

История машинного обучения

Воронцов Константин Вячеславович

vokov@forecsys.ru

<http://www.MachineLearning.ru> «User:Vokov»

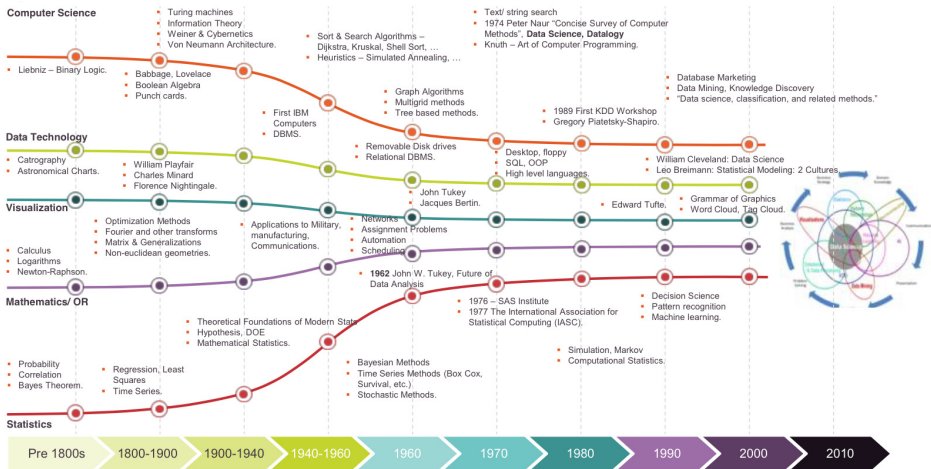
МФТИ • 4 мая 2017

- 1 Предпосылки. Докомпьютерная эпоха**
 - О какой науке мы говорим
 - Эмпирическая индукция и статистика
 - Искусственный нейрон
- 2 Основные вехи и крупные прорывы**
 - Нейронные сети и композиции моделей
 - Алгоритмы поиска закономерностей в данных
 - Теория статистического обучения и регуляризация
- 3 Современные задачи и направления исследований**
 - Состояние науки машинного обучения
 - Бум искусственного интеллекта и машинного обучения
 - Методология решения прикладных задач

Разные названия то ли одной большой науки, то ли её частей

- Статистический анализ данных (Statistical Data Analysis)
- Искусственный интеллект (Artificial Intelligence) — 1955
- Распознавание образов (Pattern Recognition)
- **Машинное обучение (Machine Learning)** — 1959
- Статистическое обучение (Statistical Learning)
- Интеллектуальный анализ данных (Data Mining) — 1989
- Knowledge Discovery in Databases — 1989
- Науки о данных (Data Science) — 1997
- Бизнес-аналитика (Business Intelligence, Business Analytics)
- Предсказательная аналитика (Predictive Analytics) — 2007
- Большие данные (Big Data) — 2008
- Аналитика больших данных (Big Data Analytics)

Предпосылки Data Science



<http://www.kdnuggets.com/2015/02/history-data-science-infographic.html>

Принцип эмпирической индукции

«Не следует полагаться на сформулированные аксиомы и формальные базовые понятия, какими бы привлекательными и справедливыми они не казались. Законы природы нужно «расшифровывать» из фактов опыта. Следует искать правильный метод анализа и обобщения опытных данных; здесь логика Аристотеля не подходит в силу её абстрактности, оторванности от реальных процессов и явлений.»



Фрэнсис Бэкон
(1561–1626)

Таблицы открытия: множество случаев x , когда

- свойство *у присутствовало* $y(x) = 1$
- свойство *у отсутствовало* $y(x) = 0$
- наблюдалось изменение *степени* свойства $y(x)$

Фрэнсис Бэкон. Новый органон. 1620.

Восстановление зависимостей по эмпирическим данным

Дано:

объекты $x_i = (f_1(x_i), \dots, f_n(x_i))$ и ответы $y_i = y(x_i)$, $i = 1, \dots, \ell$
 $f_j(x)$ — признаки объекта x , $j = 1, \dots, n$

Найти:

функцию $a(x, w)$, восстанавливающую зависимость $y(x)$

Критерий: минимум эмпирического риска

$$\sum_{i=1}^{\ell} \mathcal{L}(a(x_i, w), y_i) \rightarrow \min_w,$$

где $\mathcal{L}(a, y)$ — функция потерь от ошибки a при ответе y .

Основные типы задач обучения с учителем:

- регрессия: $y_i \in \mathbb{R}$, $\mathcal{L}(a, y) = (a - y)^2$
- классификация: $y_i \in \{0, 1\}$, $\mathcal{L}(a, y) = [a \neq y]$

Метод наименьших квадратов (Гаусс, 1795)

Линейная модель регрессии:

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

Функционал квадрата ошибки:

$$\sum_{i=1}^{\ell} (a(x_i, w) - y_i)^2 = \|Fw - y\|^2 \rightarrow \min_w.$$

Решение системы: $w^* = (F^T F)^{-1} F^T y$.

$$F_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \quad y_{\ell \times 1} = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}, \quad w_{n \times 1} = \begin{pmatrix} w_1 \\ \dots \\ w_n \end{pmatrix}.$$

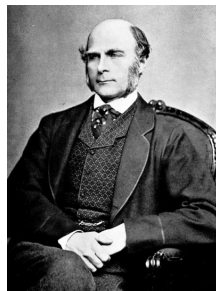
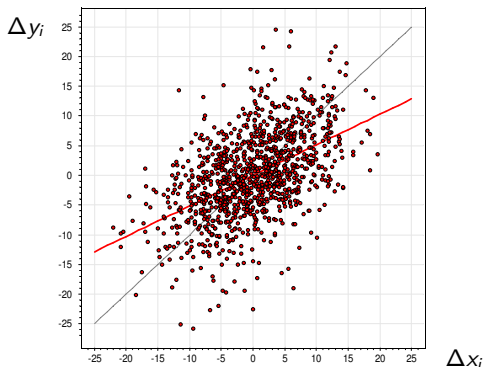


Карл Фридрих
Гаусс (1777–1855)

«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.

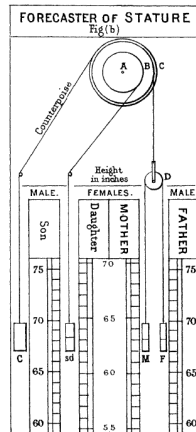
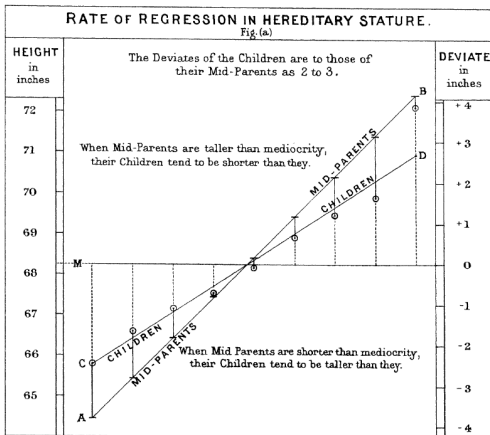
Откуда пошло название «регрессия» (Гальтон, 1886)

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.

Линейный дискриминантный анализ (Р.Фишер, 1936)

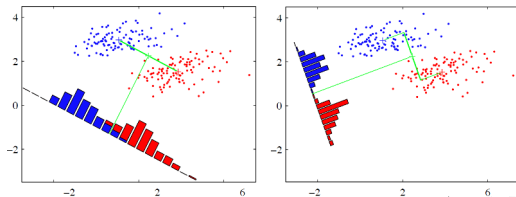
Линейная модель классификации:

$$a(x, w) = \text{sign} \left(\sum_{j=1}^n w_j f_j(x) - w_0 \right)$$

В проекции на направляющий вектор w разделяющей гиперплоскости вероятность ошибки минимальна:

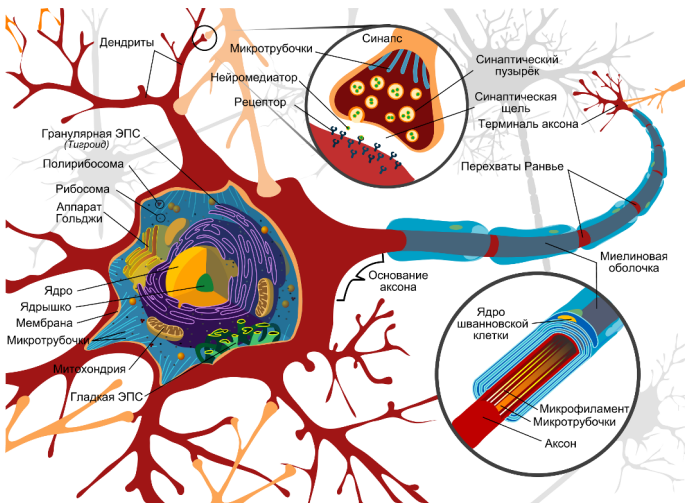


Рональд Фишер
(1890–1962)



Fisher R. A. The use of multiple measurements in taxonomic problems. 1936.

Нервная клетка — естественный нейрон

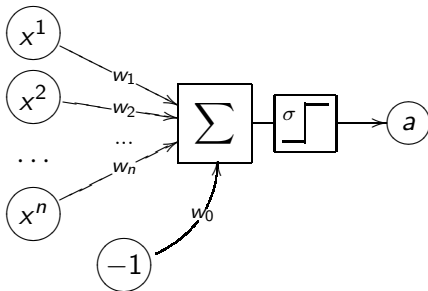


Модель МакКаллока–Питтса — искусственный нейрон

Линейная модель нейрона (1943):

$$a(x, w) = \sigma \left(\sum_{j=1}^n w_j x^j - w_0 \right),$$

где $\sigma(z)$ — функция активации,
например, $\text{sign}(z)$ или $\text{arctanh}(z)$



Уоррен МакКаллок
(1898–1969)



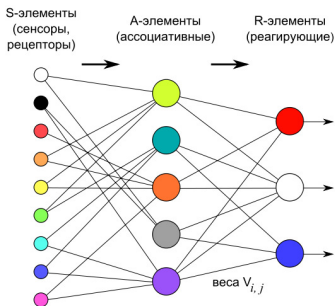
Вальтер Питтс
(1923–1969)

Перцептрон Розенблатта (1957) и теорема Новикова (1960)

Mark-1 — первый нейрокомпьютер (1960)

Обучение — метод коррекции ошибки

Архитектура — двухслойная сеть



Фрэнк Розенблатт
(1928–1971)

Розенблатт Ф. Принципы нейродинамики. Перцептроны и теория механизмов мозга. 1965 (1962).

Novikoff A. B. J. On convergence proofs on perceptrons. 1962.

Основные вехи развития теории нейронных сетей

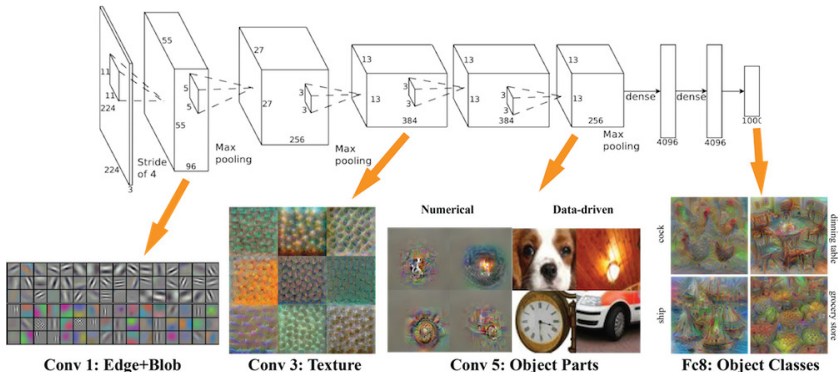
- Развитие нейронных сетей затормозилось на десятилетие
Минский М., Пайперт С. Персепоны. 1971 (1969).
- BackProp — метод обратного распространения ошибок
Галушкин А. И. Синтез многослойных систем распознавания образов. 1974.
Werbos P. J. Beyond regression: new tools for prediction and analysis in the behavioral sciences. 1974.
LeCun Y. Une procedure d'apprentissage pour reseau a seuil assymetrique. 1985.
Parker D. B. Learning-logic: Casting the cortex of the human brain in silicon. 1985.
Rummelhart D., Hinton G., Williams R. Learning internal representations by error propagation. 1986.
- Сети радиальных базисных функций
Башкиров О. А., Браверман Э. М., Мучник И. Б. Алгоритмы обучения машин распознаванию зрительных образов, основанные на использовании потенциальных функций. 1964.
Broomhead D. S., Lowe. D. Multivariable functional interpolation and adaptive networks complex systems. 1988.
- Самоорганизующиеся сети Кохонена
Kohonen T. Self-organized formation of topologically correct feature maps. 1982.
- Рекуррентные сети Хопфилда
Hopfield. Neural networks and physical systems with emergent collective computational abilities. 1982.
- Свёрточные сети
LeCun, Bottou, Bengio, Haffner. Gradient-based learning applied to document recognition. 1998.
- Глубокие сети (сети с большим числом слоёв)
Ивахненко А. Г., Лапа В. Г. Кибернетические предсказывающие устройства. 1965.
Rina Dechter. Learning while searching in constraint-satisfaction problems. 1986.
Hinton G.E. Learning multiple layers of representation. 2007.

Глубокие нейронные сети для обработки изображений

Цель — извлечение признаков из сырых данных.

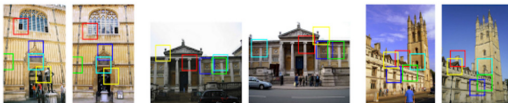
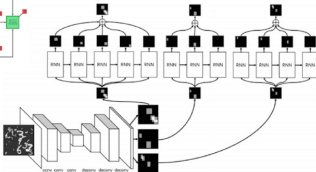
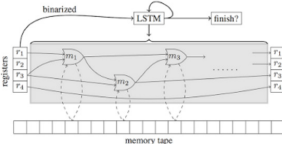
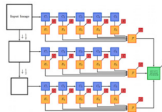
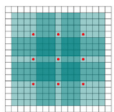
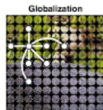
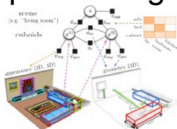
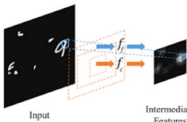
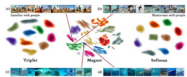
Свёрточный нейрон комбинирует признаки соседних пикселей.

Чем дальше слой, тем более крупные и сложные элементы изображений он способен распознавать.



Разнообразие приложений глубокого обучения

Deep Learning Trends @ ICLR 2016



Научная школа А. Г. Ивахненко

Метод группового учёта аргументов, МГУА (GMDH)

- 1 Самоорганизация моделей — подбор оптимальной структуры модели из огромного числа вариантов
- 2 Качество моделей оценивается в процессе перебора по совокупности разнообразных *внешних критериев*
- 3 Сотни применений, около 300 диссертаций



Алексей
Григорьевич
Ива́хненко
(1913–2007)

Ивахненко А. Г., Лапа В. Г. Кибернетические предсказывающие устройства. 1965.

Ивахненко А. Г., Зайченко Ю. П., Димитров В. Д. Принятие решений на основе самоорганизации. 1976.

Ивахненко А. Г. Индуктивный метод самоорганизации моделей сложных систем. 1982.

Композиции обучаемых моделей

- **Простое и взвешенное голосование**
Мазуров В. Д. Комитеты системы неравенств и задача распознавания. 1971.
Журавлёв Ю. И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. 1977.
Freund Y., Schapire R. E. A decision-theoretic generalization of on-line learning and an application to boosting. 1995.
Friedman G. Greedy Function Approximation: A Gradient Boosting Machine. 1999.
- **Случайный лес**
Breiman L. Random Forests. 2001.
- **Восстановление смесей распределений, EM-алгоритм**
Шлезингер М. И. О самопроизвольном различении образов. 1965.
Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM-algorithm. 1977.
- **Смеси классификаторов с областями компетентности**
Растрюгин Л. А., Эренштейн Р. Х. Коллективные правила распознавания. 1981.
Jacobs R. A., Jordan M. I., Nowlan S. J., Hinton G. E. Adaptive mixtures of local experts. 1991.

Градиентный бустинг и случайный лес — универсальные и наиболее успешные методы классификации.

Яндекс.MatrixNet — параллельная распределённая реализация Gradient Boosting над ODT (Oblivious Decision Tree).

Научная школа М. М. Бонгарда

- 1 1958: Программа «Открой закон» восстанавливала зависимость полным перебором формул
- 2 1959: Программа «Арифметика» для сокращения перебора использовала оценки информативности
- 3 1961: Программа «КоРа» перебирала информативные тройки признаков



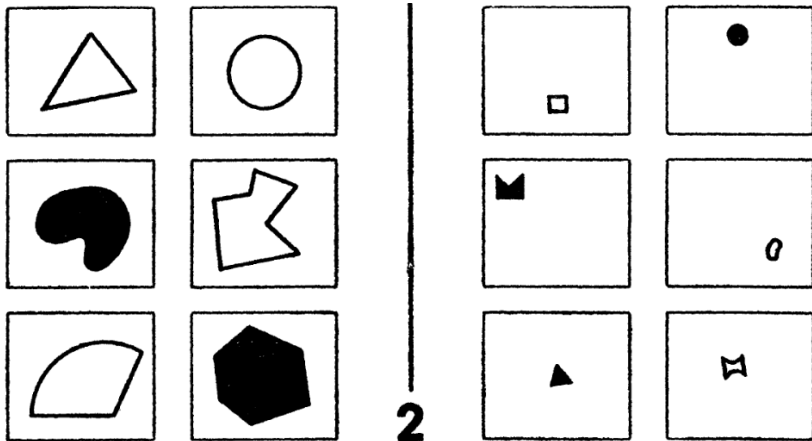
Михаил Моисеевич
Бонгард
(1924–1971)

«КоРа-3»: первое применение распознавания незрительных образов для распознавания в скважине границы нефть-вода.

Впервые применено *голосование* и *скользящий контроль*, введено понятие *предрассудка* (переобучения).

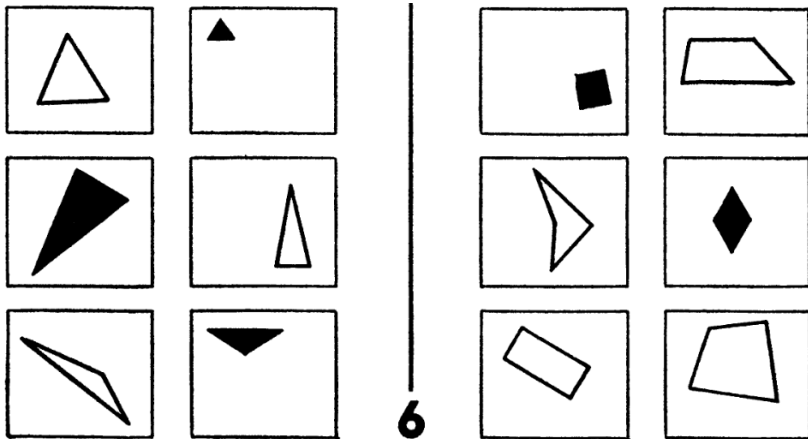
Бонгард, Вайнцвайг, Губерман, Извекова, Смирнов. Использование обучающейся программы для выявления нефтеносных пластов. 1966.

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



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

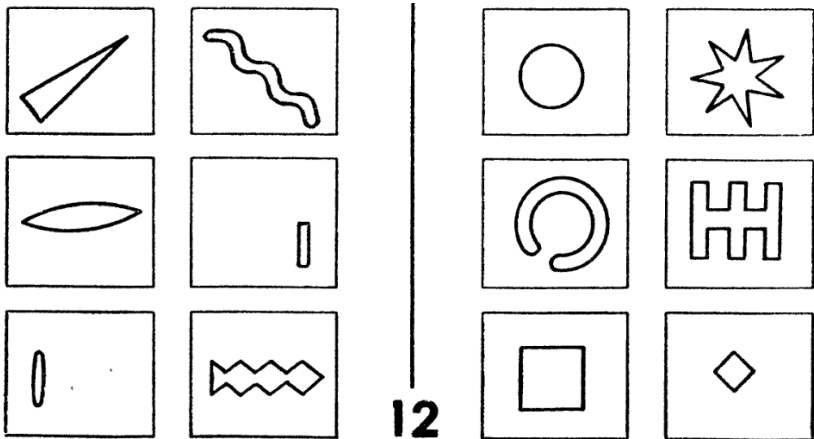
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Что даёт нам уверенность, что мы нашли верное правило?

1. Безошибочное покрытие своих и не-покрытие чужих.

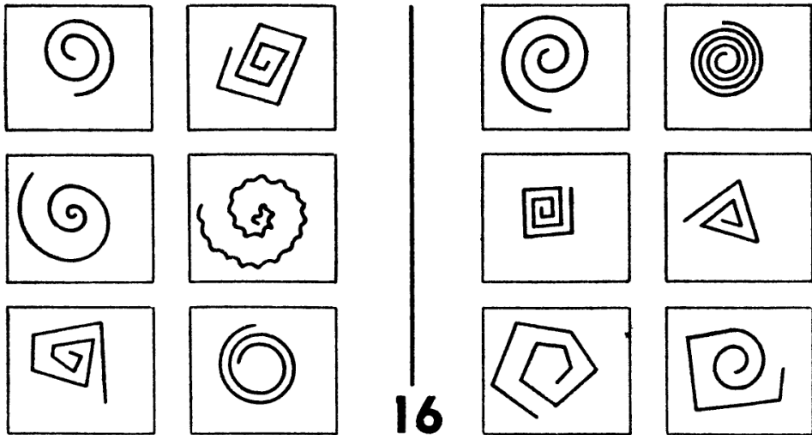
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



Что даёт нам уверенность, что мы нашли верное правило?

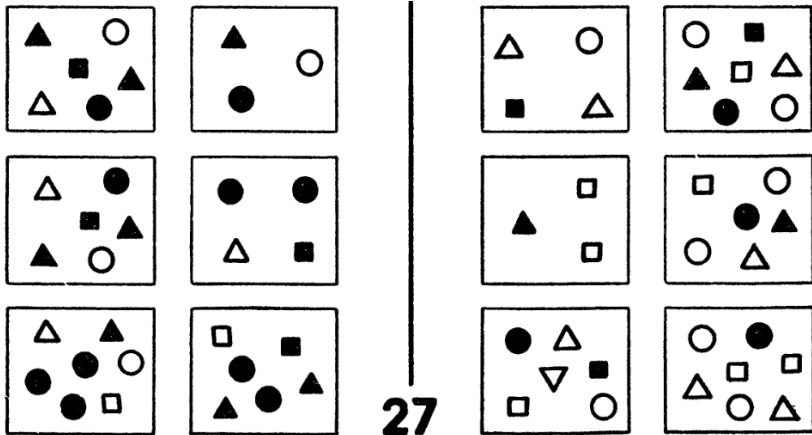
2. Простота и определённое «изящество» найденного правила.

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



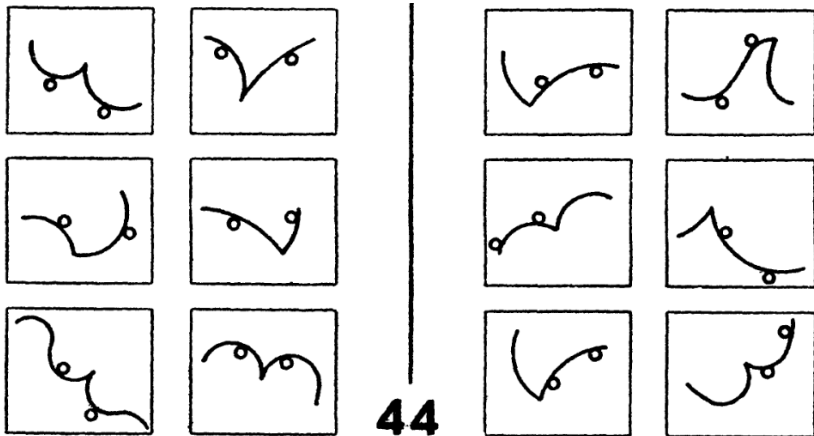
Мы решаем эти задачи почти мгновенно. Чем мы пользуемся?
Почему для компьютера они столь сложны?

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



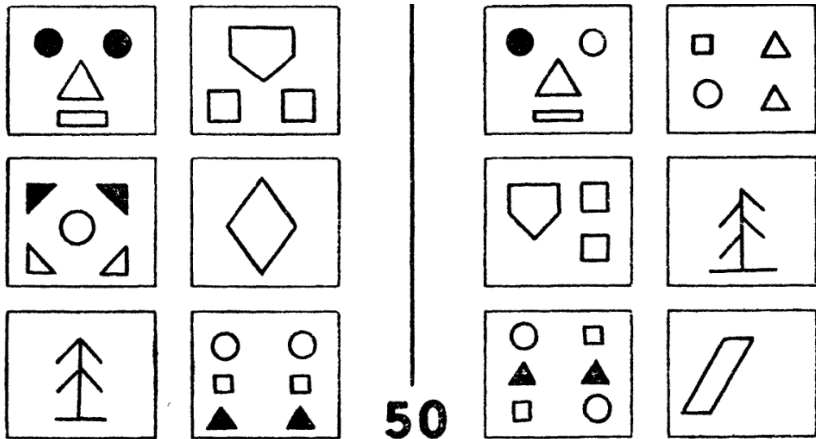
Нужно ли закладывать знания геометрии в явном виде?
Или возможно выучить геометрические понятия на примерах?

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



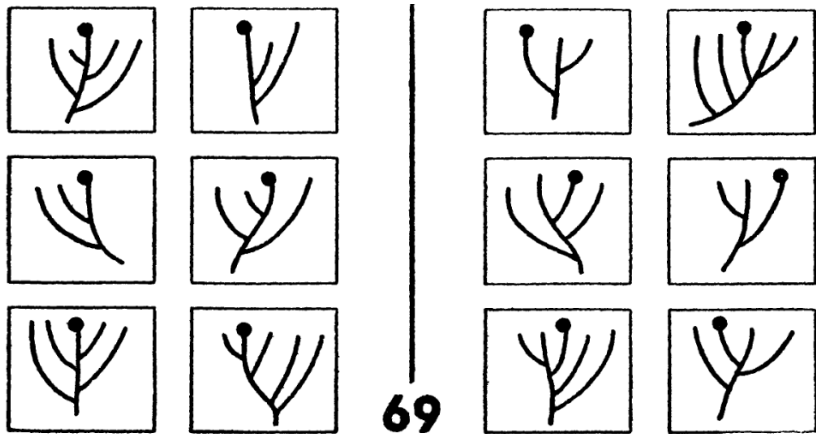
Как вычислять полезные признаки по сложным сырым данным?
Возможно ли поручить перебор признаков и моделей машине?

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!

Поиск ассоциативных правил (Association Rule Induction)

Data Mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных, доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности (Григорий Пятецкий-Шапиро, 1992)



Научная школа М. А. Айзермана

- *Гипотеза компактности*: близкие объекты, как правило, находятся в одном классе
- *Идея потенциальных функций* заимствована из физики
- *Линейные модели в пространстве близостей* $f_i(x) = K(x, x_i)$ объекта x до обучающих объектов x_i



Марк Аронович
Айзерман
(1913–1992)

М. А. Айзерман, Э. М. Браверман, Л. И. Розоноэр. Теоретические основы метода потенциальных функций в задаче об обучении автоматов разделению входных ситуаций на классы. 1964.

М. А. Айзерман, Э. М. Браверман, Л. И. Розоноэр. Метод потенциальных функций в теории обучения машин. 1970.

А. Г. Аркадьев, Э. М. Браверман. Обучение машин распознаванию образов. 1964.

Понятие обучаемости в SLT, Statistical Learning Theory

Семейство классификаторов A обучаемо:

$$P\left\{\sup_{a \in A} |P(a) - \nu(a, X^\ell)| > \varepsilon\right\} \leq \eta,$$

$P(a)$ — вероятность ошибки классификатора,
 $\nu(a, X^\ell)$ — эмпирический риск (частота
ошибки классификатора a на выборке).

Основные результаты VC-теории:

- Обосновано ограничение сложности A
- Понятие ёмкости семейства, $VCdim$
- Метод структурной минимизации риска

Вапник В. Н., Червоненкис А. Я.
Теория распознавания образов. М.: Наука, 1974.



Владимир
Наумович Вапник



Алексей Яковлевич
Червоненкис
(1938–2014)

Метод опорных векторов SVM, Support Vector Machine

Метод обобщённого портрета (1963) → SVM (1992)

- Линейный классификатор с зазором максимальной ширины
- Аппроксимация и регуляризация эмпирического риска

$$\sum_{i=1}^{\ell} (1 - \langle x_i, w \rangle y_i)_+ + \frac{1}{2C} \|w\|^2 \rightarrow \min_w$$

- Изящный переход в спрямляющее пространство

$$\langle x, x' \rangle \rightarrow K(x, x') = \langle \psi(x), \psi(x') \rangle$$

- В результате — двухслойная нейронная сеть с высокой обобщающей способностью и автоматическим выбором числа нейронов скрытого слоя

В. Н. Вапник, А. Я. Лернер. Узнавание образов при помощи обобщенных портретов. 1963.

Задачи, некорректно поставленные по Адамару

Корректно поставленная задача:

- решение существует,
- решение единственно,
- решение устойчиво
(непрерывно зависит от данных
в некоторой разумной топологии).

Задачи восстановления зависимостей
по эмпирическим данным
— всегда некорректно поставленные.

Регуляризация — это введение ограничений на модель.



Жак Саломон
Адамар
(1865–1963)

Hadamard J. Sur les problèmes aux dérivées partielles et leur signification physique. 1902.

Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. 1974.

Регуляризация линейных моделей

Регуляризатор — добавка к внутреннему критерию $Q(a, X^\ell)$, штраф за сложность (complexity penalty) модели $a \in A$:

$$Q_{\text{рег}}(a, X^\ell) = Q(a, X^\ell) + \text{штраф}(A) \rightarrow \min_{a \in A}$$

Линейные модели: $A = \{a(x) = \text{sign}\langle w, x \rangle\}$ — классификация,
 $A = \{a(x) = \langle w, x \rangle\}$ — регрессия.

L_2 -регуляризация (ридж-регрессия): штраф(w) = $\tau \sum_{j=1}^n w_j^2$

L_1 -регуляризация (LASSO): штраф(w) = $\tau \sum_{j=1}^n |w_j|$

L_0 -регуляризация (AIC, BIC): штраф(w) = $\tau \sum_{j=1}^n [w_j \neq 0]$

Регуляризаторы для упрощения модели (отбора признаков)

- Структурная минимизация риска

Вапник В. Н., Червоненкис А. Я. О равномерной сходимости частот появления событий к их вероятностям. 1971.

- Критерий Акаике

Akaike H. Information theory and an extension of the maximum likelihood principle. 1973.

- Байесовский информационный критерий

Schwarz G. E. Estimating the dimension of a model. 1978

- LASSO (least absolute shrinkage and selection operator)

Tibshirani R. Regression Shrinkage and Selection via the lasso. 1996

- LARS (least angle regression)

Efron B., Hastie T., Johnstone I., Tibshirani R. Least Angle Regression. 2004

- ElasticNet (сумма L_0 и L_1 регуляризаторов)

Hui Zou, Hastie T. Regularization and Variable Selection via the Elastic Net. 2005

- Негладкие регуляризаторы для отбора признаков

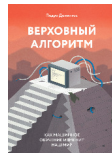
Tatarchuk A., Mottl V., Eliseyev A., Windridge D. Selectivity supervision in combining pattern recognition modalities by feature- and kernel-selective Support Vector Machines. 2008.

Tatarchuk A., Urlov E., Mottl V., Windridge D. A support kernel machine for supervised selective combining of diverse pattern-recognition modalities. 2010.

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.



Основные типы задач машинного обучения

- 1 Предварительная обработка (data preparation)
 - извлечение признаков (feature extraction)
 - отбор признаков (feature selection)
 - восстановление пропусков (missing values)
- 2 Обучение с учителем (supervised learning)
 - классификация (classification)
 - регрессия (regression)
 - ранжирование (learning to rank)
 - прогнозирование (forecasting)
- 3 Обучение без учителя (unsupervised learning)
 - кластеризация (clustering)
 - восстановление плотности (density estimation)
 - поиск ассоциативных правил (association rule learning)
- 4 Частичное обучение (semi-supervised learning)
 - трансдуктивное обучение (transductive learning)
 - одноклассовая классификация (one-class classification)
 - обучение с положительными примерами (PU-learning)

Основные типы задач машинного обучения

- 5 Обучение представлений (representation learning)
 - обучение признаков (feature learning)
 - обучение многообразий (manifold learning)
 - анализ главных компонент (principal component analysis)
 - матричные разложения (matrix factorization)
 - тематическое моделирование (topic modeling)
- 6 Глубокое обучение (deep learning)
- 7 Обучение структуры модели (structure learning)
- 8 Обучение композиций (ensemble learning)
- 9 Привилегированное обучение (learning with privileged inform.)
- 10 Динамическое обучение (online/incremental learning)
- 11 Активное обучение (active learning)
- 12 Обучение с подкреплением (reinforcement learning)
- 13 Обучение с переносом опыта (transfer learning)
- 14 Мета-обучение (meta-learning)

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 видеокладов 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

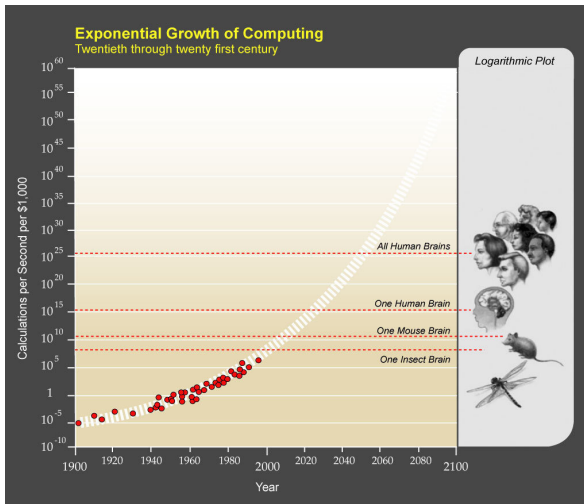
<http://abv24.com/istoriya-mashinnogo-obucheniya>

Три предпосылки этого бума

Три перехода количества в качество в нейронных сетях:

- 1 Достижения микроэлектроники**
 - процессоры, память, графические карты
 - рост вычислительных мощностей по закону Мура
 - экстраполяция: $80 \cdot 10^9$ нейронов в 2035–2050 гг.
- 2 Повсеместность и доступность IT-технологий**
 - накопление больших выборок данных
 - краудсорсинг
- 3 Развитие методов машинного обучения**
 - rectified linear unit, ReLU (V.Nair & G.Hinton, 2010)
 - быстрые SGD алгоритмы: AdaDelta (Kingma & Ba 2014)
 - dropout (G.Hinton, 2012)
 - регуляризации

Закон Мура

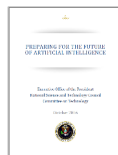


Ray Kurzweil. The singularity is near: When humans transcend biology. 2006.

Отчёты Белого Дома США, октябрь 2016

«Nations with the strongest presence in AI R&D will establish leading positions in the automation of the future»

- Цифровая и распределённая экономика
- Автоматизация и сокращение издержек
- Автономный транспорт и роботизация
- Оптимизация логистики и цепей поставок
- Оптимизация энергетических сетей
- Мониторинг сельского хозяйства
- Персональная медицина
- Персональные образовательные траектории
- Автономные системы вооружений



Preparing for the Future of Artificial Intelligence. NSTC. 2016.

Отчёты Белого Дома США. Некоторые из 23 рекомендаций

- 1 Государственным и коммерческим организациям: активнее развивать партнёрство с научными коллективами для эффективного использования данных
- 2 Развивать стандарты открытых данных для привлечения научного сообщества к решению задач
- 8 Развивать системы управления беспилотным транспортом
- 11 Вести постоянный мониторинг исследований ИИ в мире
- 13 Поддерживать фундаментальные исследования по ИИ
- 14 Развивать образовательные программы по ИИ и курсы повышения квалификации для прикладников
- 20 Развивать международную кооперацию по ИИ
- 22 Учитывать взаимовлияние ИИ и кибербезопасности

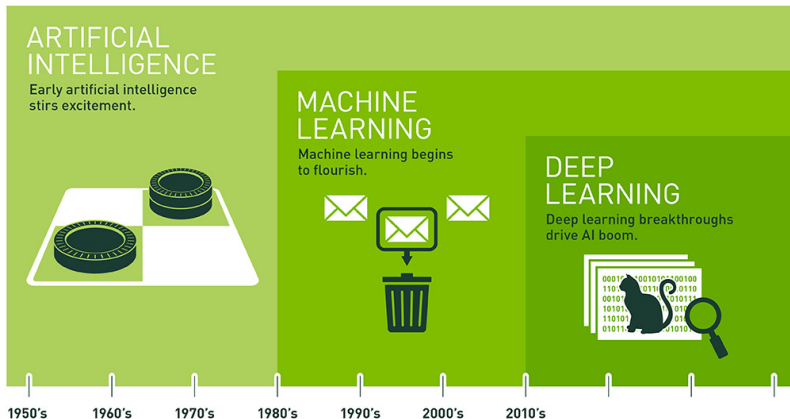
Preparing for the Future of Artificial Intelligence. NSTC. 2016.

Открытые данные. Открытый код. Открытая наука

- **Выгоды открытых данных**
 - *для компаний*: подбор технологий и исполнителей
 - *для индустрии*: бенчмаркинг задач, реклама отрасли
 - *для исследователей*: проверка теорий и технологий
 - *для студентов*: получение опыта, наполнение портфолио
- **Выгоды открытого кода**
 - снижение издержек, ускорение разработки и внедрения
 - координация усилий исследователей и разработчиков
 - снижение технологических барьеров для выхода на рынок
- **Конкурсы анализа данных**
 - www.NetflixPrize.com – первый крупный конкурс, \$1 млн. (2006-2009)
 - www.kaggle.com – самая известная платформа
 - DataRing.ru – отечественная конкурсная платформа

Будущее машинного обучения

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



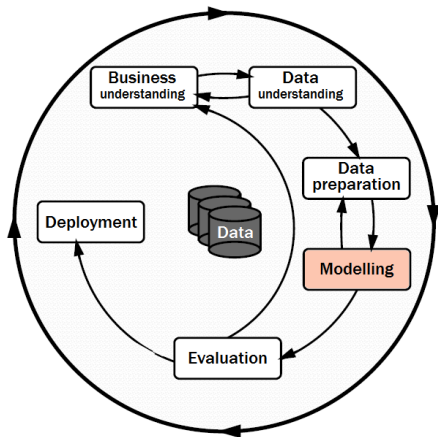
Особенности данных и постановок прикладных задач

- разнородные (признаки измерены в разных шкалах)
- неполные (измерены не все, имеются пропуски)
- неточные (измерены с погрешностями)
- противоречивые (объекты одинаковые, ответы разные)
- избыточные (сверхбольшие, не помещаются в память)
- недостаточные (объектов меньше, чем признаков)
- неструктурированные (нет признаковых описаний)

- заказчик не знает точно, чего хочет
- критерии качества нетривиальны или неясны
- заказчик не заботится о качестве своих данных

Межотраслевой стандарт интеллектуального анализа данных

CRISP-DM: Cross Industry Standard
Process for Data Mining (1999)



Компании-инициаторы:

- SPSS
- Teradata
- Daimler AG
- NCR Corp.
- OHRA

Шаги процесса:

- понимание бизнеса
- понимание данных
- предобработка данных
- инженерия признаков
- разработка моделей
- настройка параметров
- оценивание качества
- внедрение

Полезные ссылки

- www.MachineLearning.ru — русскоязычная вики
- www.kdnuggets.com — главный сайт датамайнеров
- www.datasciencecentral.com — 72 000 датамайнеров
- www.kaggle.com — конкурсы анализа данных
- DataRing.ru — отечественная конкурсная платформа
- archive.ics.uci.edu/ml — UCI ML Repository (349 datasets)
- ru.coursera.org/learn/machine-learning — курс Эндрю Блэна
- 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

Литература

- *Домингос П.* Верховный алгоритм. 2016. 336 с.
- *Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning. Springer, 2014. 739 p.
- *Bishop C. M.* Pattern recognition and machine learning. Springer, 2006. 738 p.
- *Мерков А. Б.* Распознавание образов. Введение в методы статистического обучения. 2011. 256 с.
- *Мерков А. Б.* Распознавание образов. Построение и обучение вероятностных моделей. 2014. 238 с.
- *Коэльо Л. П., Ричарт В.* Построение систем машинного обучения на языке Python. 2016. 302 с.
- Машинное обучение (курс лекций, К. В. Воронцов). www.MachineLearning.ru. 2004–2017.