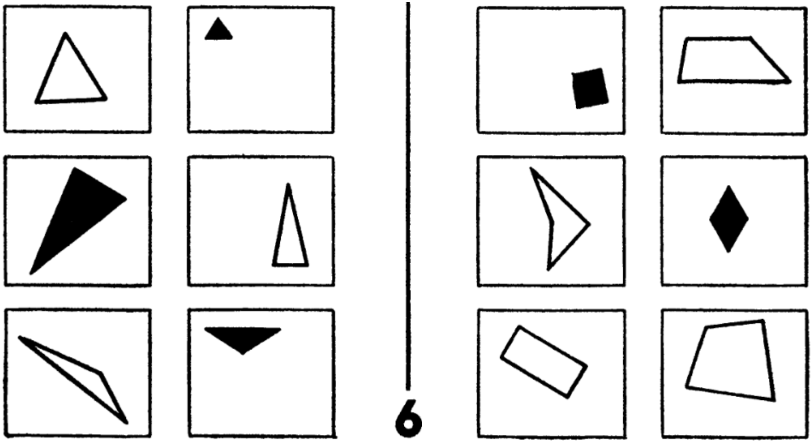
- Время от времени делать большие случайные шаги;
- Мультистарт.

Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Обучающая выборка: по 6 объектов каждого из двух классов.
Требуется найти правило классификации.

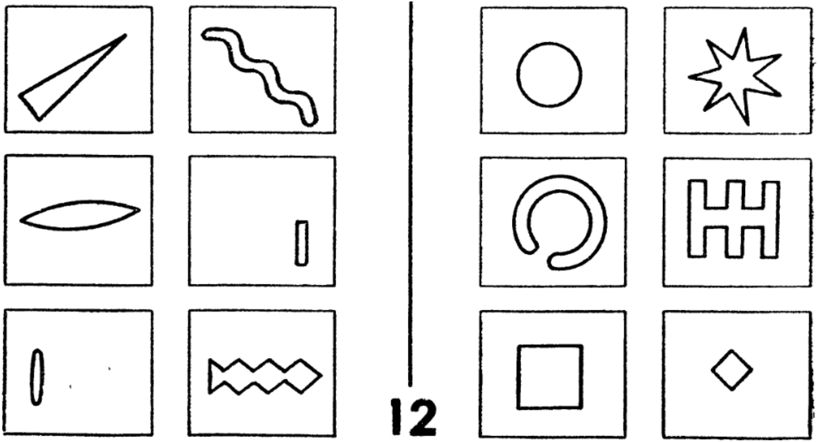
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Что даёт нам уверенность, что мы нашли верное правило?

1. Безошибочная классификация примеров обучающей выборки.

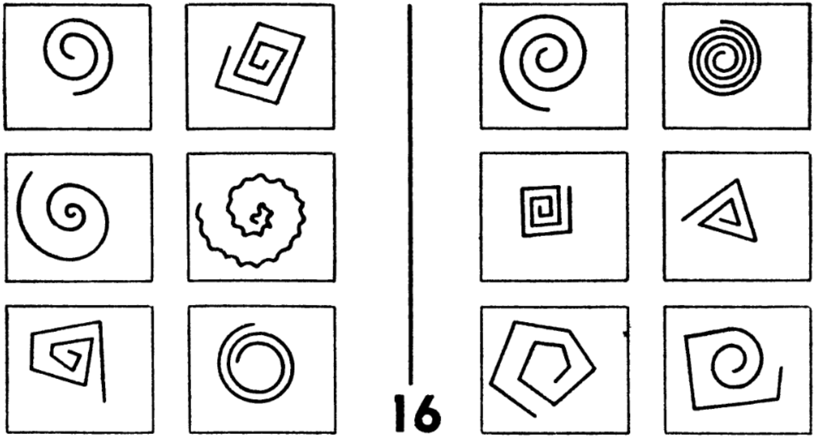
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Что даёт нам уверенность, что мы нашли верное правило?

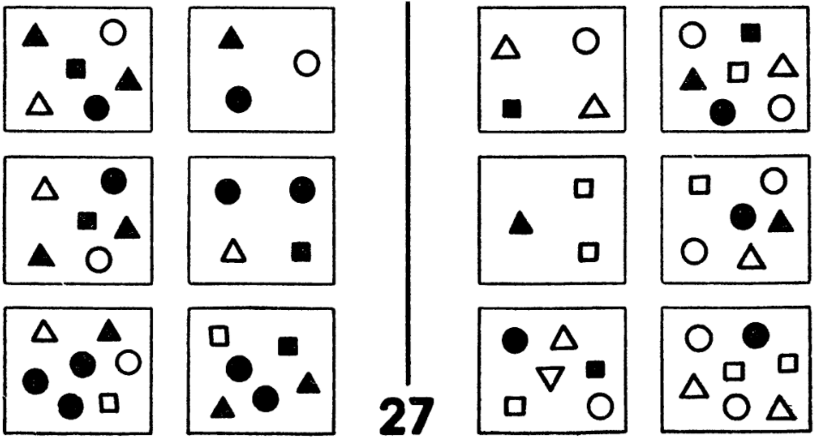
2. Простота и определённое «изящество» найденного правила.

Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Мы решаем эти задачи почти мгновенно. Чем мы пользуемся? Почему для компьютера они столь сложны?

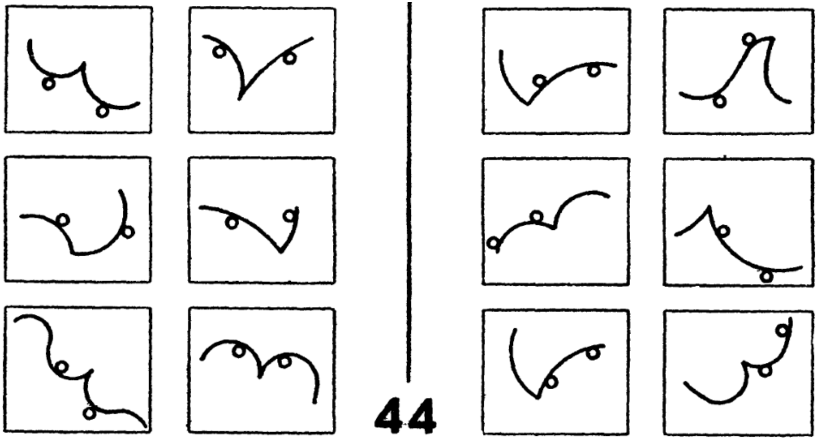
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



27

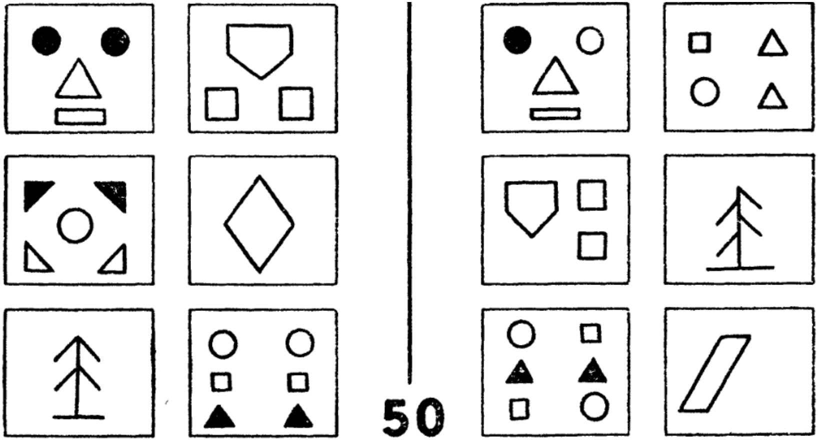
Нужно ли закладывать знания геометрии в явном виде?
Или возможно выучить геометрические понятия на примерах?

Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Как вычислять полезные признаки по сложным сырым данным?
Возможно ли поручить перебор признаков и моделей машине?

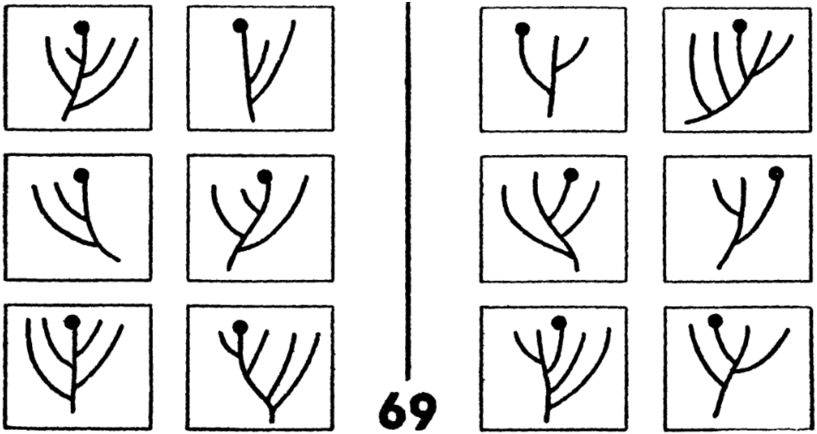
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Каков риск выбрать по данным неверное правило, *предрассудок*?

Как этот риск зависит от числа примеров и сложности правил?

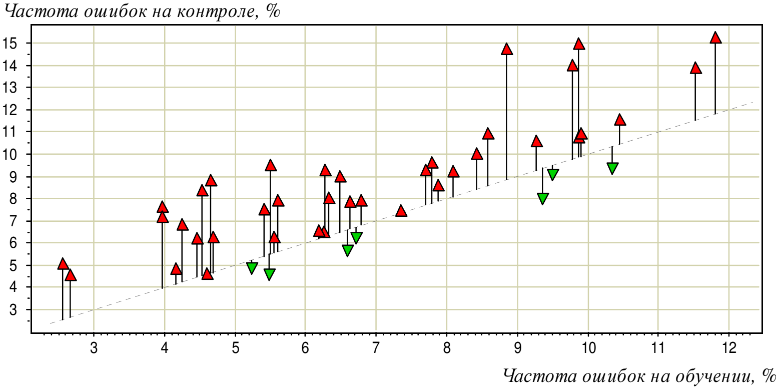
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Эти вопросы составляют основу машинного обучения сегодня.
М.М.Бонгард поставил все эти проблемы в середине 60-х!

Пример. Переобучение в задаче медицинской диагностики

Задача предсказания отдалённого результата хирургического лечения атеросклероза. Точки — различные алгоритмы.



Имеется систематическое смещение точек выше биссектрисы

Проблема переобучения. Пример

Зависимость $y(x) = \dfrac{1}{1+25x^2}$ на отрезке $x \in [-2, 2]$.
Признаковое описание объекта x: $(1, x^1, x^2, \dots, x^n)$.

Модель полиномиальной регрессии

$$a(x, w) = w_0 + w_1 x + \dots + w_n x^n \quad \text{— полином степени } n.$$

Метод наименьших квадратов:

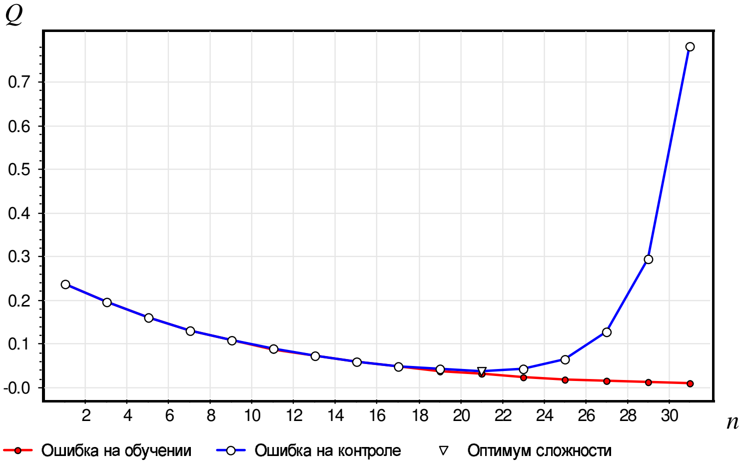
$$Q(w, X^\ell) = \sum_{i=1}^{\ell} (w_0 + w_1 x_i + \dots + w_n x_i^n - y_i)^2 \to \min_{w_0,\dots,w_n}.$$

Обучающая выборка: $X^\ell = \bigl\{x_i = 4\frac{i-1}{\ell-1} - 2 \bigm| i = 1, \dots, \ell\bigr\}$.
Контрольная выборка: $X^k = \bigl\{x_i = 4\frac{i-0.5}{\ell-1} - 2 \bigm| i = 1, \dots, \ell-1\bigr\}$.

Что происходит с $Q(w^*, X^\ell)$ и $Q(w^*, X^k)$ с ростом n?

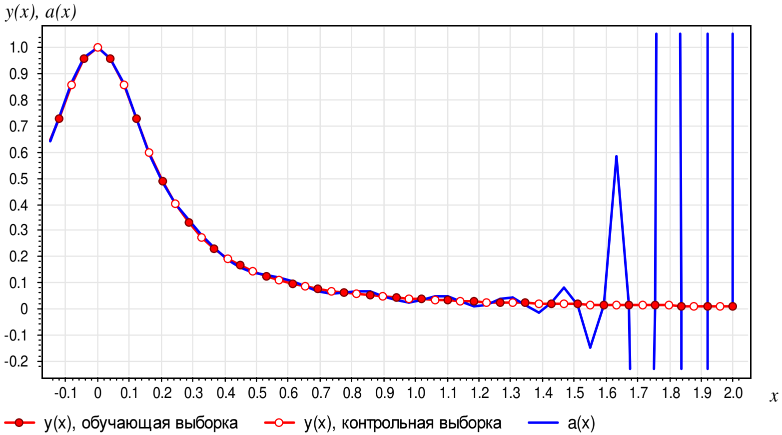
Пример переобучения: эксперимент при $\ell = 50$, $n = 1..31$

Переобучение — это когда $Q(w^*, X^k) \gg Q(w^*, X^\ell)$:



Пример переобучения: эксперимент при $\ell = 50$

$$y(x) = \frac{1}{1+25x^2}; \quad a(x) \text{ — полином степени } n = 38$$



Переобучение — ключевая проблема машинного обучения

Линейная зависимость (мультиколлинеарность) признаков:

- пусть построена модель: $a(x, w) = \sum\limits_j w_j x^j$

  и оказалось, что $\sum\limits_j \beta_j x^j = 0$ для всех x, при некотором β
- тогда задача неустойчива, решений бесконечно много:
  $a(x, w) = \sum\limits_j (w_j + \gamma\beta_j) x^j$ для любого γ

Проявления переобучения:

- слишком большие веса $|w_j|$ разных знаков
- $Q(X^\ell) \ll Q(X^k)$

Методы уменьшения переобучения:

- регуляризация — ограничения на w
- трансформация признаков (метод главных компонент)
- отбор признаков

Основные школы машинного обучения

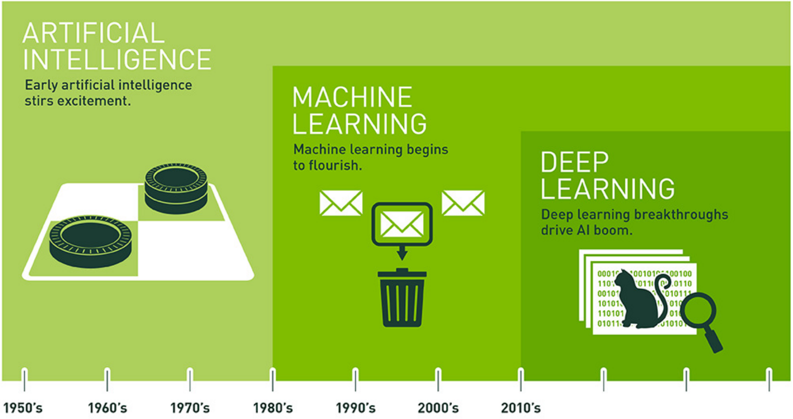
1. *символизм* – поиск логических закономерностей
   - Decision Tree, Rule Induction
2. *коннекционизм* – обучаемые нейронные сети
   - BackPropagation, Deep Belief Nets, Deep Learning
3. *эволюционизм* – саморазвитие сложных моделей
   - Genetic Algorithms, Genetic Programming
4. *байесионизм* – оценивание распределений параметров
   - Naive Bayes, Bayesian Networks, Graphical Models
5. *аналогизм* – «близким объектам близкие ответы»
   - kNN, RBF, SVM, Kernel Smoothing

⊕ *композиционизм* – кооперация моделей
   - Weighted Voting, Boosting, Bagging, Stacking, Random Forest, Яндекс.MatrixNet

Домингос П. Верховный алгоритм. 2016. 336 с.

Будущее машинного обучения

Вытеснит ли глубокое обучение все остальные методы?
Это «грубая сила» или новый способ моделирования?
Возможно ли заменить моделирование вычислениями?



Полезные ссылки

- www.kaggle.com — конкурсы анализа данных
- www.kdnuggets.com — главный сайт датамайнеров
- www.MachineLearning.ru — русскоязычная вики
- www.datasciencecentral.com — 72 000 датамайнеров
- archive.ics.uci.edu/ml — UCI ML Repository (349 datasets)
- ru.coursera.org/learn/machine-learning — курс Эндрю Ына
- ru.coursera.org/learn/vvedenie-mashinnoe-obuchenie — курс Воронцова от ВШЭ и ШАД Яндекс
- ru.coursera.org/specializations/machine-learning-data-analysis — специализация от МФТИ и ШАД Яндекс

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov