

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

Задача обучения по прецедентам

X — множество *объектов* (точнее, их информационных описаний)

Y — множество *ответов* (оценок, предсказаний или прогнозов)

$y: X \rightarrow Y$ — неизвестная зависимость (target function)

Дано:

$\{x_1, \dots, x_\ell\} \subset X$ — *обучающая выборка* (training sample)

$y_i = y(x_i)$, $i = 1, \dots, \ell$ — известные ответы

Найти:

$a: X \rightarrow Y$ — 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

$f_j: X \rightarrow D_j, j = 1, \dots, n$ — 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 $(f_1(x), \dots, f_n(x))$ — признаковое описание объекта x

Матрица «объекты–признаки» (feature data)

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

Задачи классификации (classification):

- $Y = \{-1, +1\}$ — классификация на 2 класса
- $Y = \{1, \dots, M\}$ — на M непересекающихся классов
- $Y = \{0, 1\}^M$ — на M классов, которые могут пересекаться

Задачи восстановления регрессии (regression):

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

Задачи ранжирования (ranking, learning to rank):

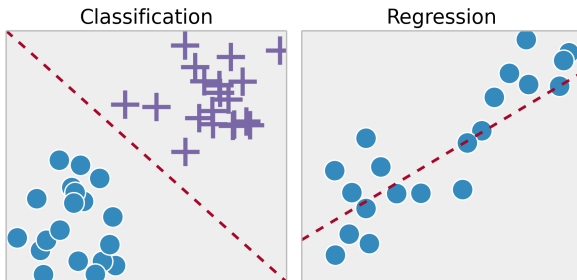
- Y — конечное упорядоченное множество

Задачи обучения без учителя (unsupervised learning):

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

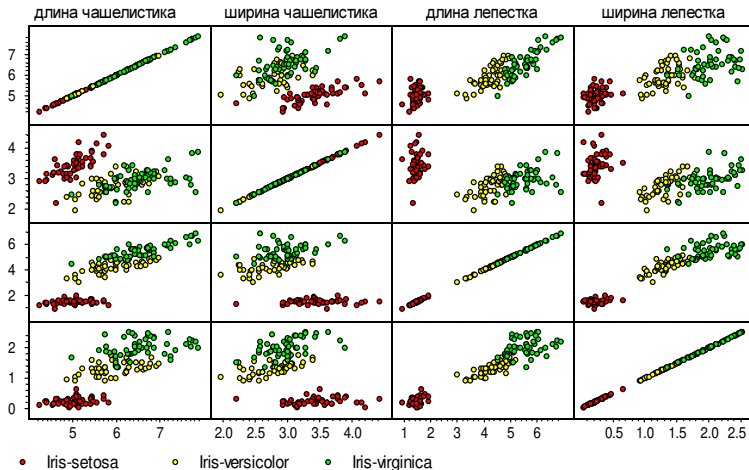
- = обучение по прецедентам
- = восстановление зависимостей по эмпирическим данным
- = предсказательное моделирование
- = аппроксимация функций по заданным точкам

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



Пример: задача классификации цветков ириса [Фишер, 1936]

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



Модель (predictive model) — параметрическое семейство функций

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

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

Пример.

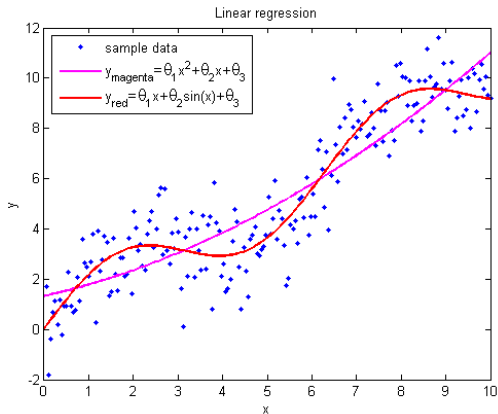
Линейная модель с вектором параметров $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$:

$g(x, \theta) = \sum_{j=1}^n \theta_j f_j(x)$ — для регрессии и ранжирования, $Y = \mathbb{R}$

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

Пример: задача регрессии, синтетические данные

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



- генерация признаков (feature generation) обогащает модель
- на практике очень важно «правильно угадать модель»

Этап обучения (train):

Метод обучения (learning algorithm) $\mu: (X \times Y)^\ell \rightarrow A$
по выборке $X^\ell = (x_i, y_i)_{i=1}^\ell$ строит алгоритм $a = \mu(X^\ell)$:

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

Этап применения (test):

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

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

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

Функции потерь для задач классификации:

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

Функции потерь для задач регрессии:

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

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

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

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 $f_j(x)$, $j = 1, \dots, n$;

линейная модель регрессии: $g(x, \theta) = \sum_{j=1}^n \theta_j f_j(x)$, $\theta \in \mathbb{R}^n$;

квадратичная функция потерь: $\mathcal{L}(a, x) = (a(x) - y(x))^2$.

Метод наименьших квадратов — частный случай ERM:

$$\mu(X^\ell) = \arg \min_{\theta} \sum_{i=1}^{\ell} (g(x_i, \theta) - y_i)^2$$

Пример Рунге. Аппроксимация функции полиномом

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

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

Модель полиномиальной регрессии

$a(x, \theta) = \theta_0 + \theta_1 x + \dots + \theta_n x^n$ — полином степени n

Обучение методом наименьших квадратов:

$$Q(\theta, X^\ell) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \dots + \theta_n x_i^n - y_i)^2 \rightarrow \min_{\theta_0, \dots, \theta_n}$$

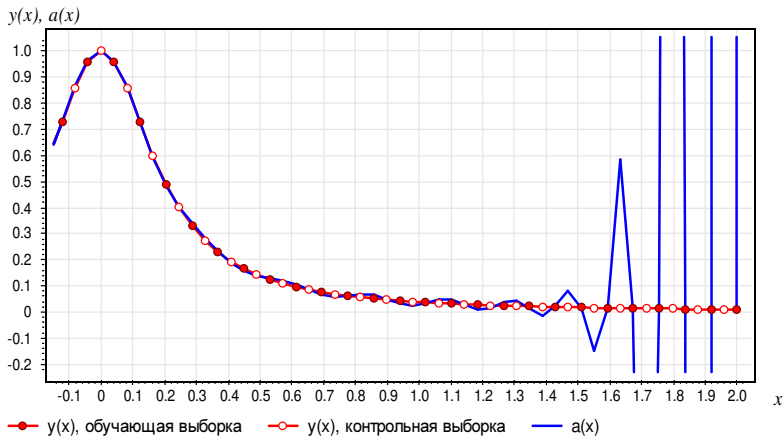
Обучающая выборка: $X^\ell = \{x_i = 4 \frac{i-1}{\ell-1} - 2 \mid i = 1, \dots, \ell\}$

Контрольная выборка: $X^k = \{x_i = 4 \frac{i-0.5}{\ell-1} - 2 \mid i = 1, \dots, \ell-1\}$

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

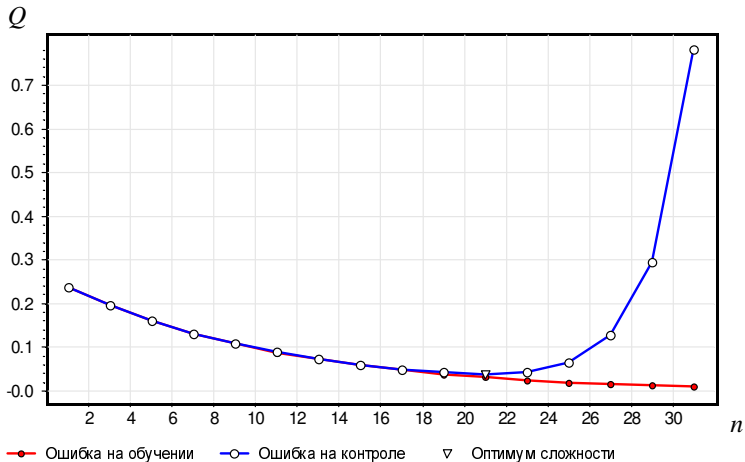
Пример Рунге. Переобучение при $n = 38$, $\ell = 50$

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

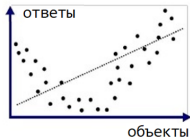


Пример Рунге. Зависимость Q от степени полинома n

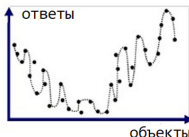
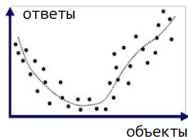
Переобучение — это когда $Q(\mu(X^\ell), X^k) \gg Q(\mu(X^\ell), X^\ell)$:



Проблемы недообучения и переобучения

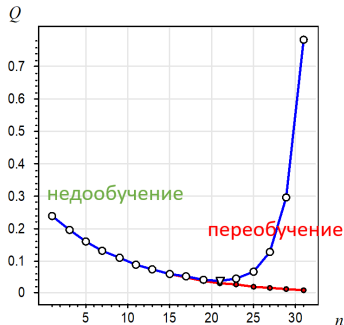


недообучение



переобучение

- **Недообучение** (underfitting):
модель слишком проста,
недостаточное число
параметров n
- **Переобучение** (overfitting):
модель слишком сложна,
избыточное число
параметров n



❶ Из-за чего возникает переобучение?

- избыточные параметры в модели $g(x, \theta)$ «расходятся» на чрезмерно точную подгонку под обучающую выборку
- выбор a из A производится по неполной информации X^ℓ

❷ Как обнаружить переобучение?

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

❸ 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):

$$\text{HO}(\mu, X^\ell, X^k) = Q(\mu(\textcolor{red}{X}^\ell), \textcolor{blue}{X}^k) \rightarrow \min$$

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

$$\text{LOO}(\mu, X^L) = \frac{1}{L} \sum_{i=1}^L \mathcal{L}(\mu(\textcolor{red}{X}^L \setminus \{x_i\}), \textcolor{blue}{x}_i) \rightarrow \min$$

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

$$\text{CV}(\mu, X^L) = \frac{1}{|P|} \sum_{p \in P} Q(\mu(\textcolor{red}{X}_p^\ell), \textcolor{blue}{X}_p^k) \rightarrow \min$$

где P — множество разбиений $X^L = \textcolor{red}{X}_p^\ell \sqcup \textcolor{blue}{X}_p^k$

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.

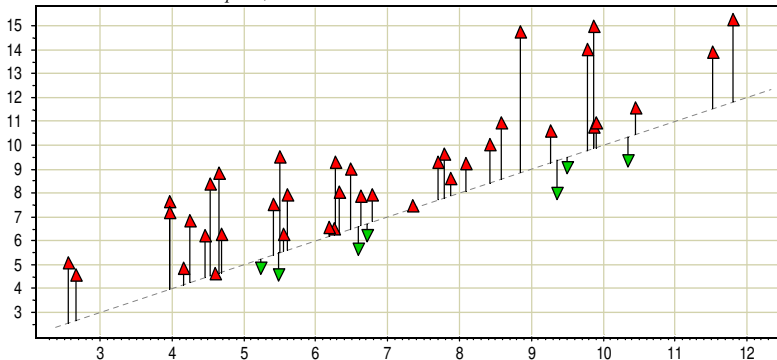
Особенности задачи:

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

Задача медицинской диагностики. Пример переобучения

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

Частота ошибок на контроле, %



Частота ошибок на обучении, %

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.

Задача кредитного скоринга

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 $P(y(x) = \text{bad})$.

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.

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.

Идентификация личности по отпечаткам пальцев



Идентификация личности по радужной оболочке глаза



Особенности задач:

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

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

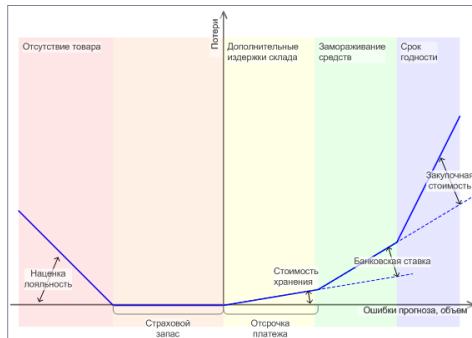
Объект — тройка $\langle \text{товар, магазин, день} \rangle$.

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).

Задача ранжирования поисковой выдачи

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

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

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 ⟨пользователь, объявление, баннер⟩.

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 (браузер, 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)

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 признаков 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)