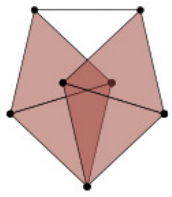
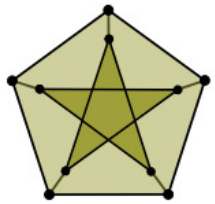
Теория и практика машинного обучения
• Лекция 2 •
Методы классификации и регрессии

Воронцов Константин Вячеславович
МФТИ • МГУ • ВШЭ • ВЦ РАН • Яндекс • FORECSYS



Комбинаторика и алгоритмы
для школьников



• Летняя школа — 2014 •
22 августа 2014

Содержание

Задача статистического (машинного) обучения с учителем

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

Дано: векторы $x_i = (x_i^1, \ldots, x_i^n)$ — объекты обучающей выборки,
$y_i = f(x_i)$ — правильные ответы, $i = 1, \ldots, \ell$:

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

Найти: классификатор $a(x)$, способный давать правильные
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:

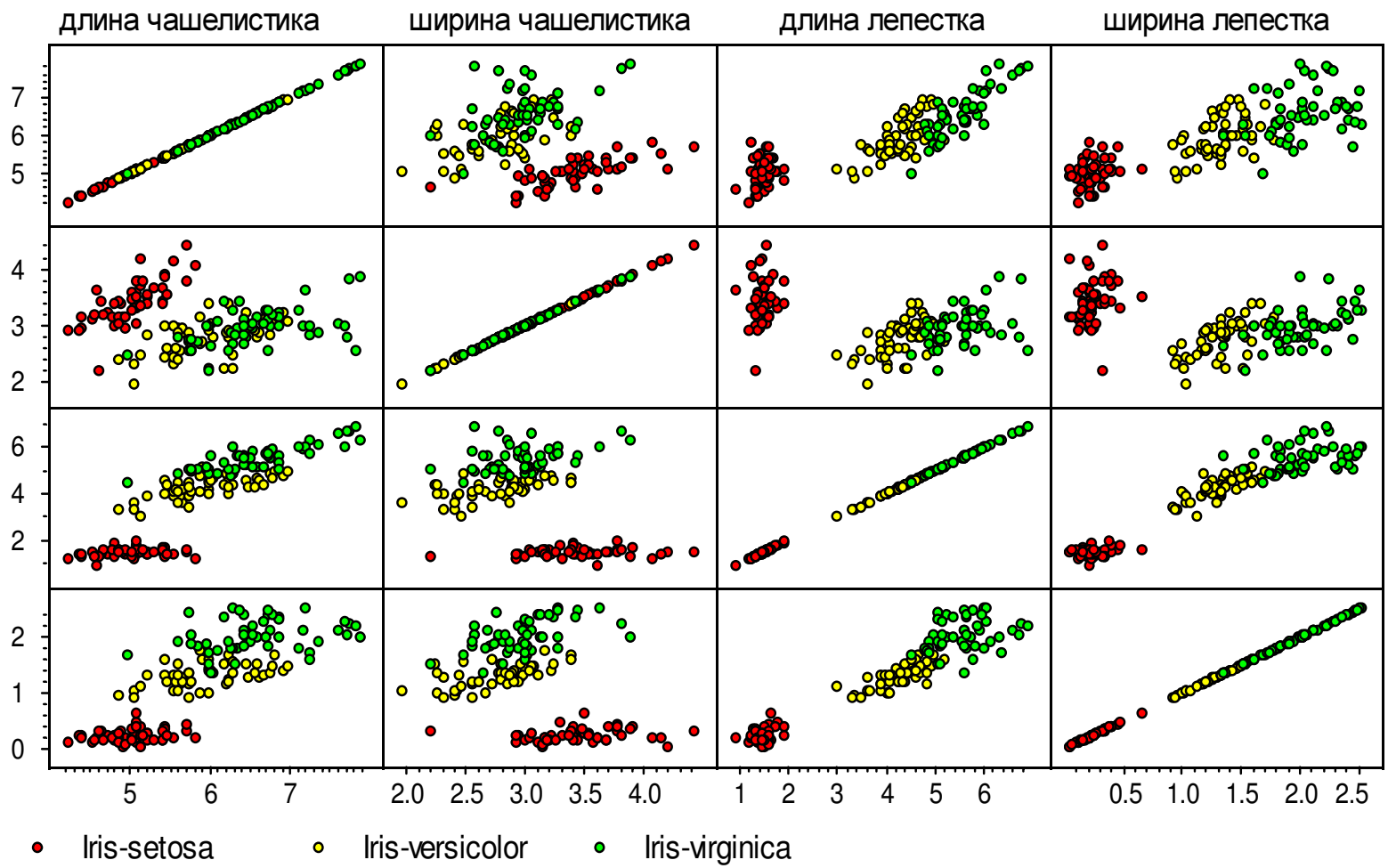
- $D_j = \{0, 1\}$ — *бинарный* признак;
- $|D_j| < \infty$ — *номинальный* признак;
- D_j упорядочено — *порядковый* признак;
- $D_j = \mathbb{R}$ — *количественный* признак.

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

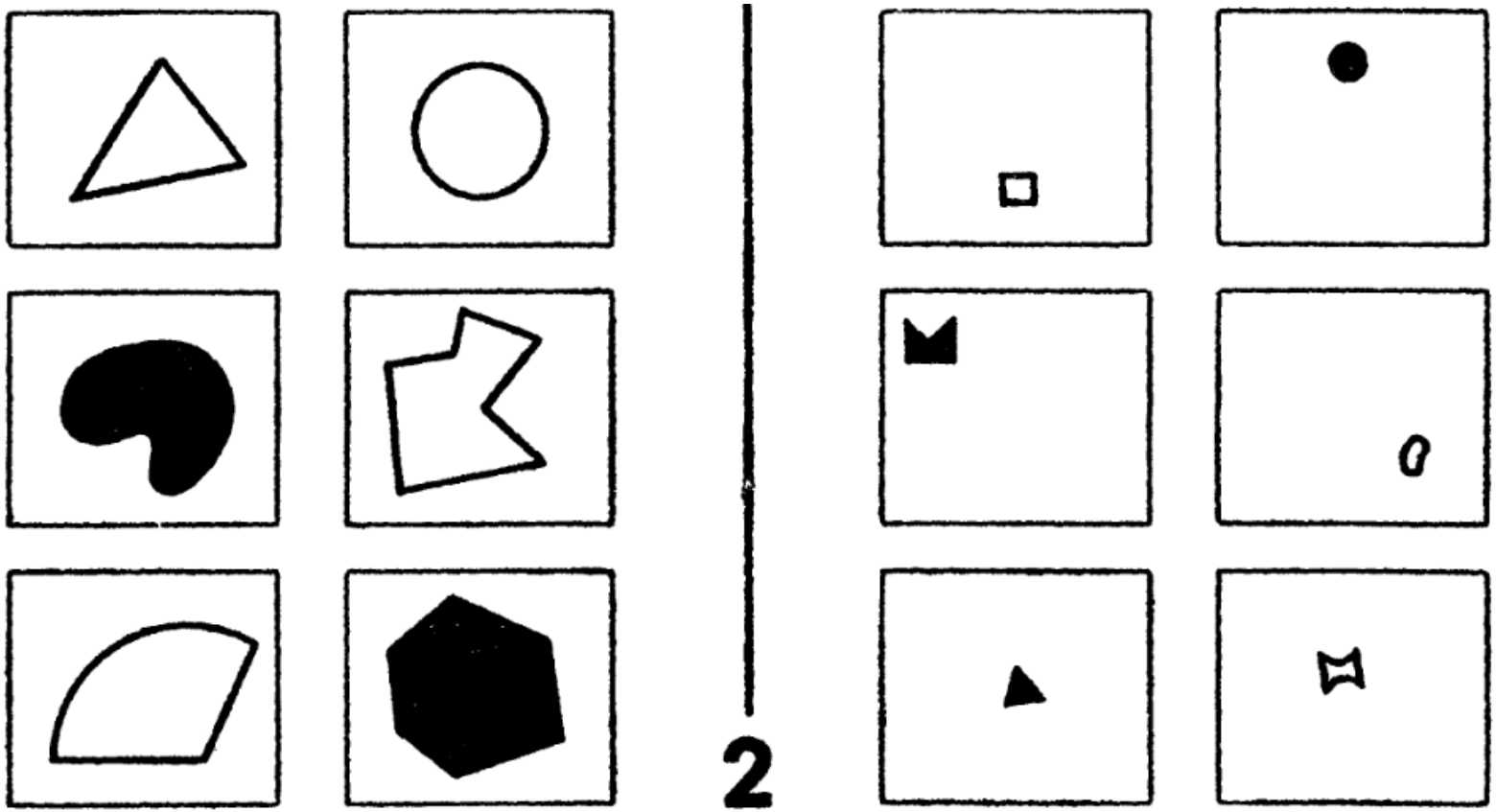
- $Y = \{0, 1\}$ или $Y = \{-1, +1\}$ — *классификация* на 2 класса;
- $Y = \{1, \ldots, M\}$ — на M непересекающихся классов;
- $Y = \{0, 1\}^M$ — на M классов, которые могут пересекаться;
- $Y = \mathbb{R}$ — задача восстановления регрессии;
- Y упорядочено — задача ранжирования (learning to rank).

Пример: задача классификации цветков ириса [Фишер, 1936]

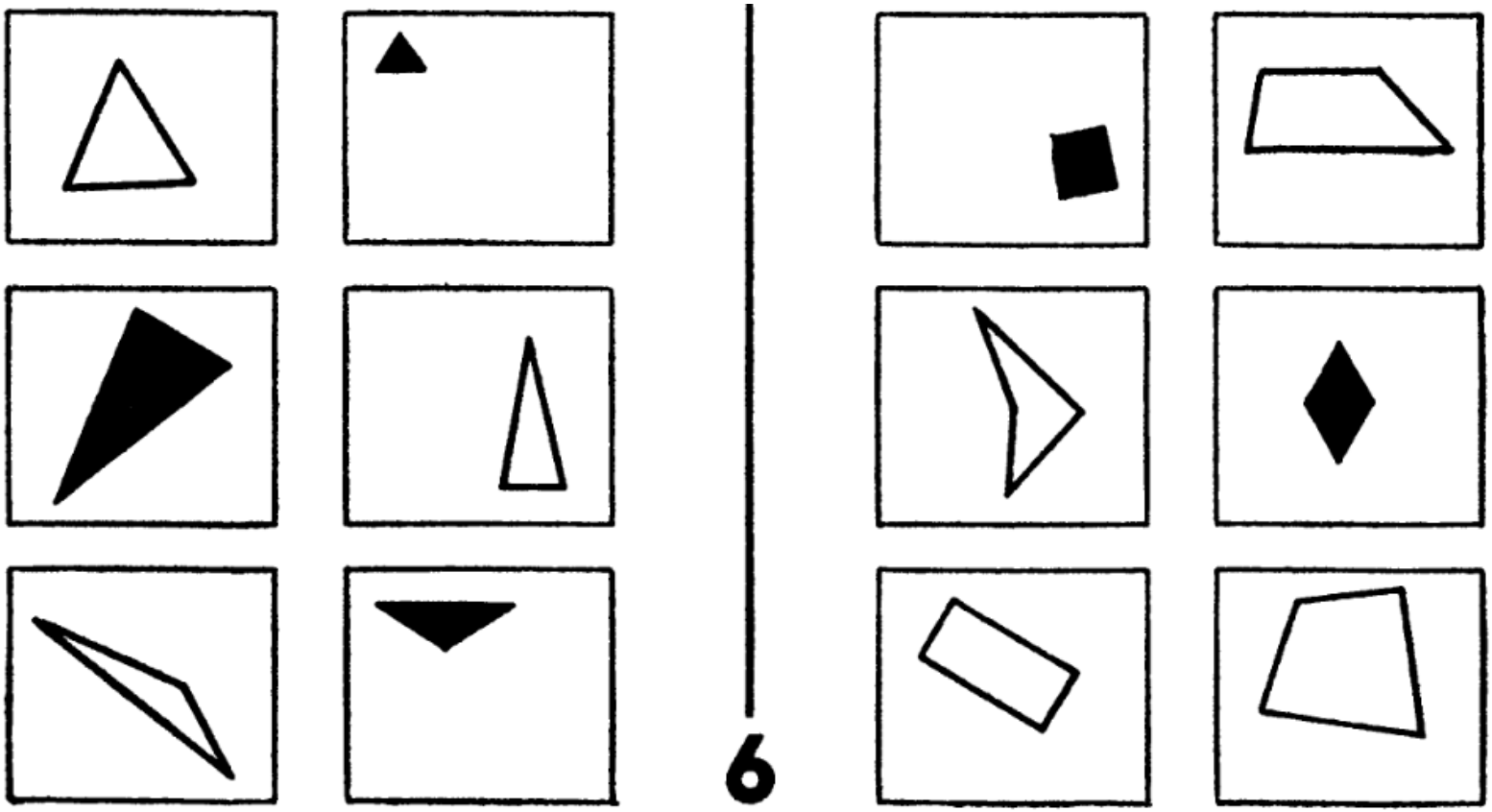
$n = 4$ признака, $|Y| = 3$ класса, длина выборки $\ell = 150$.



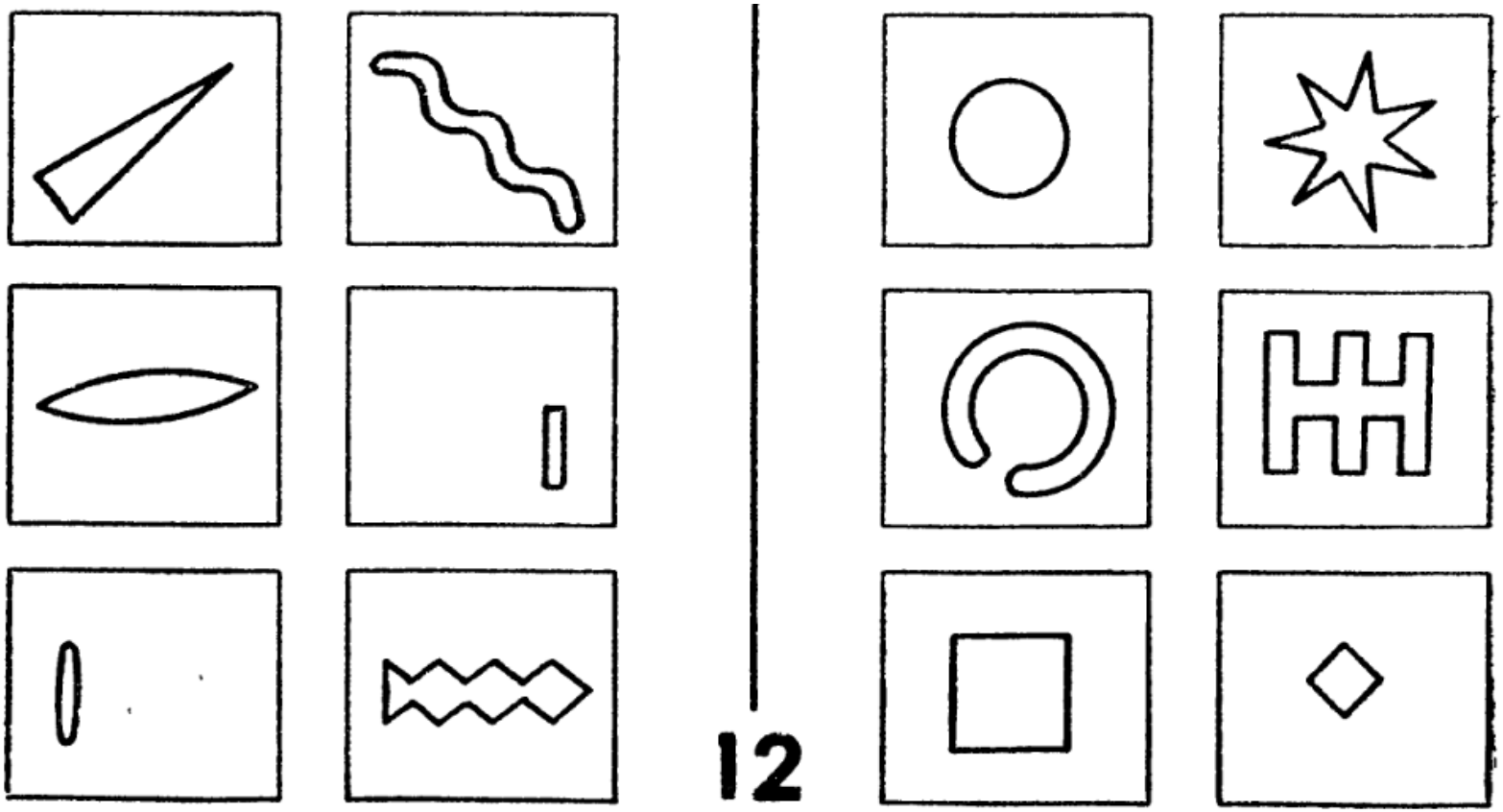
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



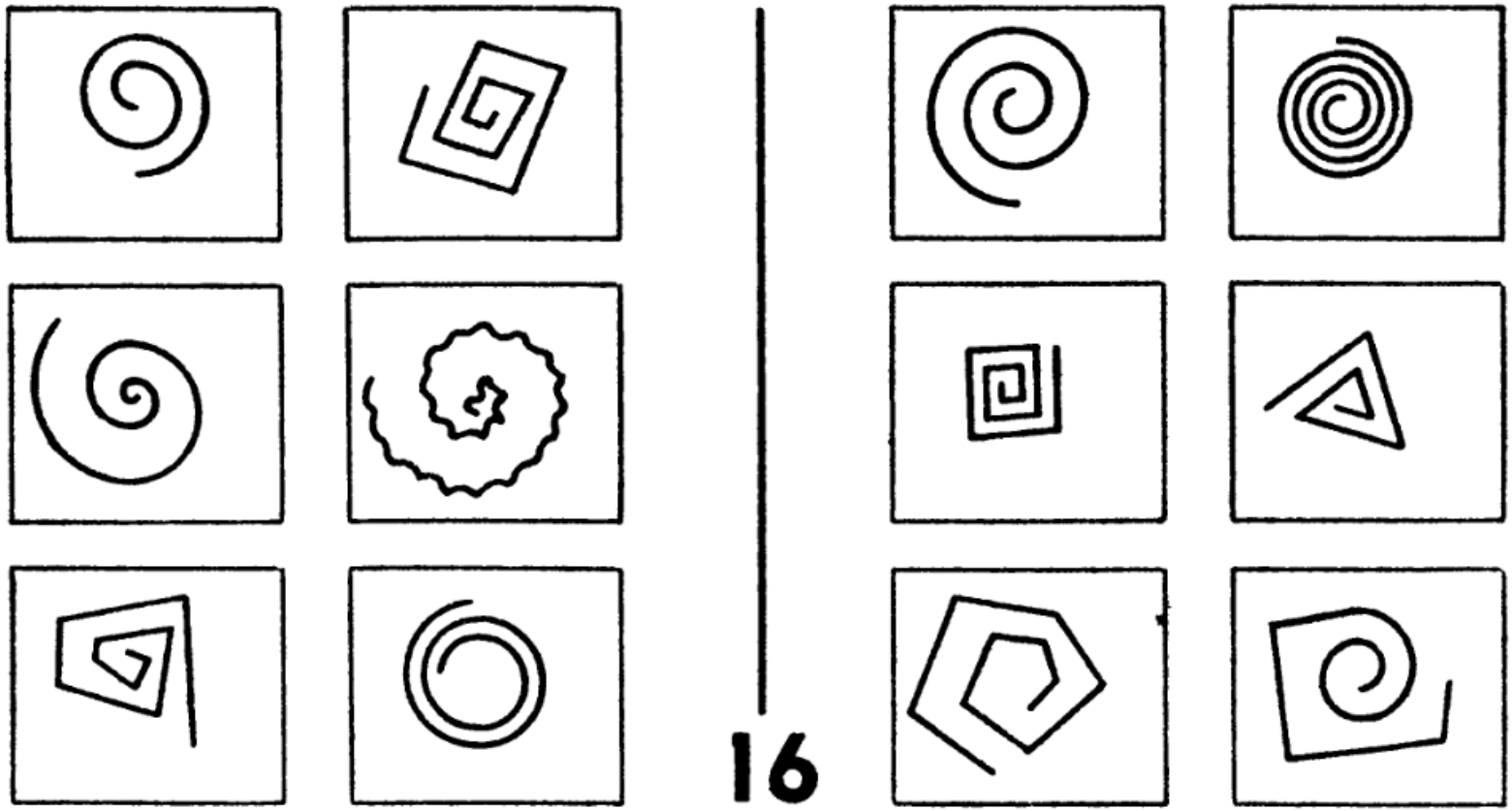
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



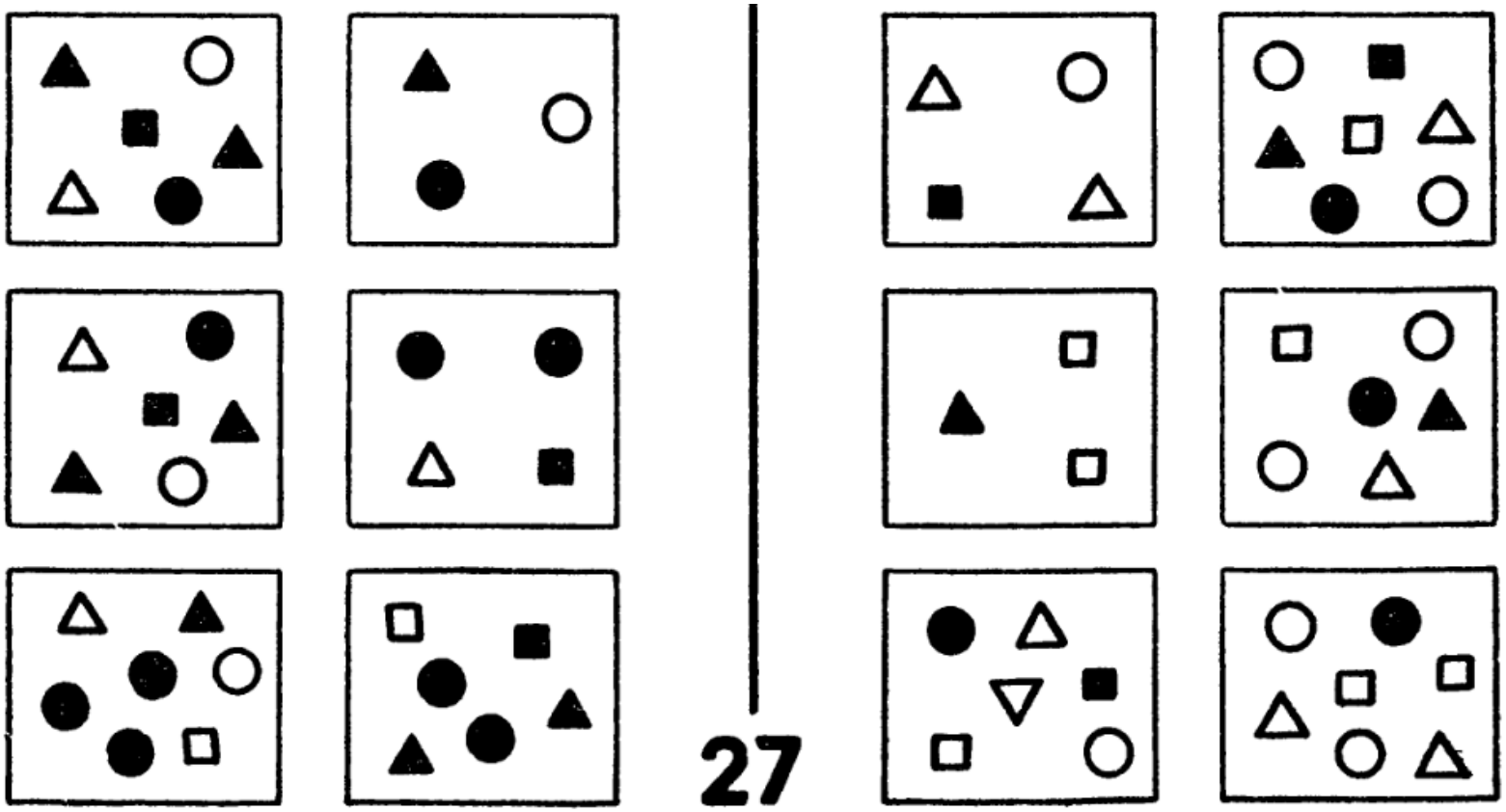
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



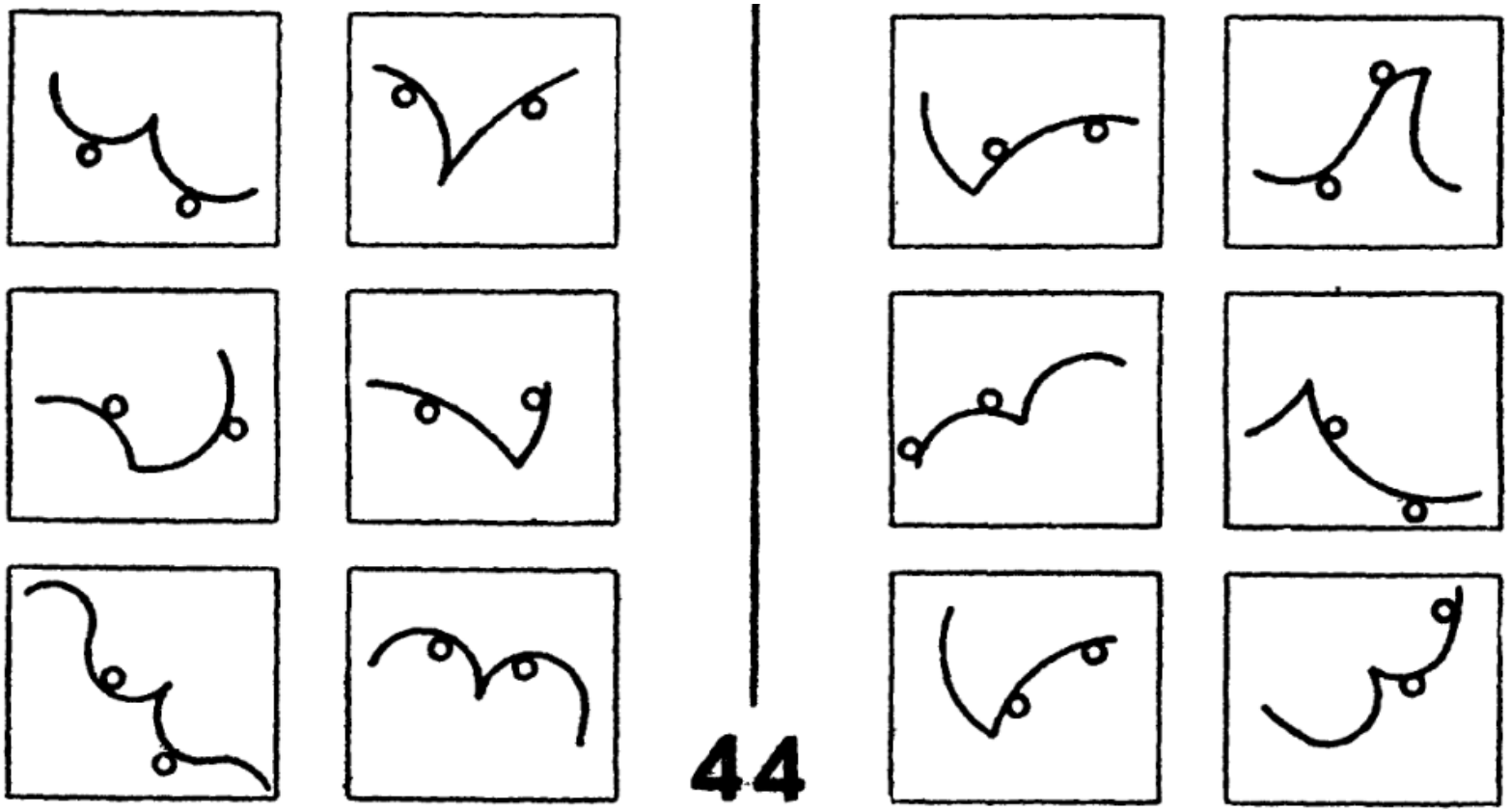
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



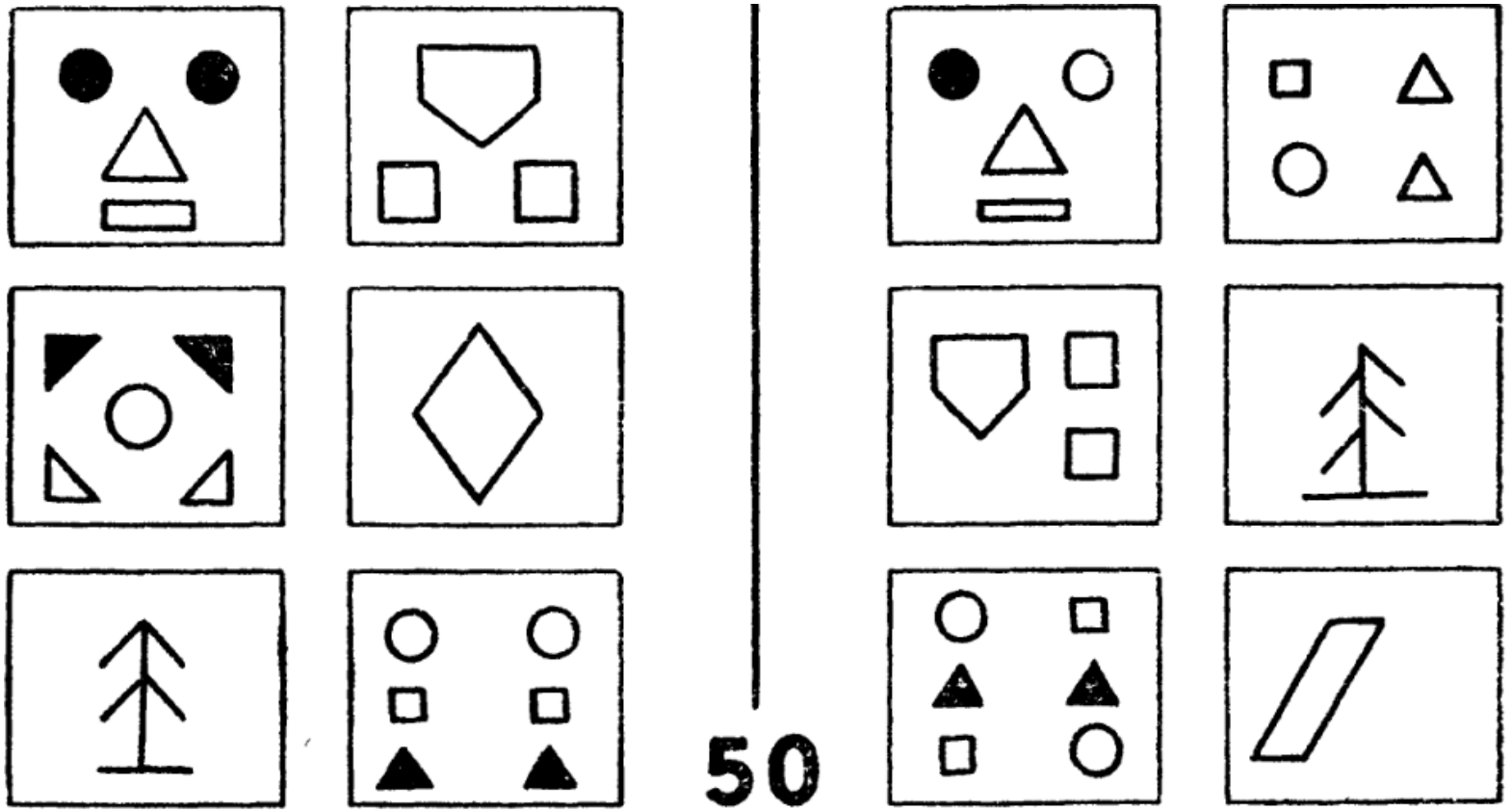
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



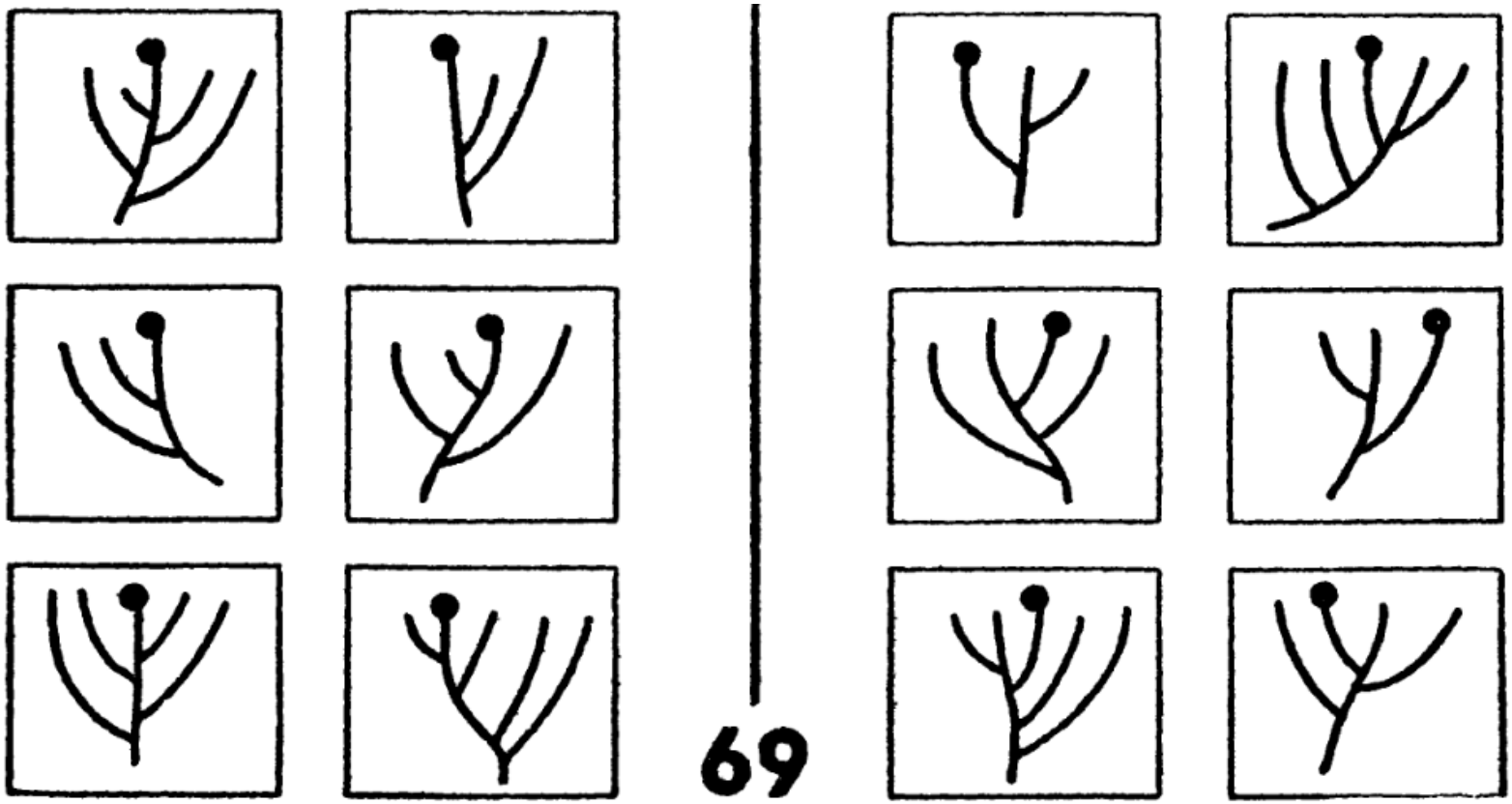
Тесты М. М. Бонгарда [Проблема узнавания, 1967]



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



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



Задачи медицинской диагностики

Объект — пациент в определённый момент времени.

Классы: диагнозы или способы лечения или исходы.

Примеры признаков:

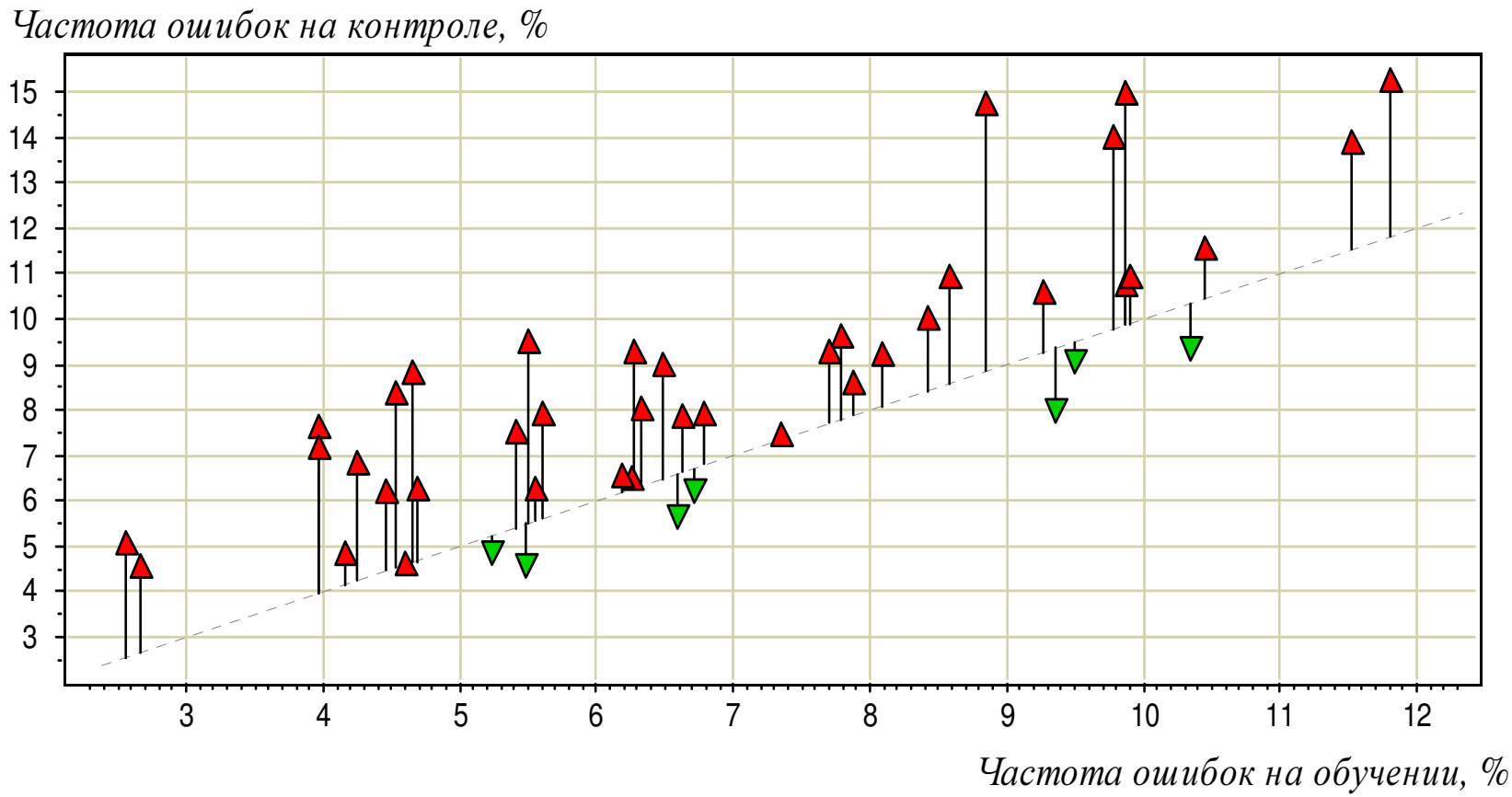
- **бинарные:** пол, головная боль, слабость, тошнота, и т. д.
- **порядковые:** тяжесть состояния, желтушность, и т. д.
- **количественные:** возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т. д.

Особенности задачи:

- обычно много «пропусков» в данных;
- нужен интерпретируемый алгоритм классификации;
- нужна оценка вероятности ошибки.

Пример переобучения. Реальная задача классификации

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



Задача кредитного скоринга

Объект — заявка на выдачу кредита.

Классы — bad или good.

Примеры признаков:

- **бинарные:** пол, наличие телефона, и т. д.
- **номинальные:** место проживания, профессия, работодатель, и т. д.
- **порядковые:** образование, должность, и т. д.
- **количественные:** возраст, зарплата, стаж работы, доход семьи, сумма кредита, и т. д.

Особенности задачи:

- нужно оценивать вероятность дефолта $P(\text{bad})$.

Задача предсказания оттока клиентов

Объект — абонент в определённый момент времени.

Классы — уйдёт или не уйдёт в следующем месяце.

Примеры признаков:

- **бинарные:** корпоративный клиент, включение услуг, и т. д.
- **номинальные:** тарифный план, регион проживания, и т. д.
- **количественные:** длительность разговоров (входящих, исходящих, СМС, межгород, и т. д.), частота оплаты, и т. д.

Особенности задачи:

- нужно строить признаки по потоку действий абонентов;
- нужно оценивать вероятность ухода;
- сверхбольшие выборки.

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 регрессии: прогноз стоимости недвижимости

Объект — квартира в Москве.

Примеры признаков:

- **бинарные:** наличие балкона, лифта, мусоропровода, охраны, гаража, чердака, 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 $\langle$запрос, документ$\rangle$.

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

Примеры признаков:

- **количественные:**
  частота слов запроса в документе,
  число ссылок на документ,
  число кликов на документ: всего, по данному запросу, и т. д.

Особенности задачи:

- нужно строить признаки по разнородным сырым данным;
- оптимизируется не число ошибок, а качество ранжирования;
- сверхбольшие выборки.

Обучение регрессии — это оптимизация

Задача регрессии, $Y = \mathbb{R}$

1. Выбираем *модель регрессии*, например, линейную:
$$a(x, w) = \langle x, w \rangle = \sum_{j=1}^{n} x^j w_j, \qquad x, w \in \mathbb{R}^n$$
2. Выбираем функцию потерь, например, квадратичную:
$$\mathscr{L}(a, y) = (a - y)^2$$
3. Минимизируем потери *методом наименьших квадратов*:
$$Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} \big(a(x_i, w) - y_i\big)^2 \to \min_w$$
4. Проверяем прогностическую (обобщающую) способность:
$$\tilde{Q}(w) = \frac{1}{k} \sum_{i=1}^{k} \big(a(\tilde{x}_i, w) - \tilde{y}_i\big)^2$$

Обучение классификации — тоже оптимизация

Задача классификации, $Y = \{-1, +1\}$

1. Выбираем *модель классификации*, например, линейную:
$$a(x, w) = \text{sign}\langle x, w\rangle$$
2. Выбираем функцию потерь, например, *бинарную*:
$$\mathscr{L}(a, y) = \big[a(x_i, w)y_i < 0\big]$$
3. Минимизируем *частоту ошибок* на обучающей выборке:
$$Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell}\big[a(x_i, w)y_i < 0\big] \to \min_w$$
4. Проверяем прогностическую (обобщающую) способность:
$$\tilde{Q}(w) = \frac{1}{k}\sum_{i=1}^{k}\big[a(\tilde{x}_i, w)\tilde{y}_i < 0\big]$$

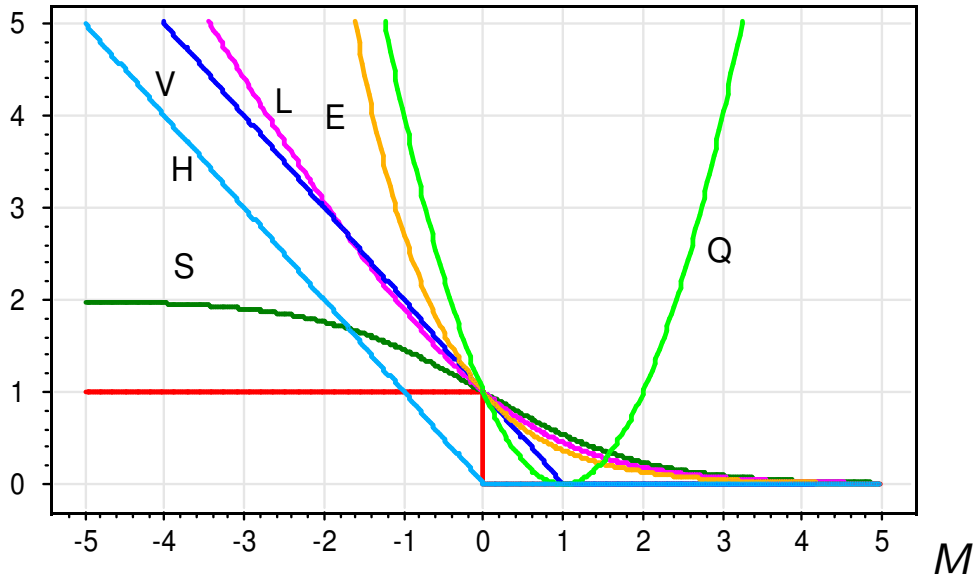
Обучение классификации — сглаживание функции потерь

Задача классификации, $Y = \{-1, +1\}$

1. Выбираем *модель классификации*, например, линейную:
$$a(x, w) = \text{sign}\langle x, w\rangle$$
2. Мажорируем пороговую функцию потерь непрерывной:
$$\big[M_i < 0\big] \leqslant \mathscr{L}\big(M_i\big), \qquad M_i = \langle x_i, w\rangle y_i \;\; — \;\; \text{отступ (margin)}$$
3. Минимизируем *сглаженную частоту ошибок*:
$$Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathscr{L}\big(\langle x_i, w\rangle y_i\big) \to \min_w$$
4. Проверяем прогностическую (обобщающую) способность:
$$\tilde{Q}(w) = \frac{1}{k}\sum_{i=1}^{k} \big[\langle \tilde{x}_i, w\rangle \tilde{y}_i < 0\big]$$

Непрерывные аппроксимации пороговой функции потерь

Часто используемые непрерывные функции потерь $\mathscr{L}(M)$:



$V(M) = (1 - M)_+$ — кусочно-линейная (SVM)
$H(M) = (-M)_+$ — кусочно-линейная (Hebb's rule)
$L(M) = \log_2(1 + e^{-M})$ — логарифмическая (Logistic Regression)
$Q(M) = (1 - M)^2$ — квадратичная (Fisher's Linear Discriminant)
$S(M) = 2(1 + e^M)^{-1}$ — сигмоидная (Artificial Neural Network)
$E(M) = e^{-M}$ — экспоненциальная (AdaBoost)
$[M < 0]$ — пороговая функция потерь.

Общие подходы к решению оптимизационных задач

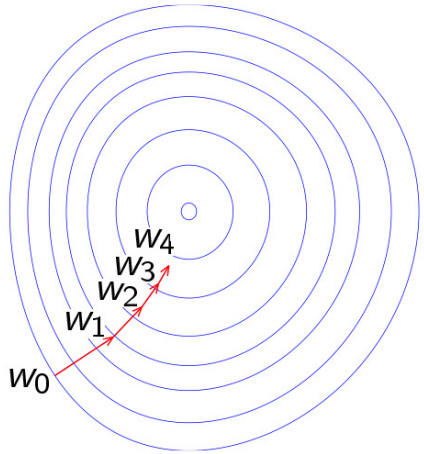
Аналитический метод (наименьших квадратов):
Если w — точка минимума *гладкой* функции $Q(w)$, то

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

Это система n уравнений с n неизвестными.

Численный метод (градиентного спуска):

1 начальное приближение w^0, $t := 0$;
2 **повторять**
3 $\quad w_j^{t+1} := w_j^t - h^t \cdot \frac{\partial Q(w^t)}{\partial w_j}$, $j = 1, \ldots, n$;
4 $\quad t := t + 1$;
5 **пока** процесс не сойдётся;

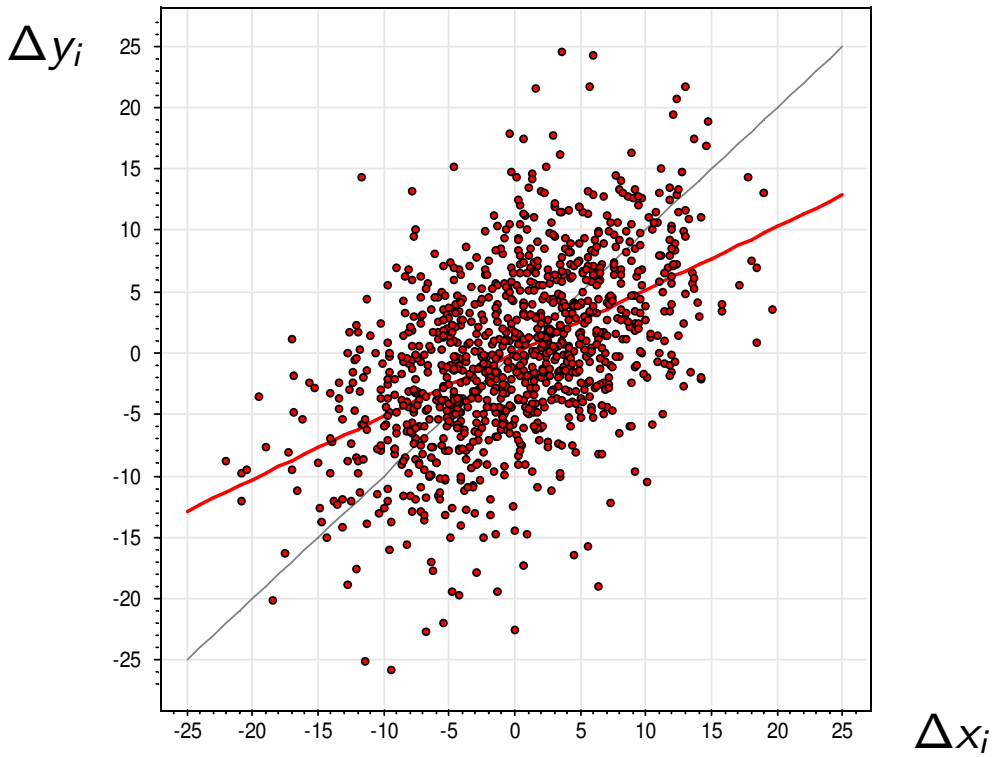


Откуда пошло название «регрессия»

Исследование наследственности роста [Фрэнсис Гальтон, 1886]

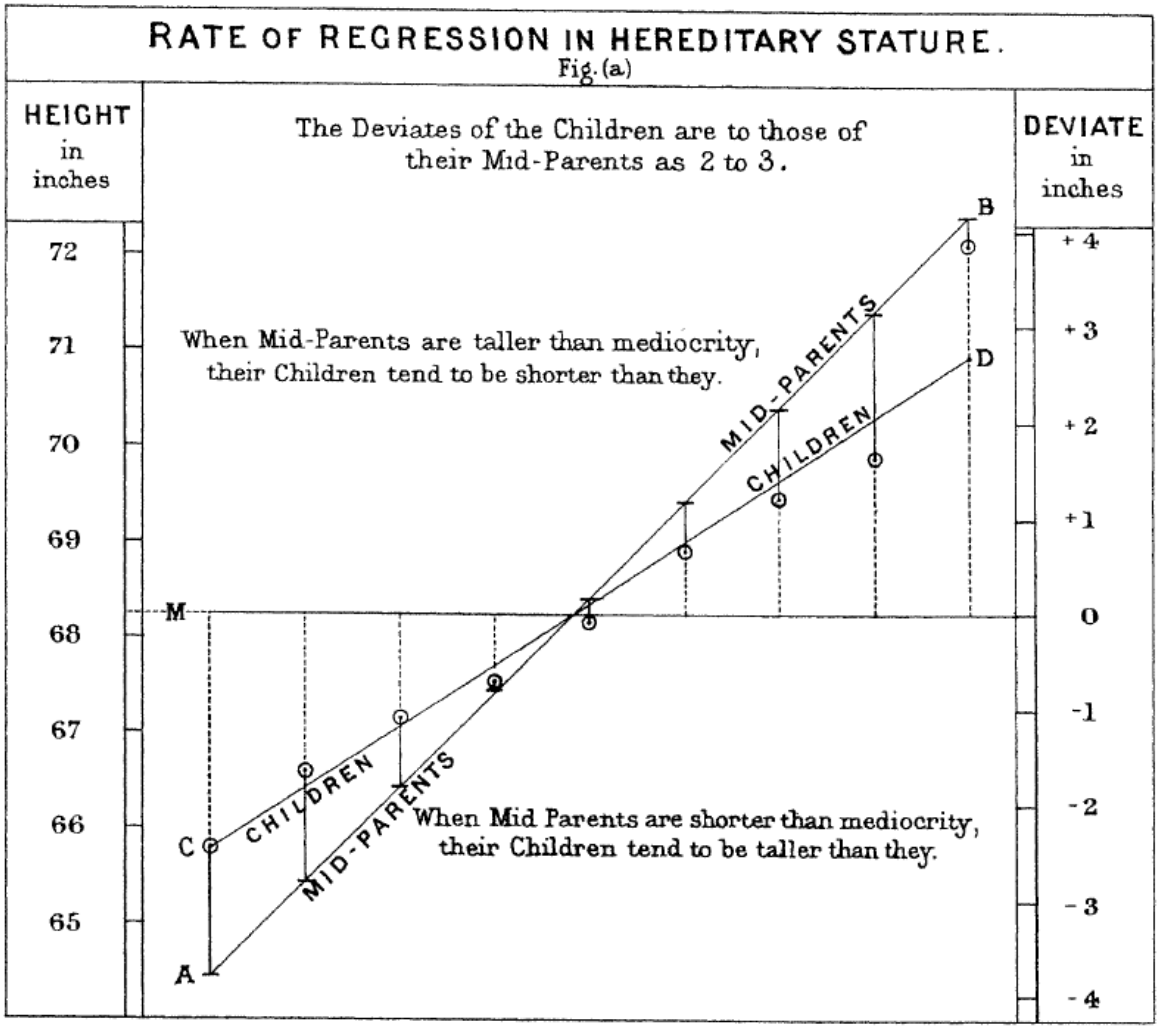
Δx_i — отклонение роста отца от среднего в популяции

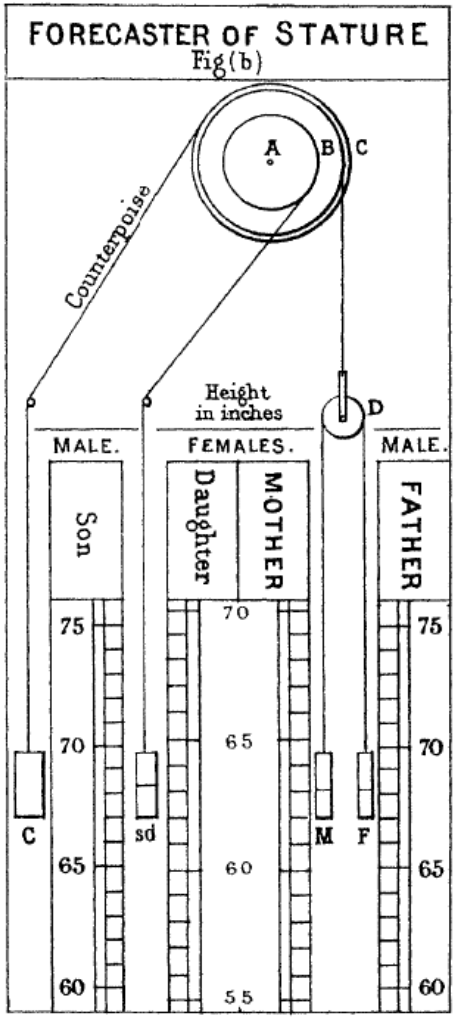
Δy_i — отклонение роста взрослого сына от среднего в популяции



Regression to mediocrity — возвращение к посредственности

Francis Galton (1886). Regression Towards Mediocrity in Hereditary Stature





Задача проведения прямой через заданные точки

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

Найти: α, β в линейной модели $y = \alpha x + \beta$

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

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:

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

Метод стохастического градиента (SG, Stochastic Gradient)

Задача классификации: $y_i \in \{-1, +1\}$, $a(x, w) = \text{sign}\langle w, x\rangle$.
Минимизация сглаженной частоты ошибок:

$$Q(w) = \sum_{i=1}^{\ell} \mathscr{L}\big(\langle w, x_i\rangle y_i\big) \to \min_w .$$

Один шаг *градиентного спуска*:

$$w^{t+1} := w^t - h^t \sum_{i=1}^{\ell} \mathscr{L}'\big(\langle w^t, x_i\rangle y_i\big)\, x_i\, y_i.$$

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

$$w^{t+1} := w^t - h^t \mathscr{L}'\big(\langle w^t, x_i\rangle y_i\big)\, x_i\, y_i.$$

Алгоритм SG (Stochastic Gradient)

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

1 инициализировать веса w_j, $j = 1, \ldots, n$;
2 **повторять**
3 выбрать случайный объект (x_i, y_i) из обучающей выборки;
4 выбрать величину градиентного шага h;
5 выполнить градиентный шаг:
$w_j := w_j - h\mathscr{L}'\big(\langle w, x_i\rangle y_i\big)x_i^j y_i$ для всех $j = 1, \ldots, n$;
6 **пока** процесс не сойдётся куда-нибудь;

Преимущества и недостатки:

⊕ можно брать какие угодно модели и функции потерь $\mathscr{L}$
⊕ хорошо работает на больших выборках
⊖ возможно застревание в локальных экстремумах

Эвристики

- Выбор начального приближения, например, так:

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

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

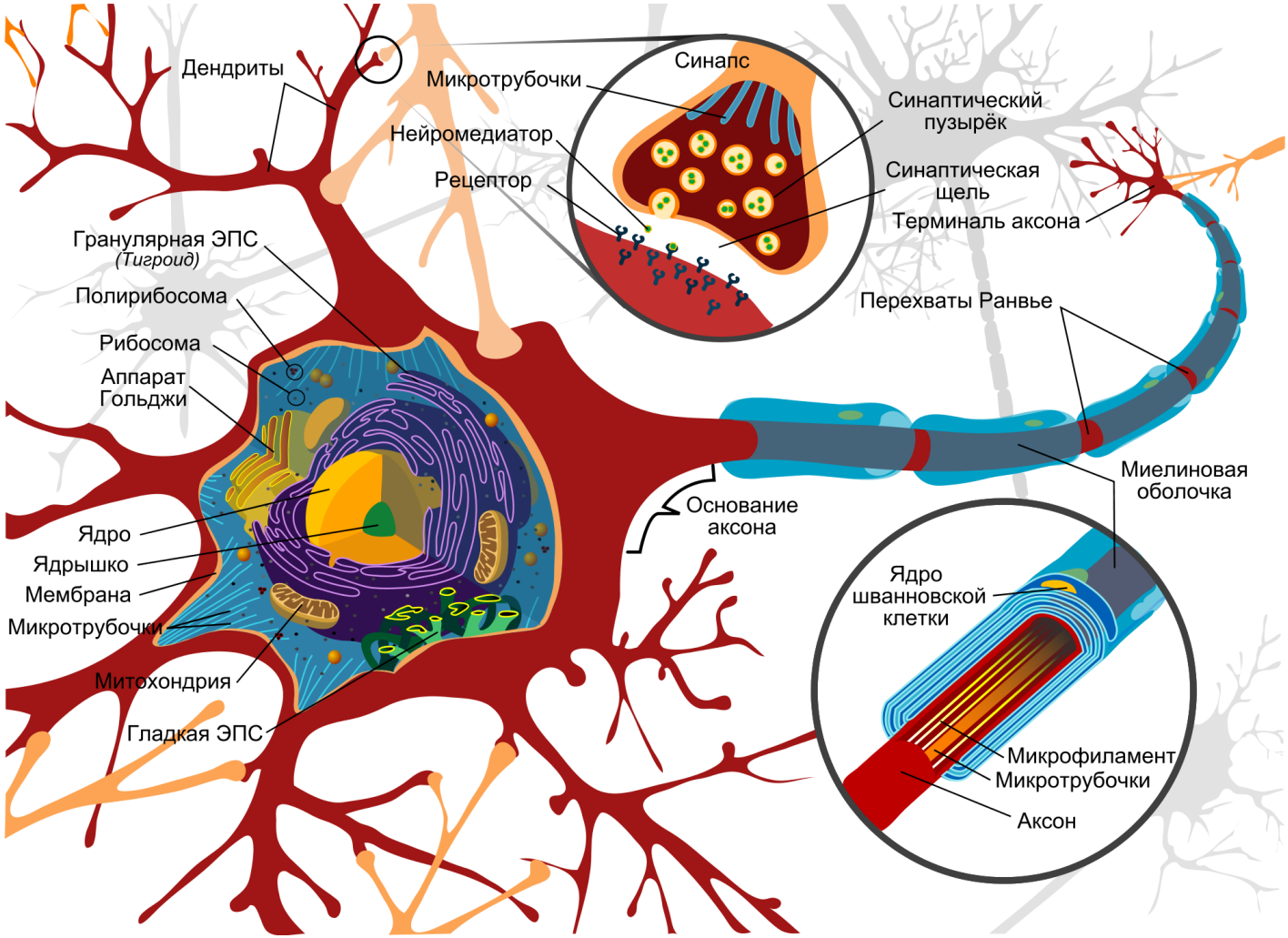
- Выбор темпа обучения (градиентного шага) h^t:
сходимость гарантируется для выпуклых $Q(w)$ при

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

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

- Выбор порядка предъявления объектов:
— случайно, но попеременно из разных классов;
— чаще брать пограничные объекты с малым $|M_i|$;

Похож ли нейрон на линейный классификатор?

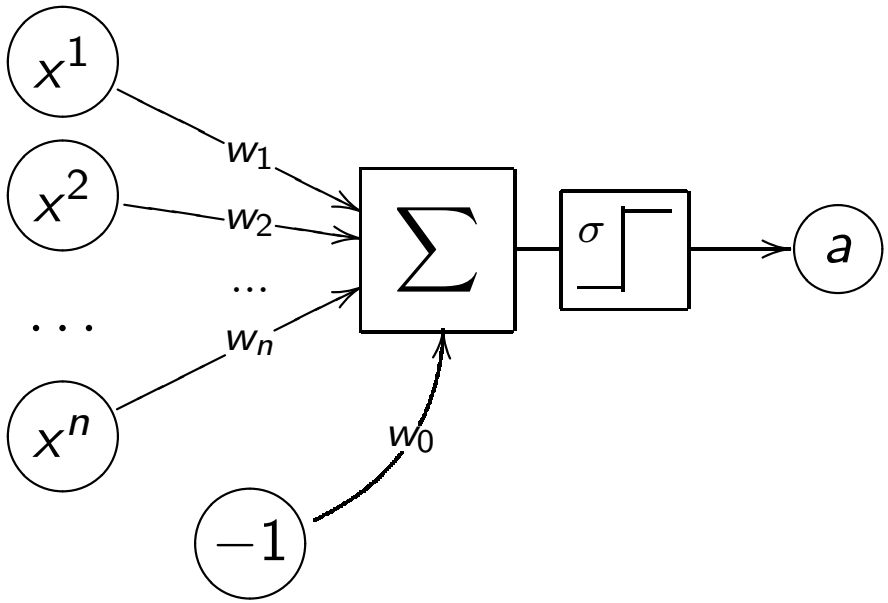


Математическая модель нейрона

Линейная модель нейрона МакКаллока-Питтса [1943]:

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

где $\sigma(z)$ — функция активации (например, $\operatorname{sign} z$ или $\operatorname{arcth} z$).



Частные случаи метода стохастического градиента

Задача регрессии:

Дельта-правило (delta-rule) [Видроу и Хофф, 1960]

$y_i \in \mathbb{R}$, $\mathscr{L}(a, y) = (a - y)^2$, $a(x, w) = \langle w, x \rangle$;

$\Delta_i = \big(\langle w, x_i \rangle - y_i\big)$ — ошибка алгоритма $a(x, w)$ на объекте x_i;

$$w := w - h\big(\langle w, x_i \rangle - y_i\big)x_i.$$

Задача классификации:

Правило Хэбба [1949]

$y_i \in \{-1, +1\}$, $a(x, w) = \text{sign}\langle w, x \rangle$, $\mathscr{L}(a, y) = (-\langle w, x \rangle y)_+$;

$$w := w + h x_i y_i \big[\langle w, x_i \rangle y_i < 0\big].$$

Правило из метода опорных векторов, SVM

$y_i \in \{-1, +1\}$, $a(x, w) = \text{sign}\langle w, x \rangle$, $\mathscr{L}(a, y) = (1 - \langle w, x \rangle y)_+$;

$$w := w + h x_i y_i \big[\langle w, x_i \rangle y_i < 1\big].$$

Причины переобучения линейных моделей

1. слишком мало объектов; слишком много признаков;
2. линейная зависимость (мультиколлинеарность) признаков:
   пусть построен классификатор: $a(x, w) = \text{sign}\langle x, w\rangle$;
   мультиколлинеарность: $\exists v \in \mathbb{R}^n$: $\forall x \;\; \langle x, v\rangle \approx 0$;
   тогда $\forall \gamma \in \mathbb{R} \quad a(x, w) \approx \text{sign}\langle x, w + \gamma v\rangle$

Последствия:

- решение неединственно и неустойчиво;
- появляются слишком большие веса $+w_j$ или $-w_j$;
- $Q(w)$ на обучении много меньше, чем на контроле;

Спасает *регуляризация* — введение дополнительного критерия:

$$\|w\|^2 = \sum_{j=1}^{n} w_j^2 \;\to\; \min\,.$$

Метод сокращения весов (weight decay)

Штраф за увеличение нормы вектора весов:

$$Q_\tau(w) = Q(w) + \frac{\tau}{2}\|w\|^2 \to \min_w .$$

Градиент:

$$\frac{\partial}{\partial w_j} Q_\tau(w) = \frac{\partial}{\partial w_j} Q(w) + \tau w_j .$$

Модификация градиентного шага:

$$w_j^{t+1} := w_j^t (1 - h^t \tau) - h^t \frac{\partial}{\partial w_j} Q(w^t).$$

Параметр регуляризации τ подбирается экспериментально, по качеству на контрольной выборке.

Явная максимизация сглаженного AUC

AUC равна доле правильно упорядоченных пар (x_i, x_s):

$$\text{AUC} = \frac{1}{\ell_0 \ell_1} \sum_{i=1}^{\ell} \sum_{s=1}^{\ell} \big[y_i < y_s\big] \big[\langle x_i, w\rangle < \langle x_s, w\rangle\big] \;\to\; \max_w,$$

Максимизация сглаженного AUC:

$$Q(w) = \sum_{i,s:\; y_i < y_s} \mathscr{L}\big(\underbrace{\langle x_s, w\rangle - \langle x_i, w\rangle}_{M_{is}}\big) \;\to\; \min_w,$$

$\mathscr{L}(M)$ — гладкая убывающая функция отступа,
M_{is} — новое понятие отступа, теперь для пар объектов.

Метод стохастического градиента по парам (x_i, x_s): $y_i < y_s$

$$w^{t+1} := w^t - h^t \mathscr{L}'\big(\langle x_s - x_i, w^t\rangle\big)(x_s - x_i);$$

Давайте соберём вместе всё, что мы узнали

... чтобы решать конкурсную задачу про ЭКГ :)

Вход: выборка $(x_i, y_i)_{i=1}^{\ell}$, $y_i \in \{-1, +1\}$;
Выход: веса $w_1, \ldots, w_n$;

1 инициализировать веса $w_j = \frac{\langle y, f_j \rangle}{\langle f_j, f_j \rangle}$, $j = 1, \ldots, n$;
2 **повторять**
3 выбрать случайную пару обучающих объектов x_i, x_s из разных классов: $y_i = -1$ (больной), $y_s = 1$ (здоровый);
4 выбрать величину градиентного шага h;
5 выполнить градиентный шаг:
для всех $j = 1, \ldots, n$
$w_j := w_j(1 - \tau h) + h\big[\langle x_s - x_i, w \rangle < 1\big](x_s^j - x_i^j)$;
6 **пока** процесс не сойдётся куда-нибудь;

Резюме

- Практических задач классификации и регрессии много!
- Обучение — это оптимизации (почти во всех методах)
- Лучшие методы классификации основаны на сглаживании пороговой функции потерь
- Метод наименьших квадратов для линейной регрессии сводится к решению системы линейных уравнений $n \times n$
- Метод стохастического градиента позволяет единообразно решать самые разные задачи, в том числе для Big Data
- Регуляризация помогает против переобучения
- Можно максимизировать непосредственно AUC

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov

Если что-то было не понятно,
не стесняйтесь подходить и спрашивать :)