Курс «Введение в машинное обучение»

Научный метод и основные понятия машинного обучения

Воронцов Константин Вячеславович
k.v.vorontsov@phystech.edu
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Введение в машинное обучение (курс лекций, К.В.Воронцов)»

МФТИ • 13 февраля 2025

Эпиграф №1

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

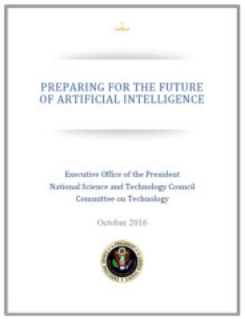
Клаус Мартин Шваб,
президент Всемирного экономического форума

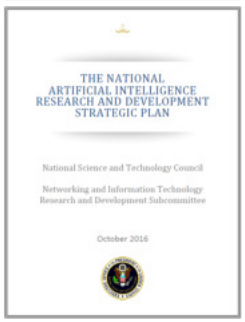
Мир наконец поверил в искусственный интеллект
Машинное обучение — новый двигатель прогресса
Машинное обучение — технология, которая меняет мир

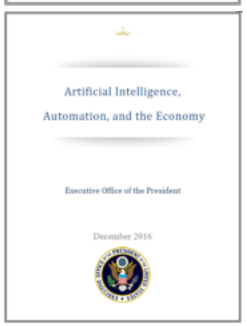
Эпиграф №2 (отчёты Белого Дома США, октябрь 2016)

«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 (Energy Tech)
- Автоматизация банковских услуг (Fin Tech)
- Автоматизация юридических услуг (Legal Tech)
- Автоматизация образовательных услуг (Ed Tech)
- Автоматизация работы с кадрами (HR Tech)
- Персональная медицина (Med Tech)
- Автоматизация в сельском хозяйстве (Agro Tech)
- Автономные системы вооружений (Mil Tech)







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

Содержание

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)

Таблица открытия: множество объектов $\{x_i : i = 1, \ldots, \ell\}$

- $f_j(x)$ — измеряемые *признаки* объектов, $j = 1, \ldots, n$
- $y_i \in \mathbb{R}$ — измеряемое значение *целевого свойства* x_i, либо $y_i \in \{0, 1\}$ — отсутствие или наличие *целевого свойства*

Фрэнсис Бэкон. Новый органон. 1620.

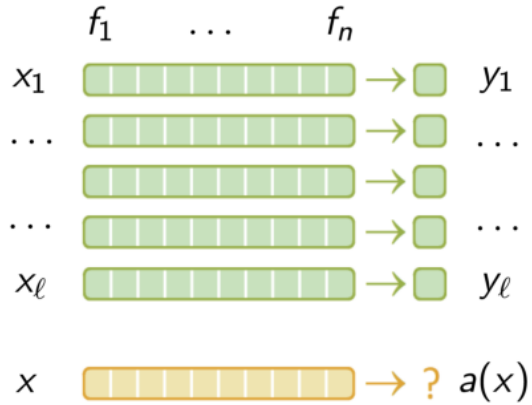
Научный метод: основные шаги и принципы

Наблюдения (эмпирический опыт, измерения, эксперименты)
Гипотеза (модель, теория) объясняет и обобщает наблюдения

- *Принцип верифицируемости* (Фрэнсиса Бэкона): гипотеза подтверждается измеримыми наблюдениями
- *Принцип фальсифицируемости* (Карла Поппера): должны существовать способы опровергнуть гипотезу
- *Принцип соответствия:* новая гипотеза или теория должна включать прежнюю как частный случай
- *Принцип минимальной достаточности* (бритва Оккама): среди всех объяснений следует выбирать самое простое
- *Принцип воспроизводимости*: сообществу должно быть предоставлено всё необходимое для повторения результата
- *Принцип честности* (Ричарда Фейнмана): гипотеза должна сопровождаться указанием её «слабых мест», возможных ошибок, противоречий, границ применимости

Восстановление зависимостей по эмпирическим данным

Дано: обучающая выборка
объектов $x_i = \big(f_1(x_i), \ldots, f_n(x_i)\big) \in X$
с *ответами* $y_i = y(x_i) \in Y$, $i = 1, \ldots, \ell$



Найти: параметры w *модели* $a(x, w)$,
приближающей зависимость $y : X \to Y$

Критерий: *минимум эмпирического риска*

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

$\mathscr{L}(a, y)$ — *функция потерь* (ответ модели a при правильном y)

Основные типы задач машинного обучения с учителем:

- $\mathscr{L}(a, y) = (a - y)^2$ в задачах регрессии, $y_i \in \mathbb{R}$
- $\mathscr{L}(a, y) = [a \neq y]$ в задачах классификации, $y_i \in \{0, 1\}$

Метод наименьших квадратов (Гаусс, 1795)

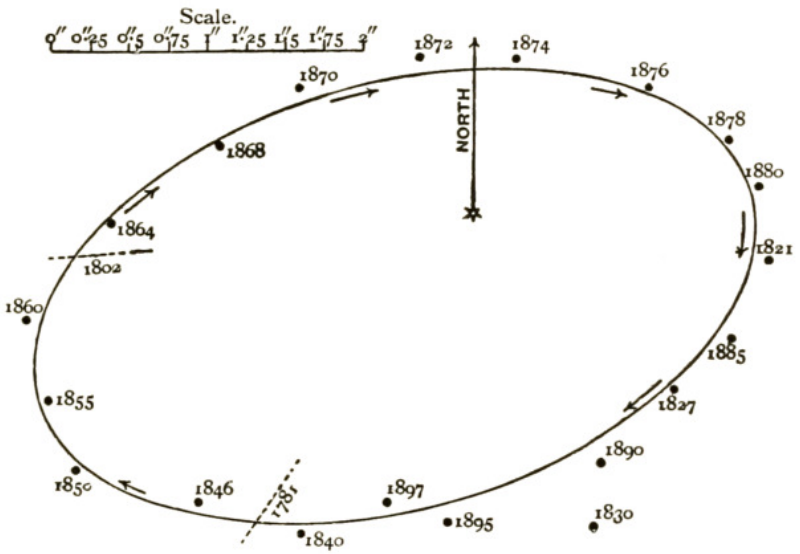
Линейная модель регрессии:

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

Метод наименьших квадратов:

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

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.

Восстановление уравнения эллипса по точкам

Дано: $(x_i, y_i)_{i=1}^{\ell}$ — точки эллипса, в декартовых координатах, измеренные с погрешностями

Найти: параметры w_{ij} $(i + j \leqslant 2)$ уравнения эллипса

$$w_{20}x^2 + w_{11}xy + w_{02}y^2 + w_{10}x + w_{01}y + w_{00} = 0$$

Критерий: метод наименьших квадратов (least squares)

$$\sum_{i=1}^{\ell}\left(w_{20}x_i^2 + w_{11}x_iy_i + w_{02}y_i^2 + w_{10}x_i + w_{01}y_i + w_{00}\right)^2 \to \min_w$$

Это модель, линейная по параметрам $w_{20}, w_{11}, w_{02}, w_{10}, w_{01}, w_{00}$

Вопрос 1: почему такая нумерация параметров?
Вопрос 2: где здесь признаки? где целевое свойство?

Восстановление уравнения эллипса по точкам: 2-й способ

Дано: $(t_i, x_i, y_i)_{i=1}^{\ell}$ — точки эллипса, добавлены измерения моментов времени t_i, тоже с погрешностями

Найти: уравнение эллипса в параметрической форме

$$\begin{cases} x(t; R, a, b, x_0) = R\cos(at+b) + x_0 \\ y(t; r, a, b, y_0) = r\sin(at+b) + y_0 \end{cases}$$

Критерий: метод наименьших квадратов (least squares)

$$\sum_{i=1}^{\ell} \big(x(t_i; R, a, b, x_0) - x_i\big)^2 + \big(y(t_i; r, a, b, y_0) - y_i\big)^2 \to \min$$

Модель линейна по параметрам R, r, x_0, y_0, не линейна по a, b

Вопросы: где здесь признаки? модель? целевое свойство?

Восстановление уравнения эллипса по точкам: 3-й способ

Дано: $(\varphi_i, \rho_i)_{i=1}^{\ell}$ — точки эллипса в полярных координатах

Найти: уравнение эллипса в полярных координатах

$$\rho(\varphi) = \frac{R}{1 - \varepsilon \cos \varphi}$$

Критерий: метод наименьших квадратов, два варианта:

$$\sum_{i=1}^{\ell} \left(\frac{R}{1 - \varepsilon \cos \varphi_i} - \rho_i \right)^2 \to \min_{R,\varepsilon}; \quad \sum_{i=1}^{\ell} \left(\varepsilon \rho_i \cos \varphi_i + R - \rho_i \right)^2 \to \min_{R,\varepsilon}$$

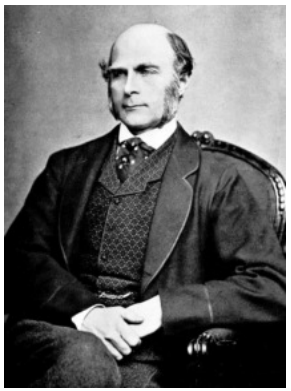
Вопрос 1: модель линейная или не линейная по параметрам?
Вопрос 2: где здесь признаки? где целевое свойство?
Вопрос 3: почему в модели два параметра, а не шесть?
Вопрос 4: какие данные на самом деле были у Гаусса?
Вопрос 5: где физически находится фокус эллипса?

История термина «регрессия» (Гальтон, 1886)

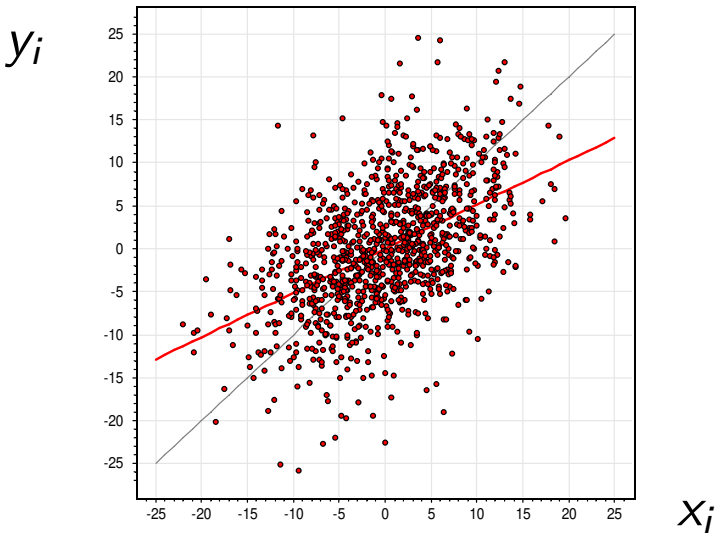
Дано: $(x_i, y_i)_{i=1}^{\ell}$ — отклонение роста отца (x_i) и взрослого сына (y_i) от среднего в популяции

Найти: модель наследственности роста $y(x) = kx$

Критерий: метод наименьших квадратов



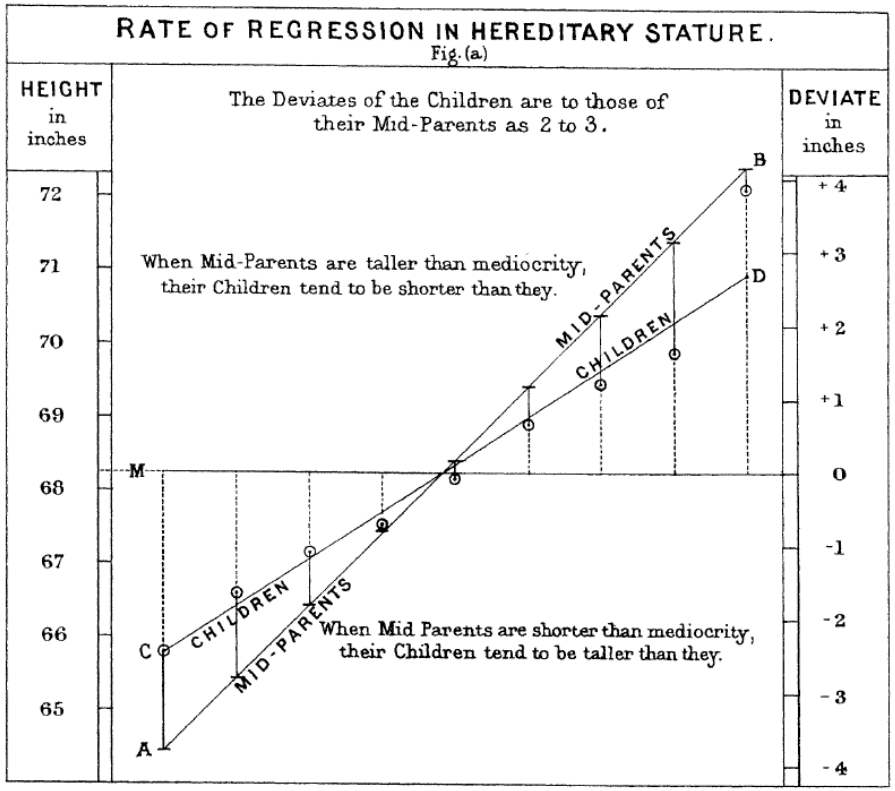
Фрэнсис Гальтон (1822–1911)

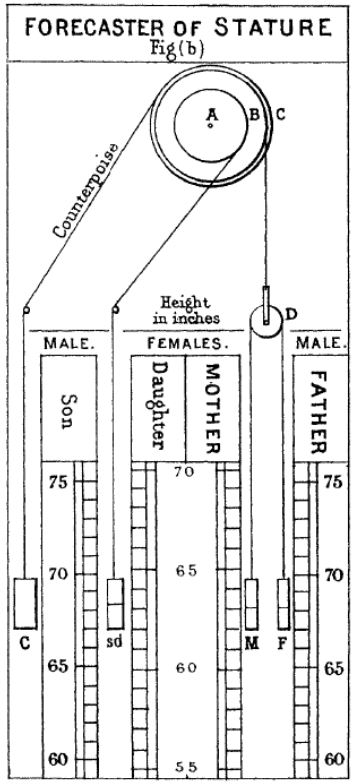


Вопрос: что было бы при $k > 1$?

История термина «регрессия» (Гальтон, 1886)

«Регрессия к посредственности» — угол наклона меньше 1
Скрытый смысл: обратный ход исследования от данных к модели

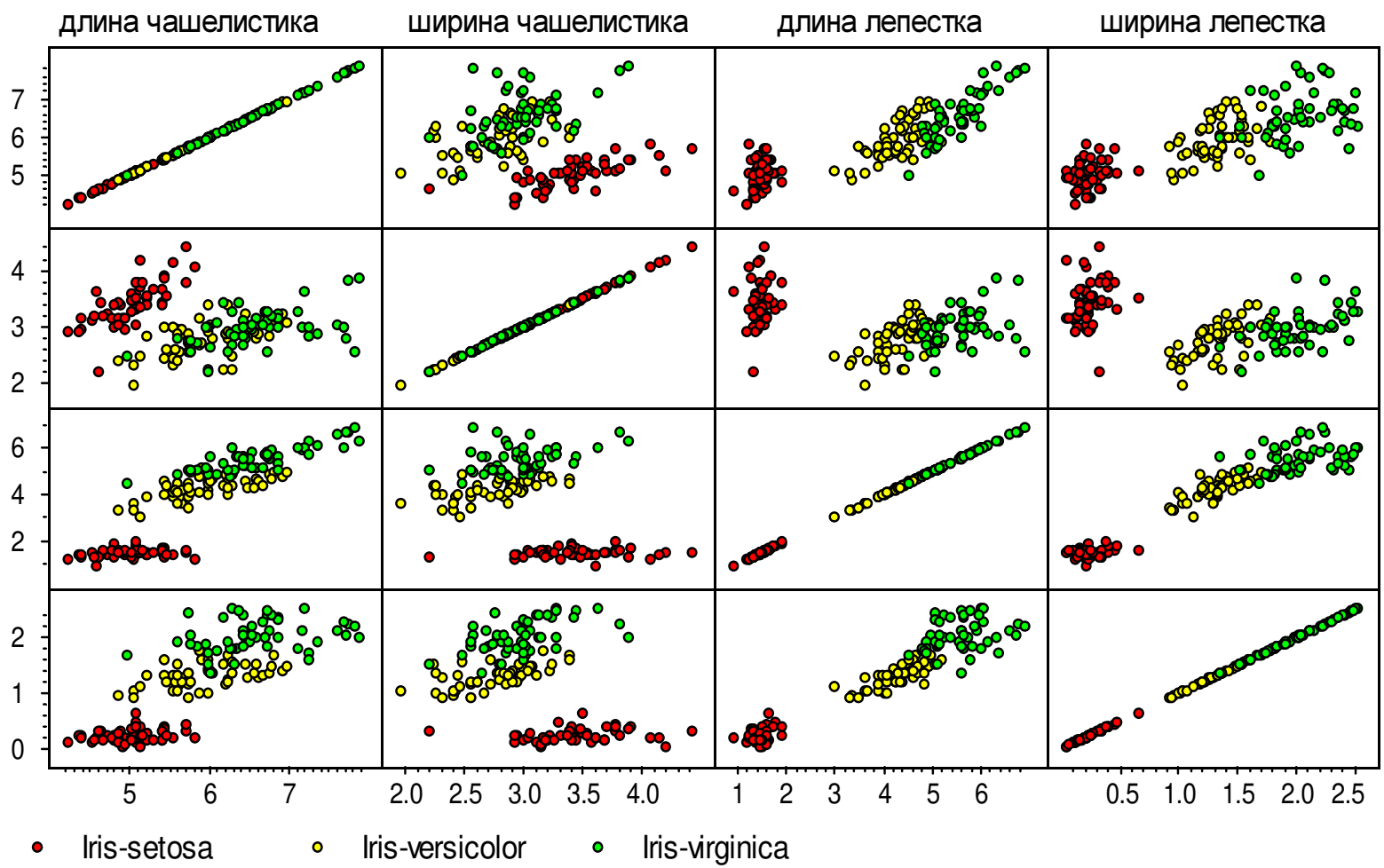




Galton F. Regression towards mediocrity in hereditary stature. 1886.

Пример: задача классификации цветков ириса [Фишер, 1936]

Дано: $n = 4$ признака, $|Y| = 3$ класса, наблюдений $\ell = 150$



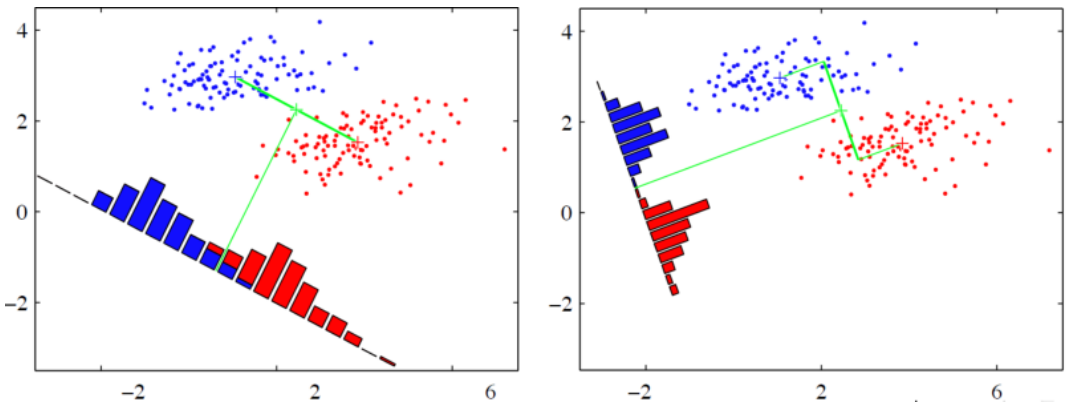
Линейный дискриминантный анализ (Р.Фишер, 1936)

Найти линейную модель классификации:

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

Критерий: 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.

Формализация постановки задачи в машинном обучении

Дано: X — пространство *объектов*
$X^\ell = \{x_1, \ldots, x_\ell\} \subset X$ — *обучающая выборка* (training sample)
$a(x, w)$, $a\colon X \times W \to Y$ — параметрическая модель, гипотеза

Найти $w \in W$ — вектор параметров модели $a(x, w)$

Критерий минимизации эмпирического риска
(empirical risk minimization, ERM):

$$\sum_{i=1}^{\ell} \mathscr{L}(w, x_i) + \mathscr{R}(w) \to \min_w$$

$\mathscr{L}(w, x)$ — *функция потерь* (loss function),
тем больше, чем хуже модель $a(x, w)$ обработала объект x

$\mathscr{R}(x)$ — *регуляризатор* для формализации дополнительных (как правило, не прецедентных) требований к модели

Как задаются объекты. Векторное признаковое описание

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

Скалярные (одномерные) типы признаков:

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

Вектор $\big(f_1(x), \ldots, f_n(x)\big)$ — *признаковое описание* объекта x

Матрица «*объекты–признаки*» (feature data)

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

Как задаются объекты. Сложно структурированные данные

Сложные типы признаков:

- текст, символьная дискретнозначная последовательность
- сигнал, непрерывнозначная последовательность
- чёрно-белое, серое изображение — 2D-матрица
- цветное, многозональное изображение — 3D-матрица
- видео, последовательность изображений
- транзакции, взаимодействия с другими объектами
- всё это вместе — мультимодальные данные

Выделение признаков (feature extraction)

- вычисление признаков по формулам (feature engineering)
- генерация векторов признаков (feature generation):
  $f(x, w')$, $f : X \times W' \to \mathbb{R}^n$ — модель векторизации объекта
  $\sum_{i=1}^{\ell} \mathscr{L}\big(w, f(x_i, w')\big) \to \min_{w,w'}$ — обучение векторизатора

Как задаются ответы. Типы задач

Задачи обучения с учителем (supervised learning):
на объектах $x_i \in X^\ell$ заданы правильные ответы $y_i = y(x_i)$

задачи классификации (classification, Y — class labels):

- $Y = \{-1, +1\}$ — на 2 класса (binary classification)
- $Y = \{1, \ldots, M\}$ — на много классов (multiclass c.)
- $Y = \{0, 1\}^M$ — на пересекающиеся классы (multilabel c.)

задачи ~~восстановления~~ регрессии (regression):

- $Y = \mathbb{R}$ или $Y = \mathbb{R}^m$

задачи ранжирования (ranking, learning to rank):

- Y — конечное упорядоченное множество

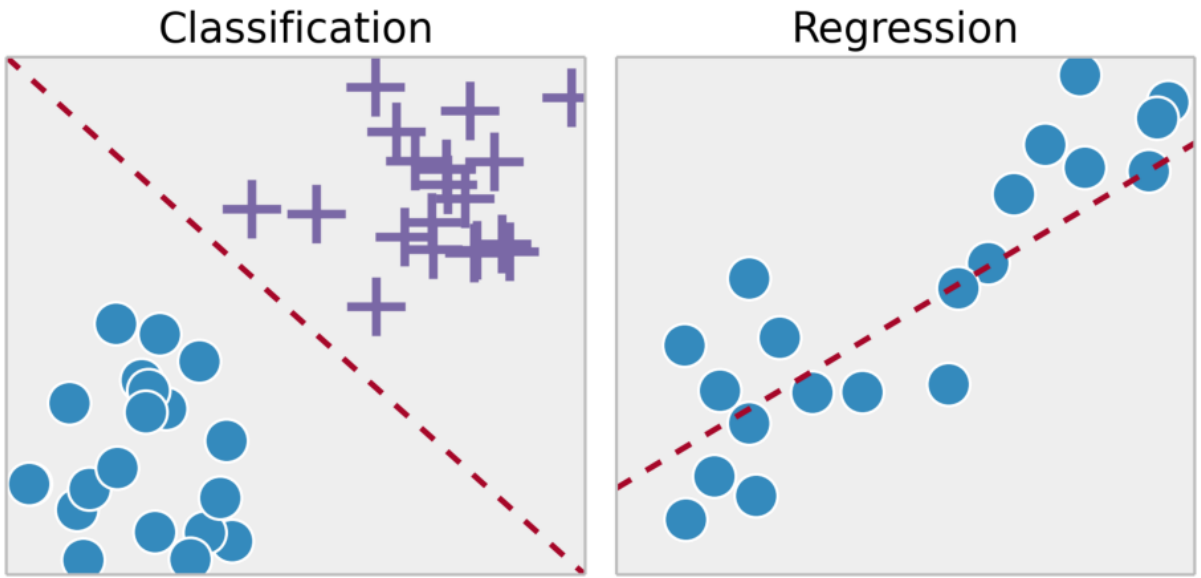
Задачи обучения без учителя (unsupervised learning):

- ответов нет, требуется что-то делать с самими объектами

Статистическое (машинное) обучение с учителем

= обучение по прецедентам
= восстановление зависимостей по эмпирическим данным
= предсказательное моделирование
= аппроксимация функций по заданным точкам

Два основных типа задач — *классификация* и *регрессия*



Как задаются предсказательные модели

Модель (predictive model) — параметрическое семейство функций

$$A = \{a(x, w) \mid w \in W\},$$

где $a\colon X \times W \to Y$ — фиксированная функция,
W — множество допустимых значений параметра w

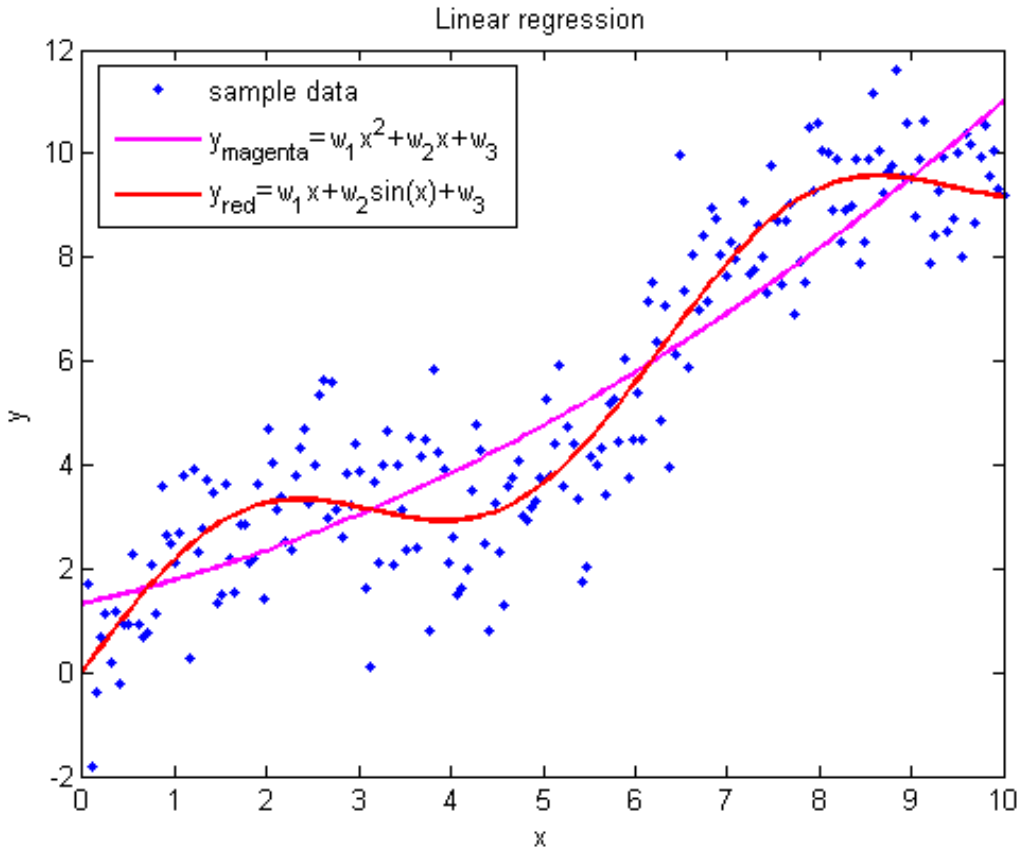
Пример.
Линейная модель с вектором параметров $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$:

$$a(x, w) = \sum_{j=1}^{n} w_j f_j(x) \quad \text{— для регрессии и ранжирования, } Y = \mathbb{R}$$

$$a(x, w) = \operatorname{sign} \sum_{j=1}^{n} w_j f_j(x) \quad \text{— для классификации, } Y = \{-1, +1\}$$

Пример: задача регрессии, синтетические данные

$X = Y = \mathbb{R}$, $\ell = 200$, $n = 3$ признака: $\{x, x^2, 1\}$ или $\{x, \sin x, 1\}$



- вычисление новых признаков может обогатить модель
- на практике очень важно «правильно угадать модель»

Алгоритм обучения, этапы обучения и применения

Этап *обучения* (train):

алгоритм обучения (learning algorithm) $\mu\colon (X \times Y)^\ell \to W$
по выборке $X^\ell = (x_i, y_i)_{i=1}^\ell$ строит функцию $a(x, w)$,
оценивая (оптимизируя) параметры модели $w \in W$:

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix} \xrightarrow{y} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix} \xrightarrow{\mu} w$$

Этап *применения* (test):

функция $a(x, w)$ для новых объектов x'_i выдаёт ответы $a(x'_i, w)$:

$$\begin{pmatrix} f_1(x'_1) & \dots & f_n(x'_1) \\ \dots & \dots & \dots \\ f_1(x'_k) & \dots & f_n(x'_k) \end{pmatrix} \xrightarrow{a} \begin{pmatrix} a(x'_1, w) \\ \dots \\ a(x'_k, w) \end{pmatrix}$$

Как задаются функции потерь

Функции потерь для задач классификации:

- $\mathscr{L}(w,x) = \big[a(x,w) \neq y(x)\big]$ — индикатор ошибки
- для модели $a(x,w) = \operatorname{sign} b(x,w)$, $Y = \{-1,+1\}$:
  $\mathscr{L}(w,x) = L\big(b(x,w)y(x)\big)$ — margin-based функция потерь,
  $L(M)$ — непрерывная невозрастающая функция от
  $M(x,w) = b(x,w)y(x)$ — *отступ (margin)* объекта x

Функции потерь для задач регрессии:

- $\mathscr{L}(w,x) = \big|a(x,w) - y(x)\big|$ — абсолютное значение ошибки
- $\mathscr{L}(w,x) = \big(a(x,w) - y(x)\big)^2$ — квадратичная ошибка

Метод наименьших квадратов — частный случай ERM:

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

Пример Рунге. Аппроксимация функции полиномом

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

Признаковое описание объекта $x \mapsto (1, x^1, x^2, \ldots, x^n)$

Модель полиномиальной регрессии

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

Обучение методом наименьших квадратов:

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

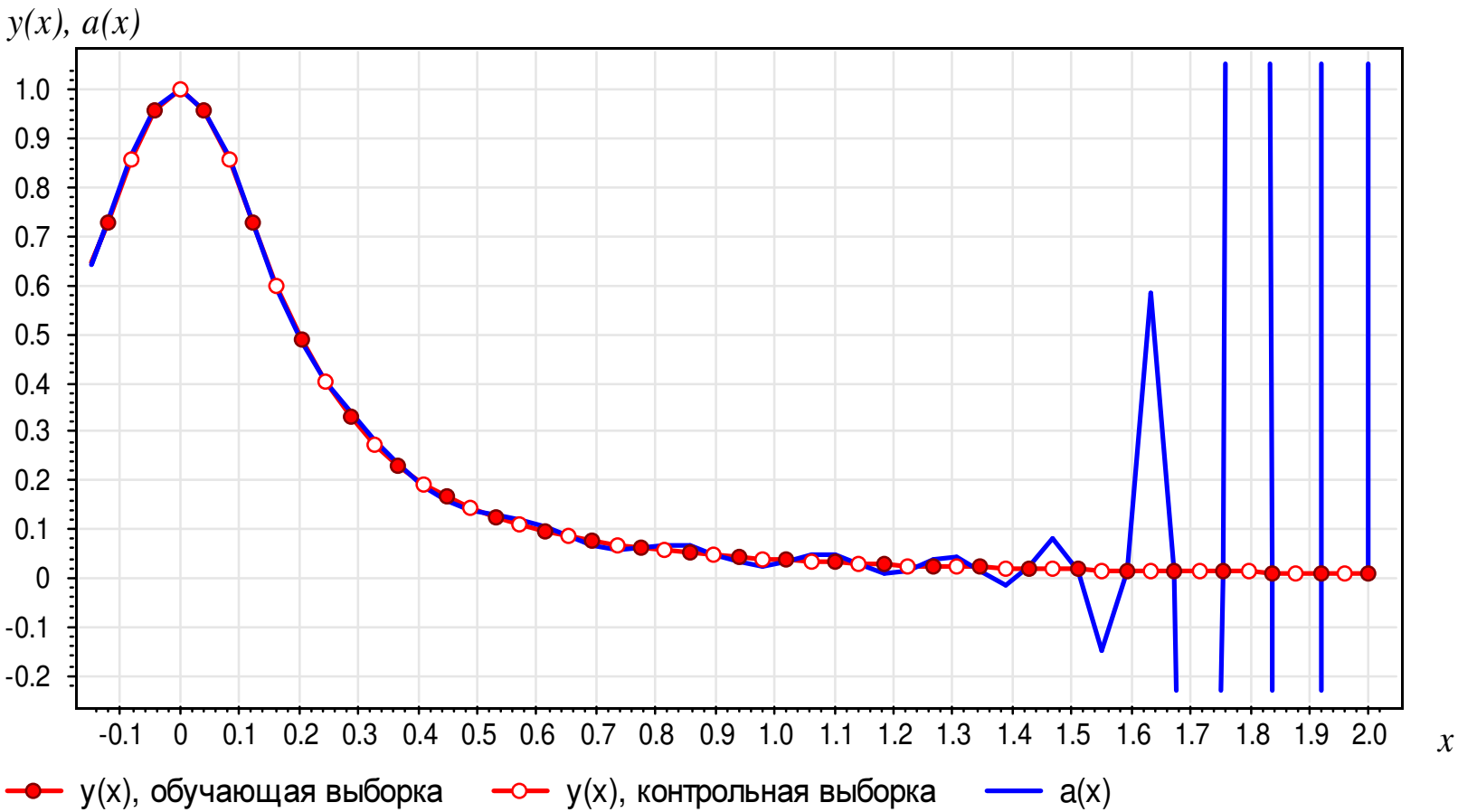
Обучающая выборка: $X^\ell = \left\{ x_i = 4\frac{i-1}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell \right\}$

Контрольная выборка: $X^k = \left\{ x_i = 4\frac{i-0.5}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell - 1 \right\}$

Что происходит с $Q\left(w, X^\ell\right)$ и $Q\left(w, X^k\right)$ при увеличении n?

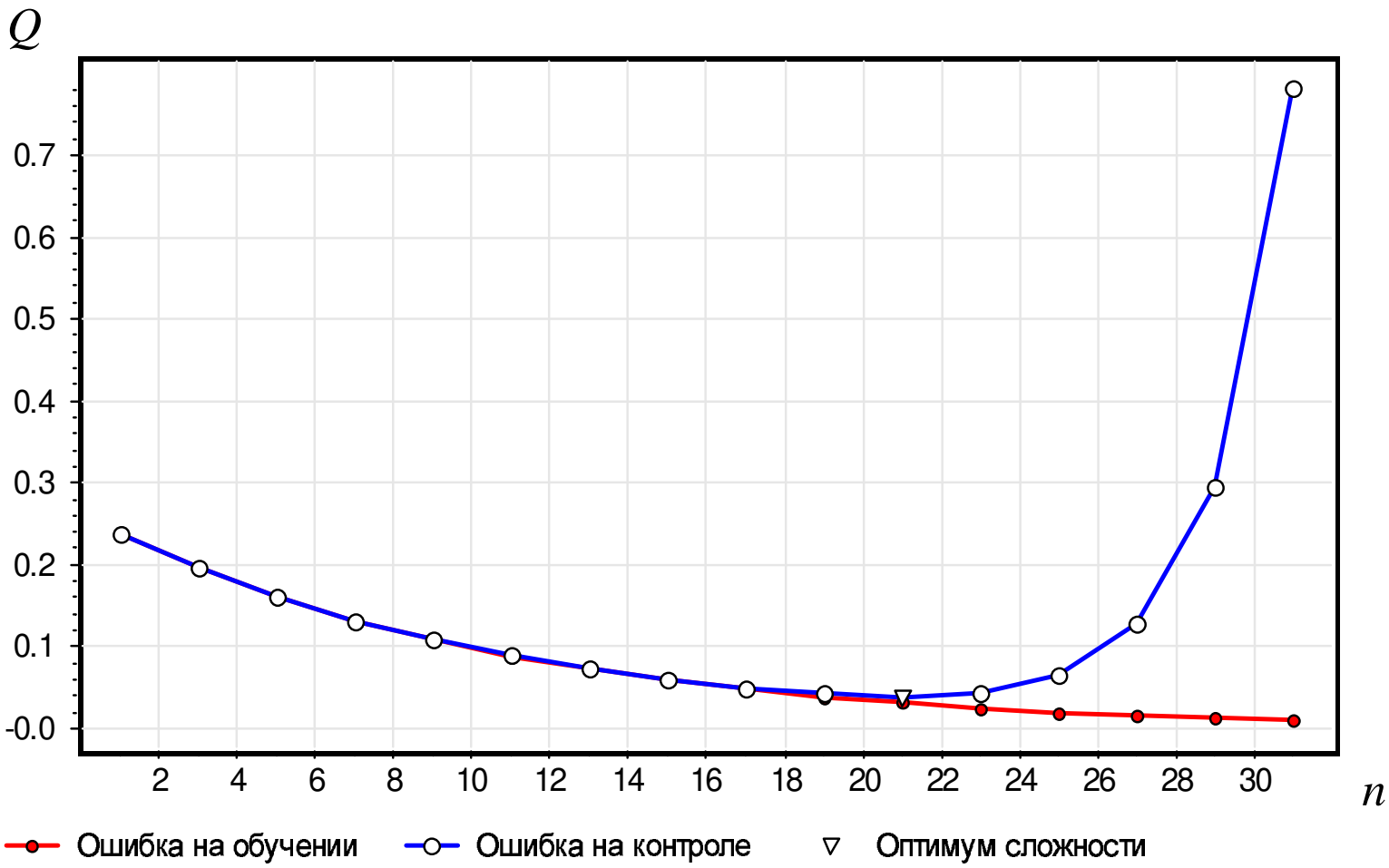
Пример Рунге. Переобучение при $n = 38$, $\ell = 50$

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

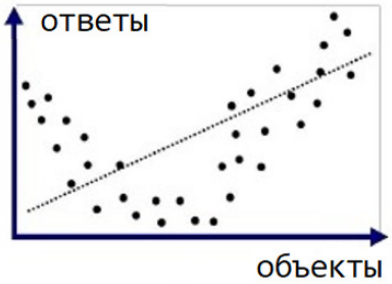


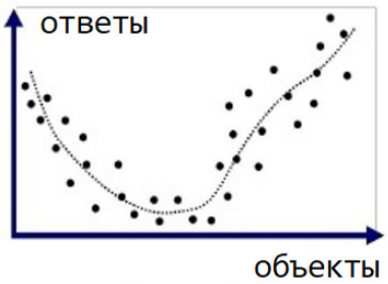
Пример Рунге. Зависимость Q от степени полинома n

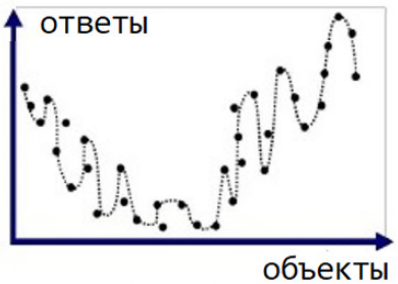
Переобучение — это когда $Q\big(\mu(X^\ell), X^k\big) \gg Q\big(\mu(X^\ell), X^\ell\big)$:



Проблемы недообучения и переобучения



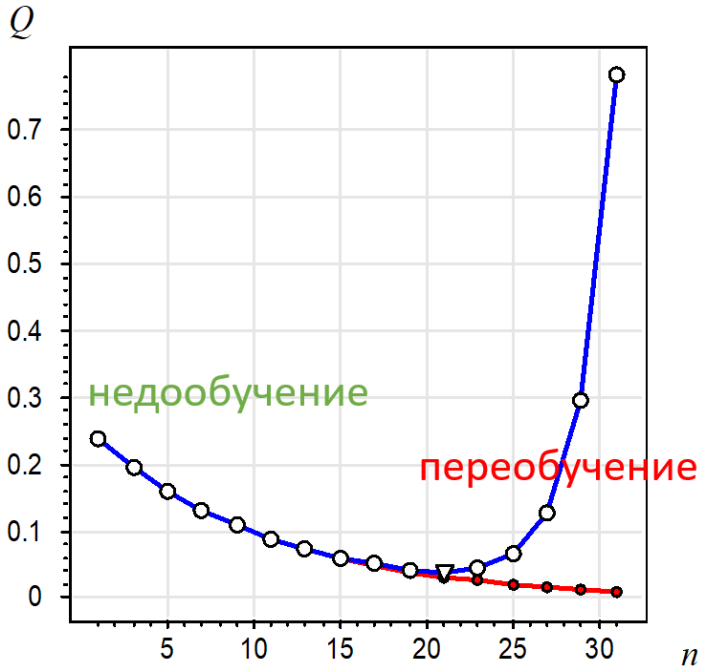




недообучение

переобучение

- Недообучение (underfitting): данных много, параметров недостаточно, модель простая, негибкая
- Переобучение (overfitting): данных мало, параметров слишком много, модель сложная, избыточно гибкая



Переобучение — ключевая проблема в машинном обучении

1. **Из-за чего возникает переобучение?**
   — избыточные параметры в модели $a(x, w)$ «расходуются» на чрезмерно точную подгонку под обучающую выборку
   — выбор a из A производится по неполной информации X^ℓ

2. **Как обнаружить переобучение?**
   — эмпирически, путём разбиения выборки на train и test (на test должны быть известны правильные ответы)

3. **Избавиться от него нельзя. Как его минимизировать?**
   — увеличивать объём обучающих данных (big data)
   — накладывать ограничения на w (регуляризация)
   — минимизировать одну из теоретических оценок
   — выбирать лучшую модель (model selection) по оценкам обобщающей способности (generalization performance)

Эмпирические оценки обобщающей способности

- Эмпирический риск на тестовых данных (hold-out):

$$\mathrm{HO}(\mu, X^\ell, X^k) = Q(\mu(X^\ell), X^k) \to \min$$

- Скользящий контроль (leave-one-out), $L = \ell + 1$:

$$\mathrm{LOO}(\mu, X^L) = \frac{1}{L}\sum_{i=1}^{L} \mathscr{L}(\mu(X^L \setminus \{x_i\}), x_i) \to \min$$

- Кросс-проверка (cross-validation), $L = \ell + k$:

$$\mathrm{CV}(\mu, X^L) = \frac{1}{|P|}\sum_{p \in P} Q(\mu(X_p^\ell), X_p^k) \to \min$$

где P — множество разбиений $X^L = X_p^\ell \sqcup X_p^k$

Задачи медицинской диагностики

Объект — пациент в определённый момент времени.

Классы: диагноз или способ лечения или исход заболевания.

Примеры признаков:

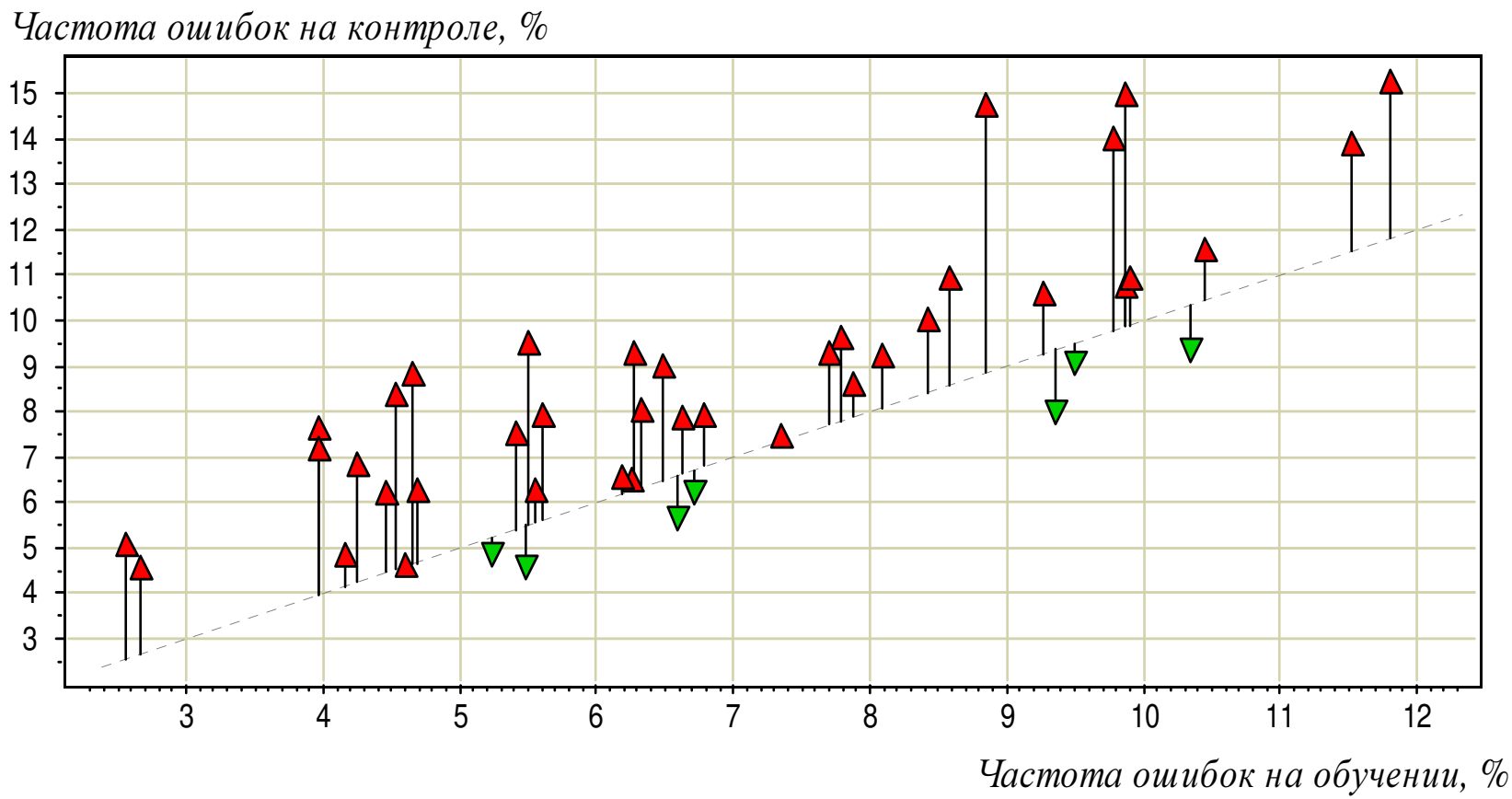
- **бинарные:** пол, головная боль, слабость, тошнота, и т. д.
- **порядковые:** тяжесть состояния, желтушность, и т. д.
- **количественные:** возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т. д.

Особенности задачи:

- обычно много «пропусков» в данных;
- нужна интерпретируемая модель классификации;
- нужно выделять *синдромы* — сочетания *симптомов*;
- нужна оценка вероятности отрицательного исхода.

Задача медицинской диагностики. Пример переобучения

Задача предсказания отдалённого результата хирургического лечения атеросклероза; точки — различные решающие правила



Задача категоризации текстовых документов

Объект — текстовый документ.

Классы — рубрики иерархического тематического каталога.

Примеры признаков:

- **номинальные:** автор, издание, год, и т. д.
- **количественные:** для каждого термина — частота в тексте, в заголовках, в аннотации, и т. д.

Особенности задачи:

- лишь небольшая часть документов имеют метки y_i;
- документ может относиться к нескольким рубрикам;
- в каждом ребре дерева свой классификатор на 2 класса.

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 ранжирования поисковой выдачи

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

Классы — релевантен или не релевантен,
разметка делается людьми — асессорами.

Примеры количественных признаков:

- частота слов запроса в документе,
- число ссылок на документ,
- число кликов на документ: всего, по данному запросу.

Особенности задачи:

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

Резюме. Философия машинного обучения

Машинное обучение — автоматизация научного метода

— *наблюдения, измерения* → выборка данных

— *гипотеза* → модель, параметрическое семейство функций

— *верифицируемость* → обучение (train) путём оптимизации

— *фальсифицируемость* → проверка (test) на новых данных

— *соответствие* → процесс постепенного усложнения модели

— *бритва Оккама* → своевременное прекращение усложнений

— *воспроизводимость* → культура open data / open source

— *честность* → анализ границ применимости, хрупкости модели

Постановка задачи — это ДНК (Дано, Найти, Критерий)

Основные понятия машинного обучения:

— объект, ответ, признак, модель, функция потерь, эмпирический риск, метод обучения, переобучение

Прикладные задачи машинного обучения:

— очень много, очень разных, во всех областях деятельности