1 **Основные понятия и обозначения**

- Данные в задачах обучения по прецедентам
- Модели и методы обучения
- Обучение и переобучение

2 **Примеры прикладных задач**

- Задачи классификации
- Задачи регрессии
- Задачи ранжирования

3 **О методологии машинного обучения**

- Особенности данных
- Межотраслевой стандарт CRISP-DM
- Эксперименты на синтетических и реальных данных

X — множество *объектов* (точнее, их информационных описаний)
Y — множество *ответов* (оценок, предсказаний или прогнозов)
$y: X \to Y$ — неизвестная зависимость (target function)

Дано:
$\{x_1, \ldots, x_\ell\} \subset X$ — *обучающая выборка* (training sample)
$y_i = y(x_i), \;\; i = 1, \ldots, \ell$ — известные ответы

Найти:
$a: X \to Y$ — алгоритм, решающую функцию (decision function), приближающую y на всём множестве X

Весь курс машинного обучения — это конкретизация:

- как задаются объекты и какими могут быть ответы
- в каком смысле «a приближает y»
- как строить функцию a

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

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 $($

f

$$F = \| \qquad \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \\ & \ldots & f_n(x_\ell) \end{pmatrix}$$

1

$($

Задачи классификации (classification):

$Y = \{-1, +1\}$ — *классификация* на 2 класса

$Y = \{1, \ldots, M\}$ — на M непересекающихся классов

$Y = \{0, 1\}^M$ — на M классов,

которые могут пересекаться

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):

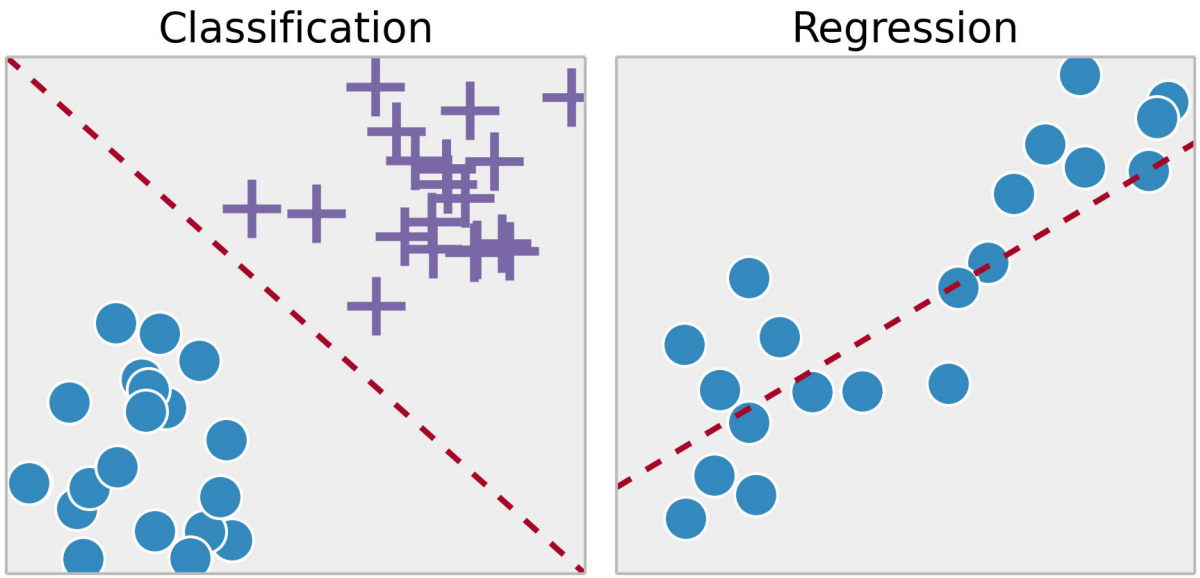
= обучение по прецедентам

= восстановление зависимостей по эмпирическим данным

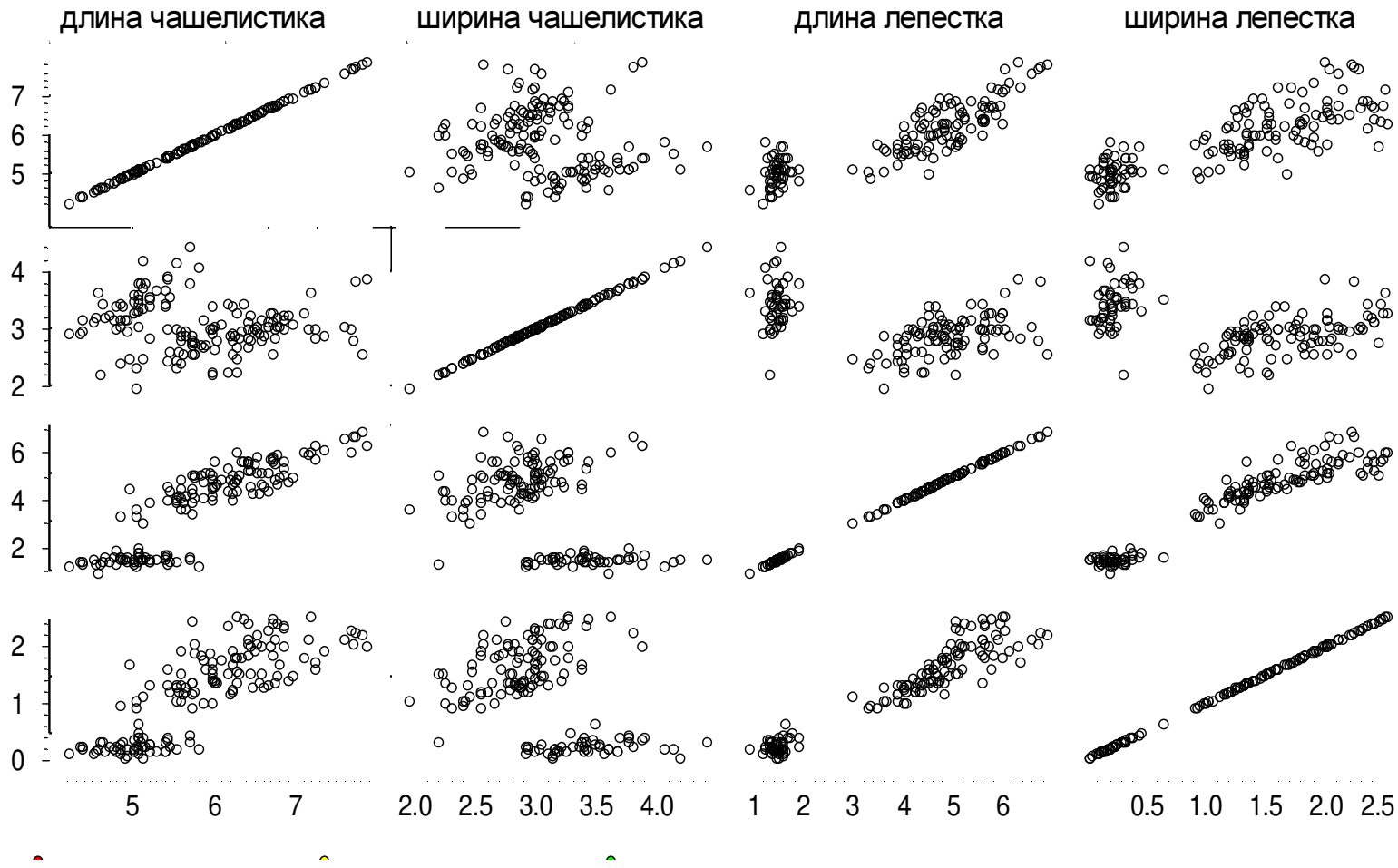
= предсказательное моделирование

= аппроксимация функций по заданным точкам

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



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



Модель (predictive model) — параметрическое семейство функций

$$A = \{g(x, \theta) \mid \theta \in \Theta\},$$

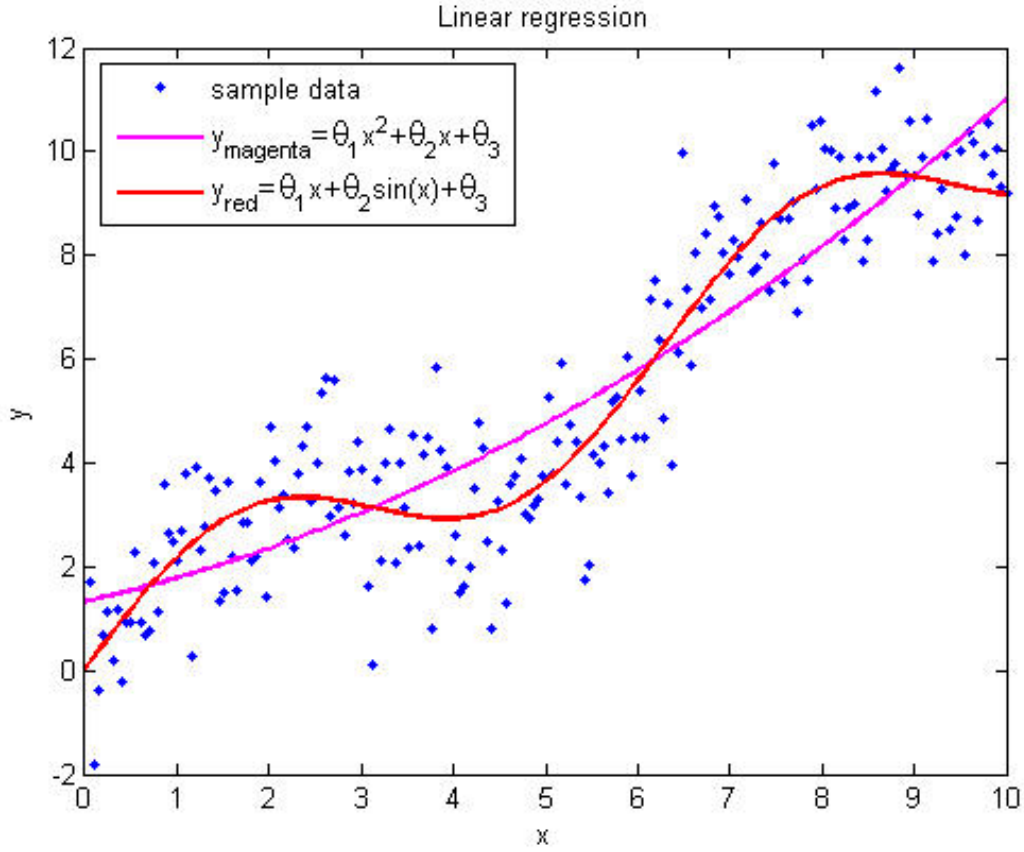
где $g : X \times \Theta \to Y$ — фиксированная функция,
Θ — множество допустимых значений параметра θ

Пример.
Линейная модель с вектором параметров $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$:

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

$g(x, \theta) = \text{sign} \sum_{j=1}^{n} \theta_j f_j(x)$ — для классификации, $Y = \{-1, +1\}$

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

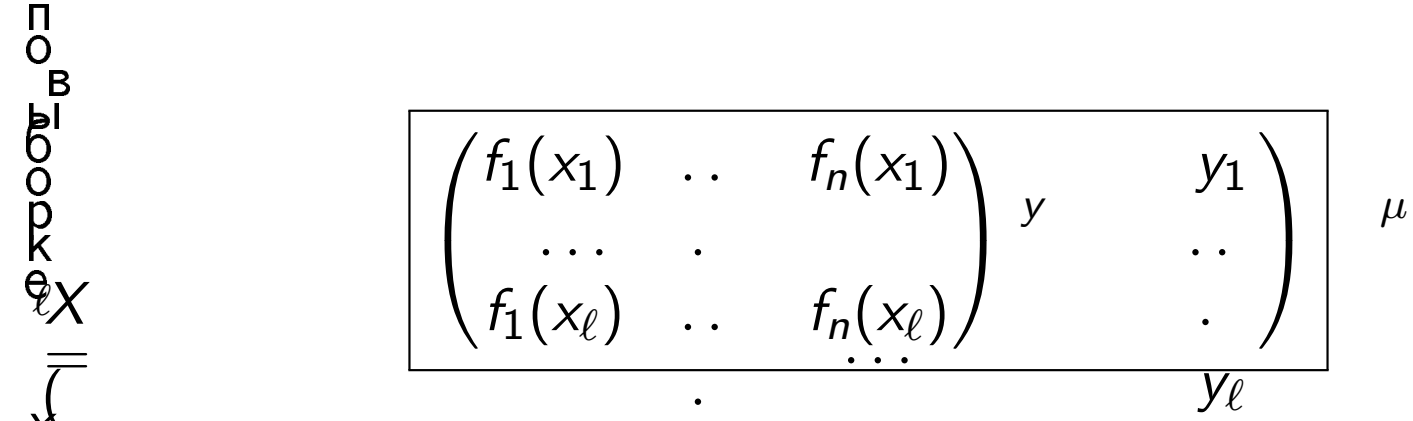


генерация признаков (feature generation) обогащает модель

на практике очень важно «правильно угадать модель»

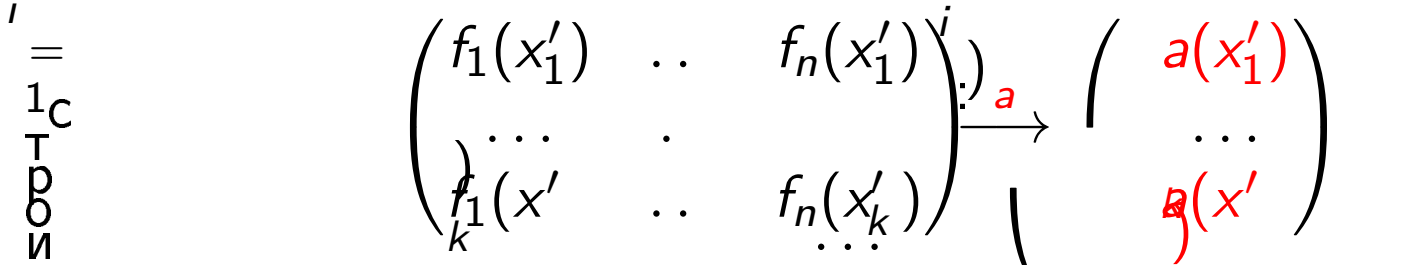
Этап *обучения* (train):

Метод обучения (learning algorithm) $\mu\colon (X \times Y)^\ell \to A$



Этап *применения* (test):

алгоритм a для новых объектов x' выдаёт ответы $a(x')$



$\mathscr{L}(a, x)$ — *функция потерь* (loss function) — величина ошибки алгоритма $a \in A$ на объекте $x \in X$

Функции потерь для задач классификации:

$\mathscr{L}(a, x) = \big[a(x) \neq y(x)\big]$ — индикатор ошибки

Функции потерь для задач регрессии:

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

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

Эмпирический риск — функционал качества алгоритма a на X^ℓ:

$$Q(a, X^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathscr{L}(a, x_i)$$

Метод минимизации эмпирического риска
(Empirical Risk Minimization, ERM):

$$\mu(X^\ell) = \arg\min_{a \in A} Q(a, X^\ell)$$

Пример: задача регрессии, $Y = \mathbb{R}$;
n числовых признаков $f_j(x)$, $j = 1, \ldots, n$;
линейная модель регрессии: $g(x, \theta) = \sum_{j=1}^{n} \theta_j f_j(x)$, $\theta \in \mathbb{R}^n$;
квадратичная функция потерь: $\mathscr{L}(a, x) = ($

a
(
x

$$\mu(X^\ell) = \arg\min_{\theta} \sum_{i=1}^{\ell} ($$

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

$$a(x, \theta) = \theta_0 + \theta_1 x + \cdots + \theta_n x^n \quad \text{— полином степени } n$$

Обучение методом наименьших квадратов:

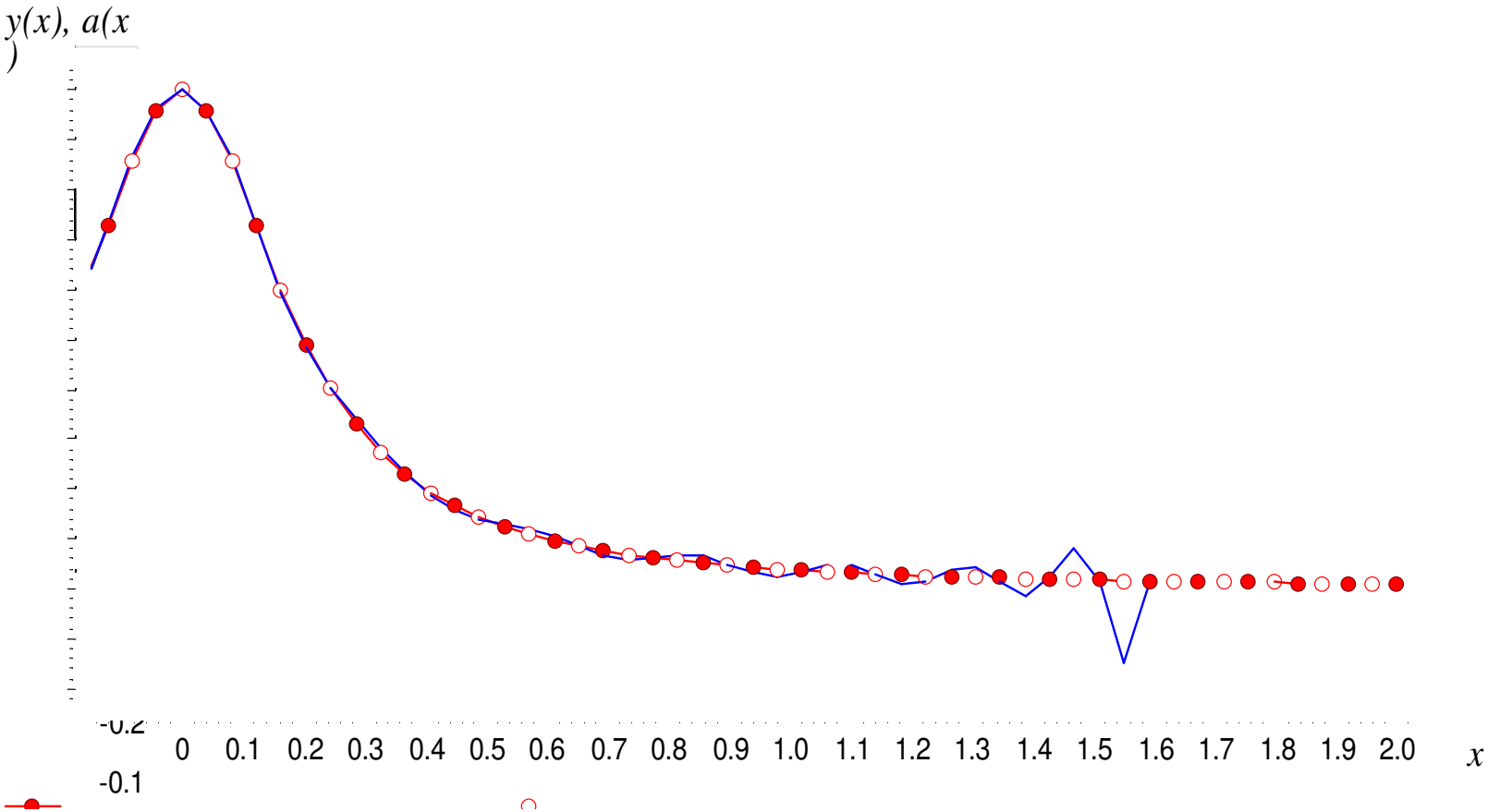
$$Q(\theta, X^\ell) = \sum_{i=1}^{\ell} \qquad ^2 \to \min$$

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

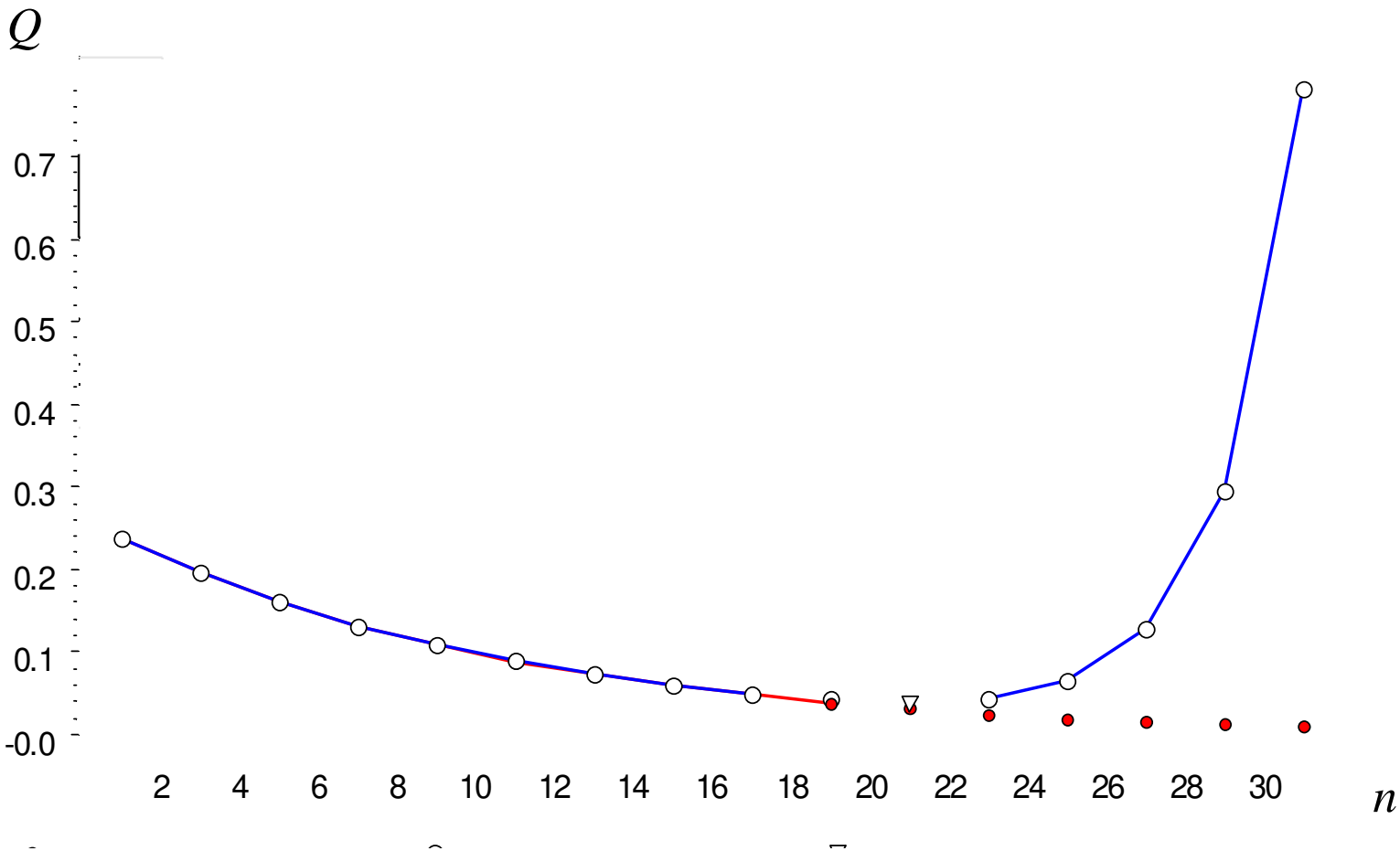
Контрольная выборка: $X^k = \left\{ x_i = 4\frac{i-0.5}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \right.$

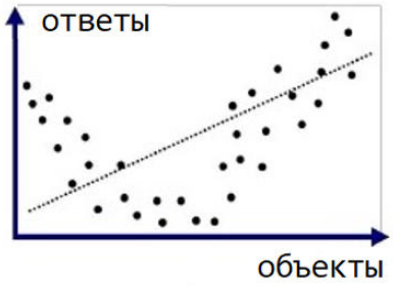
Что происходит с $Q($

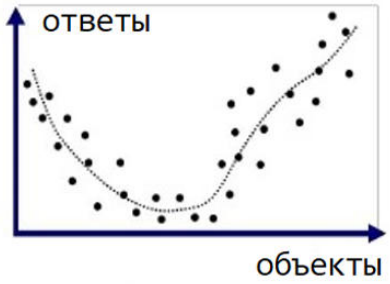
$$y(x) = \frac{1}{1 + 25x^2}; \quad a(x) \text{ — полином степени } n = 38$$

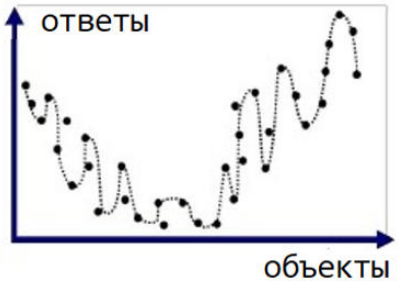


Переобучение — это когда $Q\big(\mu(X^\ell), X^k\big) \gg Q\big(\mu(X^\ell), X^\ell\big)$:







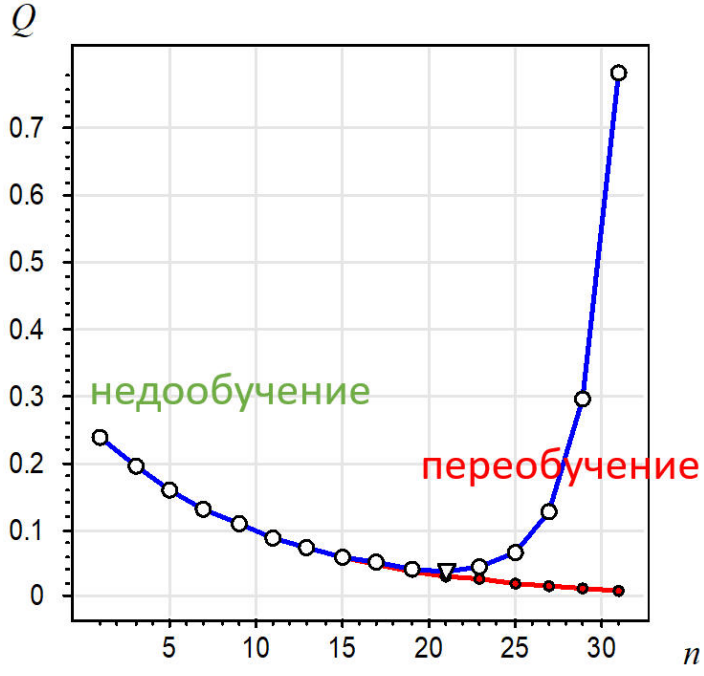


недообучение

переобучение

Недообучение (underfitting): модель слишком проста, недостаточное число параметров n

Переобучение (overfitting): модель слишком сложна, избыточное число параметров n



Из-за чего возникает переобучение?

— избыточные параметры в модели $g(x, \theta)$ «расходуются» на чрезмерно точную подгонку под обучающую выборку

— выбор a из A производится по неполной информации X^ℓ

Как обнаружить переобучение?

— эмпирически, путём разбиения выборки на train и test (на test должны быть известны правильные ответы)

Избавиться от него нельзя. Как его минимизировать?

— накладывать ограничения на θ (регуляризация)

— минимизировать одну из теоретических оценок

— выбирать модель (model selection) по оценкам обобщающей способности (generalization performance)

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 $X^L = X_p^\ell \sqcup X_p^k$

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:

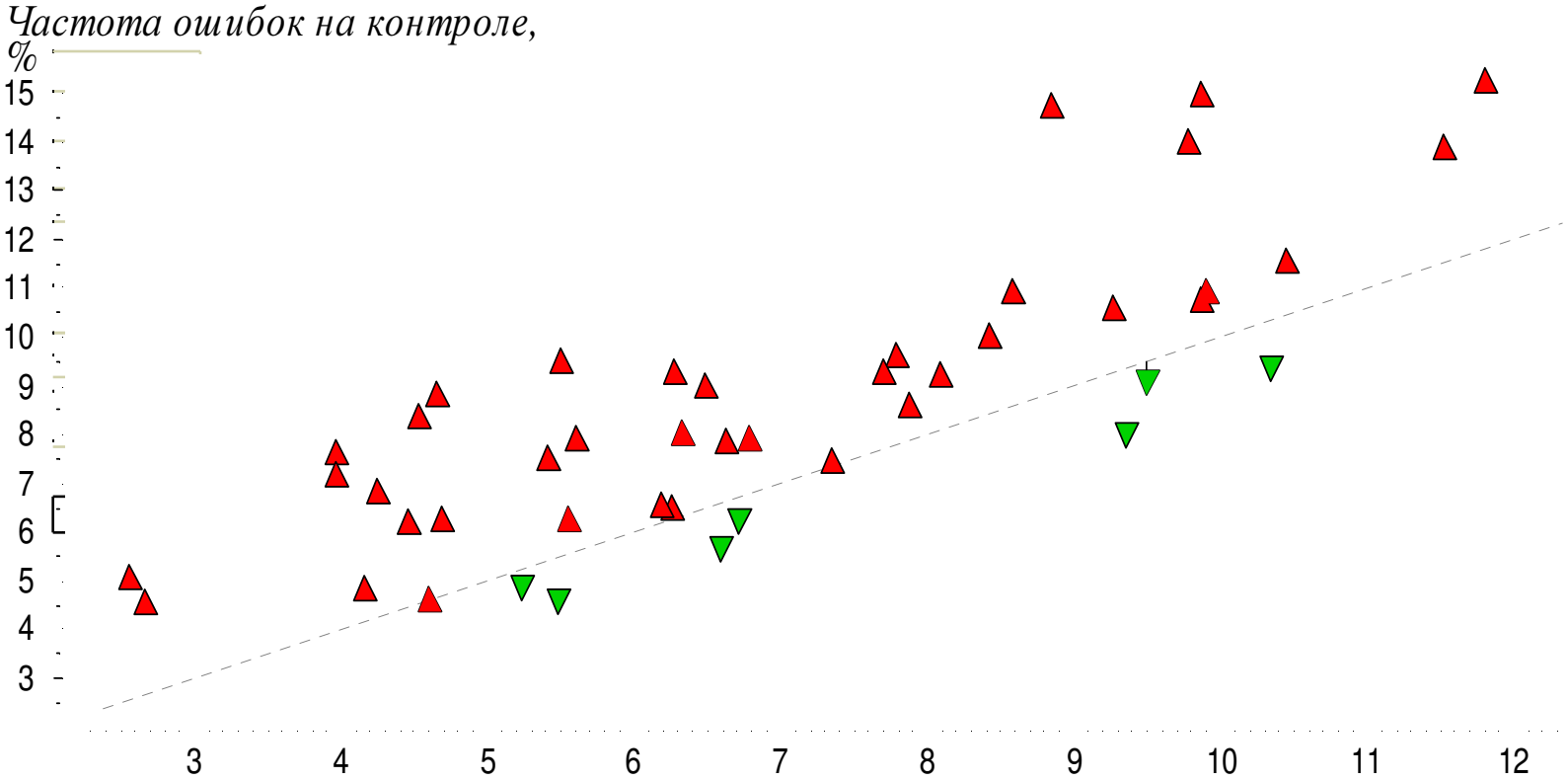
обычно много «пропусков» в данных;

нужен интерпретируемый алгоритм классификации;

нужно выделять *синдромы* — сочетания *симптомов*;

нужна оценка вероятности отрицательного исхода.

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



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.

Объект — заявка на выдачу банком 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 $P($

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, и т. д.), частота оплаты, и т. д.

Особенности задачи:

нужно оценивать вероятность ухода;

сверхбольшие выборки;

признаки приходится вычислять по «сырым» данным.

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.

Идентификация личности по отпечаткам пальцев

Идентификация личности по радужной оболочке глаза

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;

Объект — тройка $\langle$товар, магазин, день$\rangle$.

Примеры признаков:

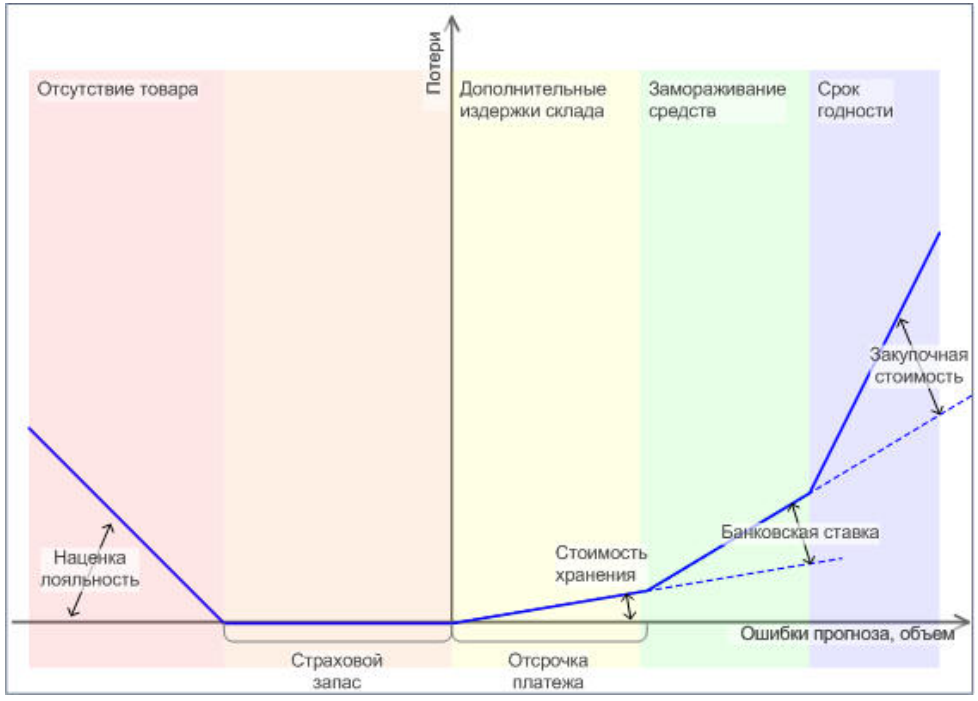
бинарные: выходной день, праздник, промоакция, и т. д.

количественные: объёмы продаж в предшествующие дни.

Особенности задачи:

функция потерь
не квадратична
и даже
не симметрична;

разреженные
данные.



Объект — место для открытия нового ресторана.

Предсказать — прибыль от ресторана через год.

Примеры признаков:

демографические данные: возраст, достаток и т.д.,

цены на недвижимость поблизости,

маркетинговые данные: наличие школ, офисов и т.д.

Особенности задачи:

мало объектов, много признаков;

разнотипные признаки;

есть выбросы;

разнородные объекты (возможно, имеет смысл строить разные модели для мелких и крупных городов).

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.

Объект — тройка $\langle$пользователь, объявление, баннер$\rangle$.

Предсказать — кликнет ли пользователь по контекстной рекламе, которую показали в ответ на его запрос на avito.ru.

Сырые данные:

все действия пользователя на сайте,

профиль пользователя (бразуер, устройство и т. д.),

история показов и кликов других пользователей по баннеру,

. . . всего 10 таблиц данных.

Особенности задачи:

признаки надо придумывать;

данных много — сотни миллионов показов;

основной критерий качества — доход рекламной площадки;

Статистический машинный перевод:

объект — предложение на естественном языке

ответ — его перевод на другой язык

Перевод речи в текст:

объект — аудиозапись речи человека

ответ — текстовая запись речи

Компьютерное зрение:

объект — изображение или видеопоследовательность

ответ — решение (объехать, остановиться, игнорировать)

Предпосылки успешного решения задач со сложными данными:

Большие и *чистые* данные (Big Data)

Глубокие нейросетевые архитектуры (Deep Learning)

Методы оптимизации для задач большой размерности

Рост вычислительных мощностей (закон Мура, GPU)

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)

Риски, связанные с постановкой задачи:

«грязные» данные
(заказчик не обеспечивает качество данных)

неясные критерии качества модели
(заказчик не определился с целями или индикаторами KPI)