Машинное обучение

ФИО преподавателя: Оцоков Шамиль Алиевич

e-mail: shamil24@mail.ru

Список литературы

1. Бурков Андрей. Машинное обучение без лишних слов.

2. Андреас Мюллер. Введение в машинное обучение с помощью Python.

3. Грас Data Science. Наука о данных с нуля, 2017 г.

4. Введение в машинное обучение.
https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie/lecture/CLOS0/formal-naia-postanovka-zadachi-mashinnogho-obuchieniia

Обучение без учителя

Коллекция неразмеченных образцов $\{x_i\}$ i=1..N.

Цель создать модель, которая преобразует исходный вектор в другой или в значение.

Примеры:

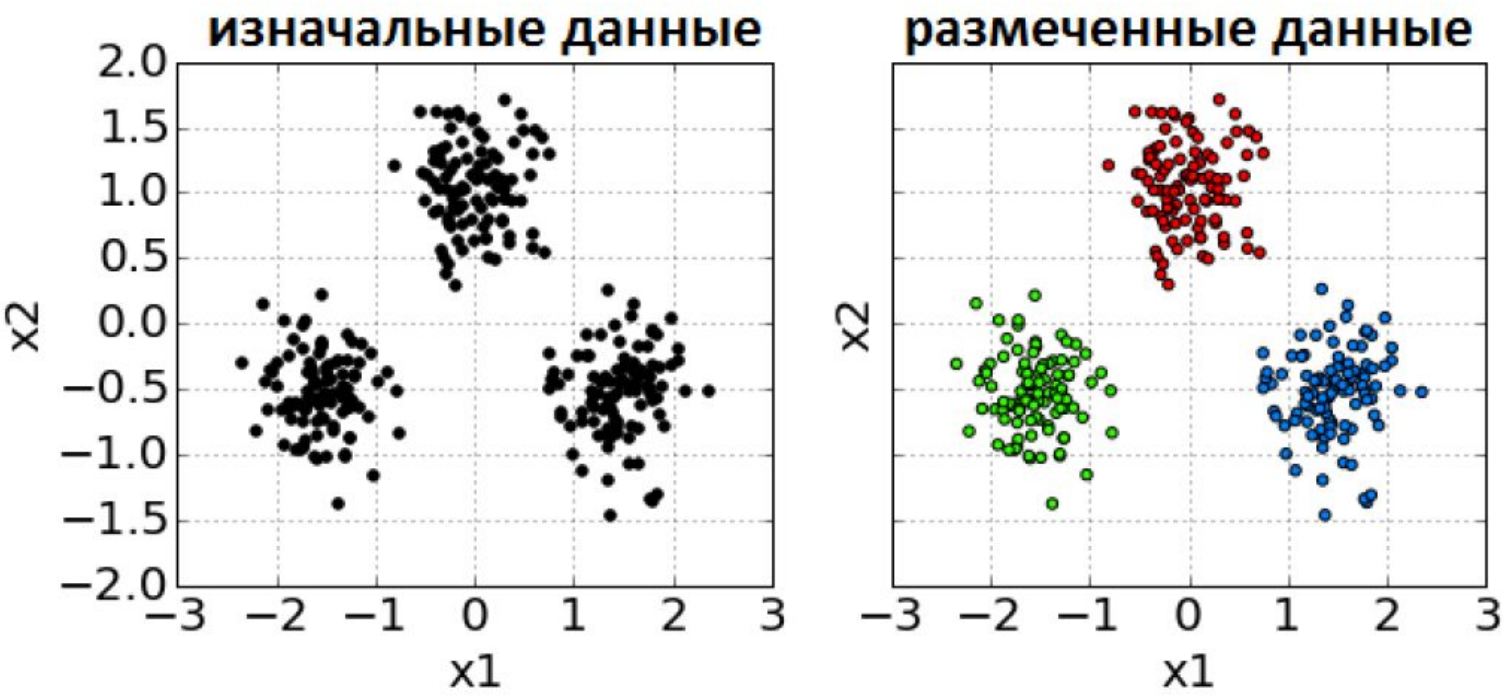
- задача кластеризации,

- уменьшение размерности

- выявление аномалий (возвращается действительное число, которое указывает, насколько
x отличается от «типичного» образца в наборе данных)

Обучение с частичным привлечением учителя

Пример кластеризации



Обучение без учителя: кластеризация

Обучение с подкреплением

Обучение с подкреплением — это раздел машинного обучения, где предполагается, что машина «живет» в определенном окружении и способна воспринимать состояние этого окружения как вектор характеристик. Машина может выполнять некоторые действия в каждом состоянии. Разные действия приносят разные вознаграждения, а также могут перевести машину в другое состояние окружения.

Цель алгоритма обучения с подкреплением — выучить линию поведения

Линия поведения — это стратегия (похожая на модель в обучении с учителем), которая принимает вектор признаков, описывающий состояние, и возвращает оптимальное действие для выполнения в этом состоянии. Действие является оптимальным, если приводит к максимальному ожидаемому среднему вознаграждению.

Обучение с подкреплением решает особый класс задач, когда решения принимаются последовательно, а цель является долгосрочной

Уменьшение размерности. Поиск и использование сходств в структуре данных аналогично алгоритмам кластеризации, но с использованием методов обучения без учителя. Цель заключается в суммировании или описании данных с использованием меньшего количества информации, чтобы набор данных становился меньше и проще в управлении. В некоторых случаях эти алгоритмы используют для задач классификации или регрессии. Вот список наиболее распространенных алгоритмов уменьшения размерности.

- Анализ основных компонентов (Principal Component Analysis — PCA).
- Факторный анализ (Factor Analysis — FA).
- Многомерное шкалирование (Multidimensional Scaling — MDS).
- Стохастическое вложение соседей с t-распределением (t-Distributed Stochastic Neighbor Embedding — t-SNE).

Ансамбли. Объединяет группу из нескольких слабых моделей в единое целое, чьи индивидуальные прогнозы каким-то образом объединяются для формирования общего прогноза. Использование ансамбля может решать некоторые задачи быстрее, эффективнее или с меньшим количеством ошибок. Вот несколько популярных алгоритмов ансамбля.

- Бустинг.
- Бутстрэп-агрегирование (бэггинг).
- AdaBoost.
- Случайный лес.
- Метод градиентного бустинга (Gradient Boosting Machine — GBM).

Задача обучения по прецедентам

X — множество *объектов*;
Y — множество *ответов*;
$y: X \to Y$ — неизвестная зависимость (target function).

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

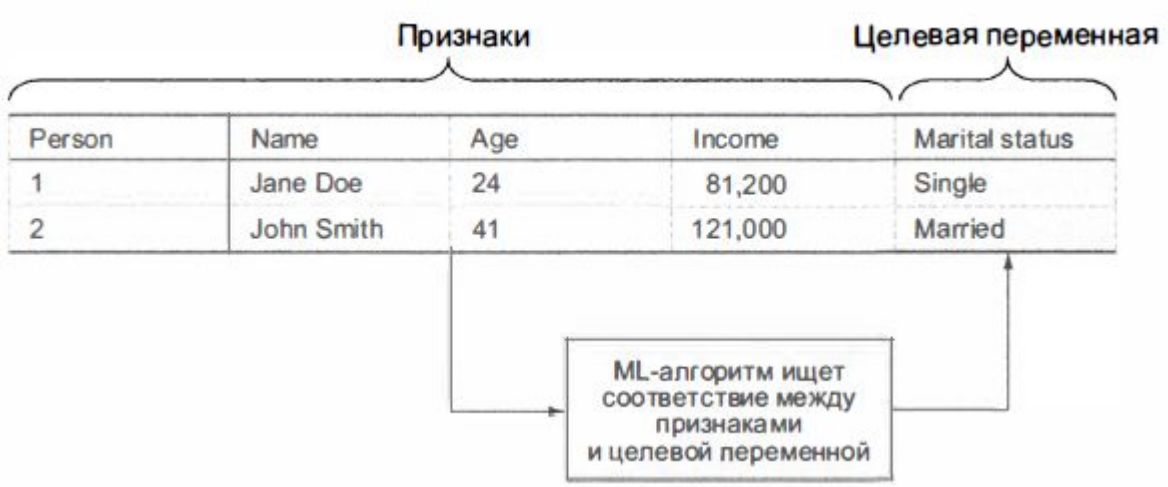
Найти:
$a: X \to Y$ — алгоритм, решающую функцию (decision function), приближающую y на всём множестве X.

Весь курс машинного обучения — это конкретизация:

- как задаются объекты и какими могут быть ответы;
- как строить функцию a;
- в каком смысле a должен приближать y.

- **Основные понятия машинного обучения:**
  объект, ответ, признак, предсказательная модель, метод обучения, эмпирический риск, переобучение.
- **Прикладные задачи машинного обучения**
  встречаются во всех областях бизнеса, науки, производства — об этом в следующей лекции

Процесс моделирования с машинным обучением



В статистике и теории вычислительных систем целевая переменная называется откликом (response), или зависимой переменной (dependent variable).

Процесс моделирования с машинным обучением

Как задаются объекты. Признаковое описание

$f_j\colon X \to D_j,\ \ j = 1, \ldots, n$ — 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 $\big(f_1(x), \ldots, f_n(x)\big)$ — *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 ответы. Типы задач

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.

Предсказательная модель

Модель (predictive model) — параметрическое семейство функций

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

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

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

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

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

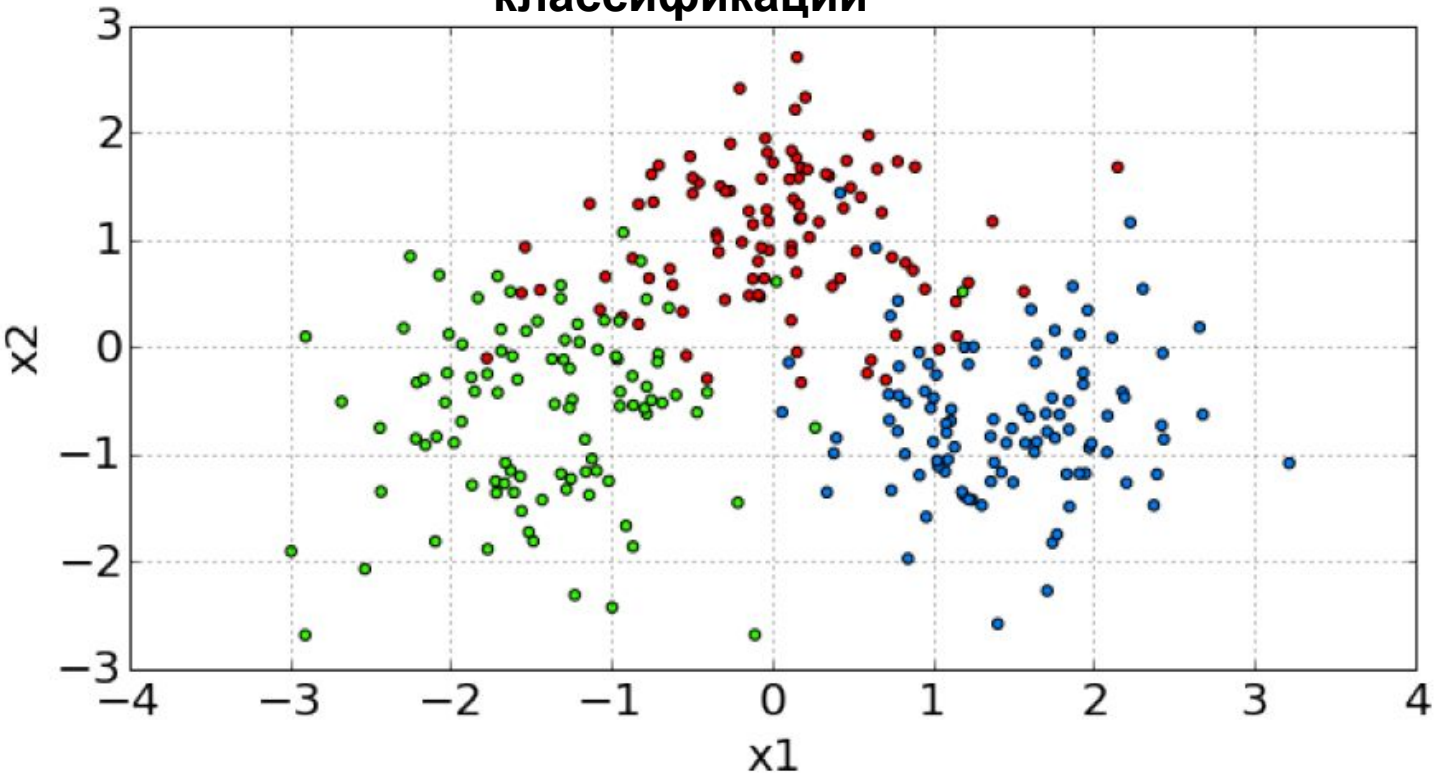
Выбор данных для алгоритмов машинного обучения

1. Включите все признаки, которые кажутся хоть как-то связанными с целевой переменной. Выполните обучение ML-модели. Если точность прогноза вас устраивает, останавливайтесь.

2. В противном случае расширьте набор, добавив туда признаки, связь которых с целевой переменной менее очевидна. Снова выполните обучение модели и оцените ее точность. Если она вас устраивает, останавливайтесь.

3. Если точность все еще неудовлетворительна, запустите для расширенного набора признаков *алгоритм отбора* (feature selection algorithm), чтобы выбрать оптимальное, сильнее всего влияющее на процесс прогнозирования подмножество.

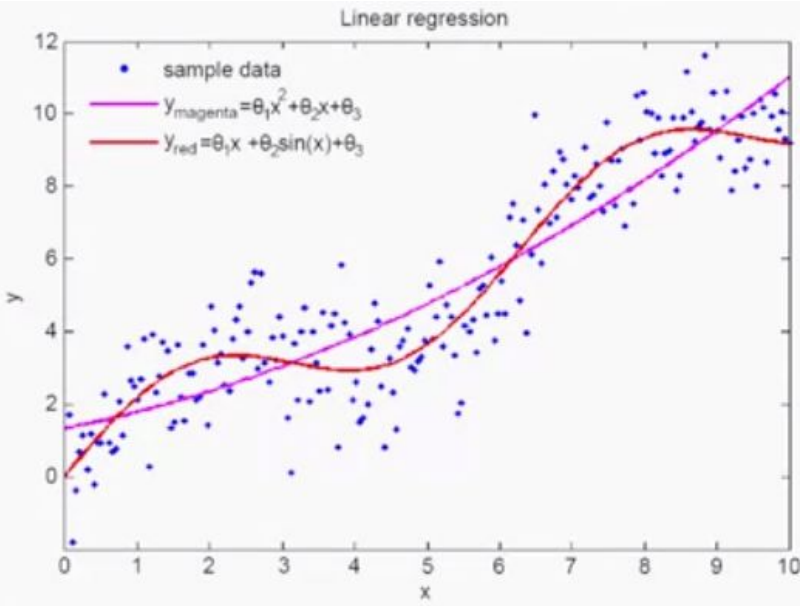
Пример классификации



Обучение с учителем: $x = (x_1, x_2)$, y обозначен цветом

Пример: задача регрессии, модельные данные

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



Вывод: признаковое описание можно задавать по-разному

Этапы обучения и применения модели

Этап обучения (train):

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

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

Этап применения (test):

алгоритм a для новых объектов $x'_1, \dots, x'_k$ выдаёт ответы $a(x'_i)$.

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

Функционалы качества

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

Функции потерь для задач классификации:

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

Функции потерь для задач регрессии:

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

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

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

Сведение задачи обучения к задаче оптимизации

Минимизация эмпирического риска (empirical risk minimization):

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

Пример: *метод наименьших квадратов* ($Y = \mathbb{R}$, $\mathscr{L}$ квадратична):

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

Метод наименьших квадратов (МНК).

Пример.

Экспериментальные данные о значениях переменных *x* и *y* приведены в таблице.

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
x_i	0	1	2	4	5
y_i	2,1	2,4	2,6	2,8	3,0

В результате их выравнивания получена функция $g(x)=\sqrt[3]{x+1}+1$

Используя **метод наименьших квадратов** , аппроксимировать эти данные линейной зависимостью $y=ax+b$ (найти параметры *a* и *b*). Выяснить, какая из двух линий лучше (в смысле метода наименьших квадратов) выравнивает экспериментальные данные. Сделать чертеж.

Суть метода наименьших квадратов (МНК).

Задача заключается в нахождении коэффициентов линейной зависимости, при которых функция двух переменных *a* и *b* $F(a,b)=\sum_{i=1}^{n}(y_i-(ax_i+b))^2$ принимает наименьшее значение. То есть, при данных *a* и *b* сумма квадратов отклонений экспериментальных данных от найденной прямой будет наименьшей. В этом вся суть метода наименьших квадратов.

Таким образом, решение примера сводится к нахождению экстремума функции двух переменных.

Вывод формул для нахождения коэффициентов.

Составляется и решается система из двух уравнений с двумя неизвестными. Находим частные производные функции $F(a,b)=\sum_{i=1}^{n}(y_i-(ax_i+b))^2$ по переменным *a* и *b*, приравниваем эти производные к нулю.

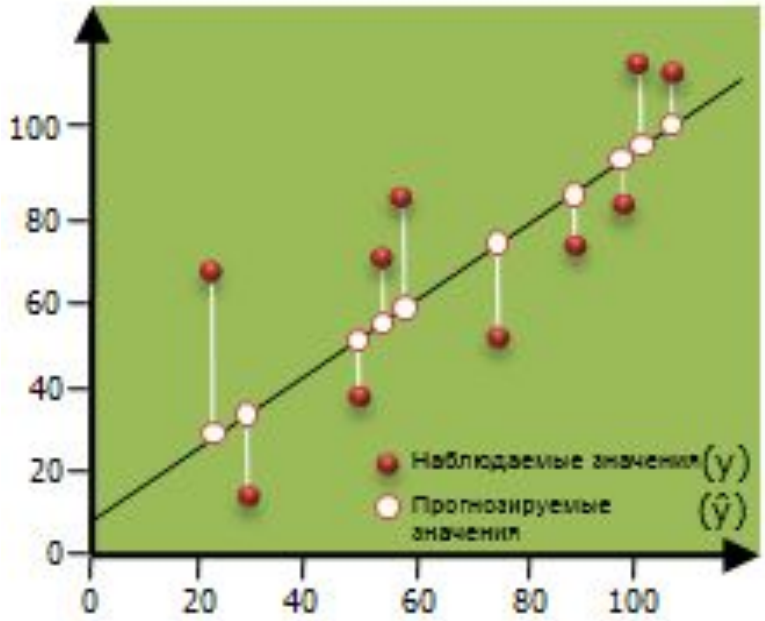
$$\begin{cases}\frac{\partial F(a,b)}{\partial a}=0\\ \frac{\partial F(a,b)}{\partial b}=0\end{cases}\Leftrightarrow\begin{cases}-2\sum_{i=1}^{n}(y_i-(ax_i+b))x_i=0\\ -2\sum_{i=1}^{n}(y_i-(ax_i+b))=0\end{cases}\Leftrightarrow$$

$$\begin{cases}a\sum_{i=1}^{n}x_i^2+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}x_iy_i\\ a\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}b=\sum_{i=1}^{n}y_i\end{cases}\Leftrightarrow\begin{cases}a\sum_{i=1}^{n}x_i^2+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}x_iy_i\\ a\sum_{i=1}^{n}x_i+nb=\sum_{i=1}^{n}y_i\end{cases}$$

Решаем полученную систему уравнений любым методом (например *методом подстановки* или методом Крамера) и получаем формулы для нахождения коэффициентов по методу наименьших квадратов (МНК).

$$\begin{cases} a = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ b = \dfrac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n} \end{cases}$$

При данных *a* и *b* функция $F(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$ принимает наименьшее значение.



Понятие обобщающей способности (generalization performance):

- найдём ли мы «закон природы» или *переобучимся*, то есть подгоним функцию $g(x_i, \theta)$ под заданные точки?
- будет ли $a = \mu(X^\ell)$ приближать функцию y на всём X?
- будет ли $Q(a, X^k)$ мало́ на новых данных — *контрольной выборке* $X^k = (x'_i, y'_i)_{i=1}^k$, $y'_i = y(x_i)$?

Восстановление зависимостей по эмпирическим данным

Задача восстановления зависимости $y = y(x)$
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,
$y_i = y(x_i)$ — правильные ответы, $i = 1, \ldots, \ell$:

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

Найти: функцию $a(x)$, способную давать правильные ответы
на *тестовых объектах* $\tilde{x}_i = (\tilde{x}_i^1, \ldots, \tilde{x}_i^n)$, $i = 1, \ldots, k$:

$$\begin{pmatrix} \tilde{x}_1^1 & \ldots & \tilde{x}_1^n \\ \ldots & \ldots & \ldots \\ \tilde{x}_k^1 & \ldots & \tilde{x}_k^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \ldots \\ a(\tilde{x}_k) \end{pmatrix}$$

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;
- как правило, недостаточный объём данных;
- нужен интерпретируемый алгоритм классификации;
- нужна оценка вероятности

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 $P(\text{bad})$.

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, и т. д.), частота оплаты, и т. д.

Особенности задачи:

- нужно оценивать вероятность ухода;
- сверхбольшие выборки;
- не ясно, какие признаки 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.

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 объемов продаж

Объект — тройка ⟨товар, магазин, день⟩.

Примеры признаков:

- **бинарные:** выходной день, праздник, промоакция, и т. д.
- **количественные:** объёмы продаж в предшествующие дни.

Задача прогнозирования объемов продаж

Особенности задачи:

- функция потерь не квадратична и даже не симметрична;
- разреженные данные.

Конкурс kaggle.com: TFI Restaurant Revenue Prediction

Объект — место для открытия нового ресторана.

Предсказать — прибыль от ресторана через год.

Примеры признаков:

- демографическими свойствами района;
- цены на недвижимость поблизости;
- маркетинговые данные: наличие школ, офисов и т.д.

Степени обученности модели

Недообученная модель

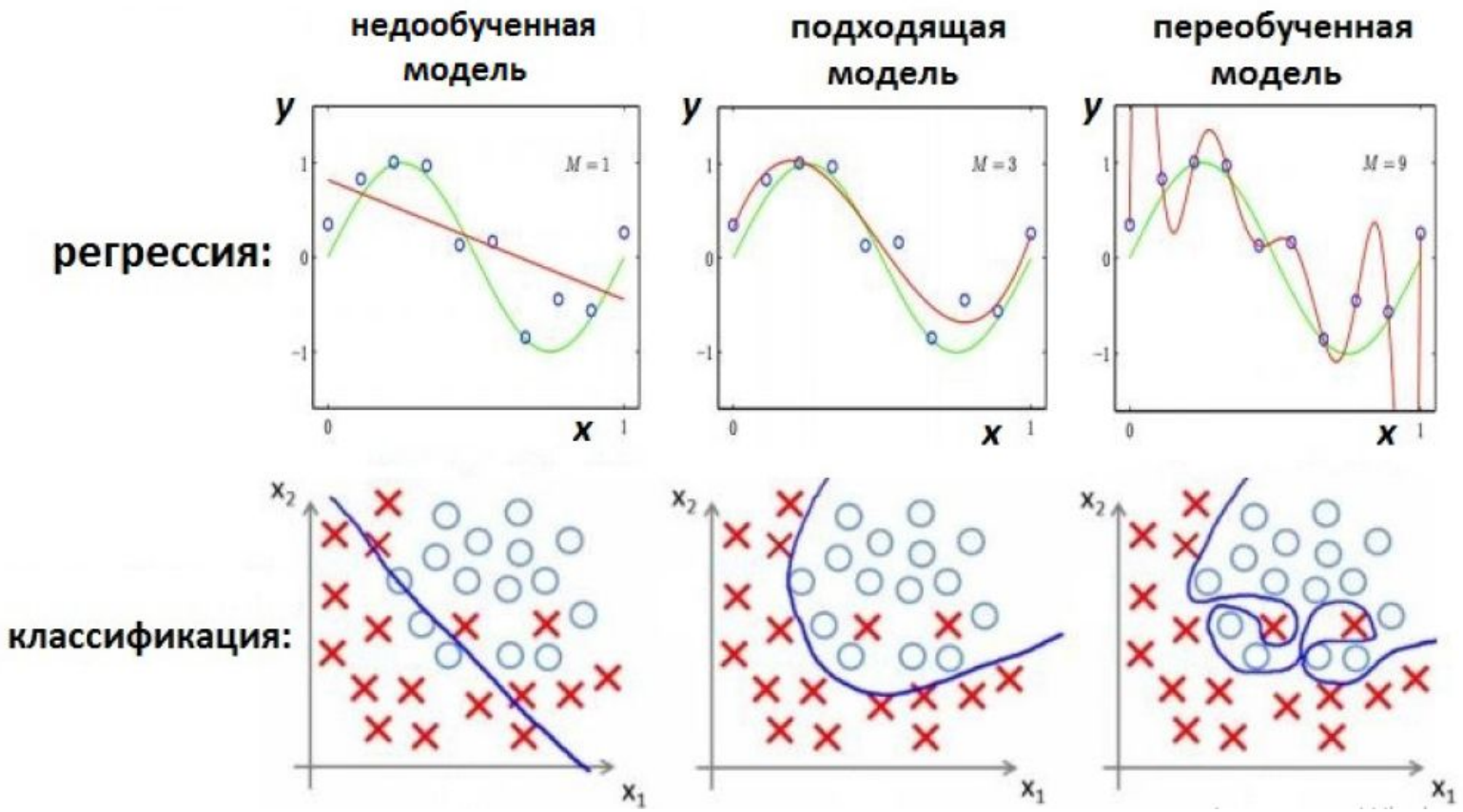
Модель, слишком сильно упрощающая закономерность $\mathcal{X} \to \mathcal{Y}$.

Переобученная модель

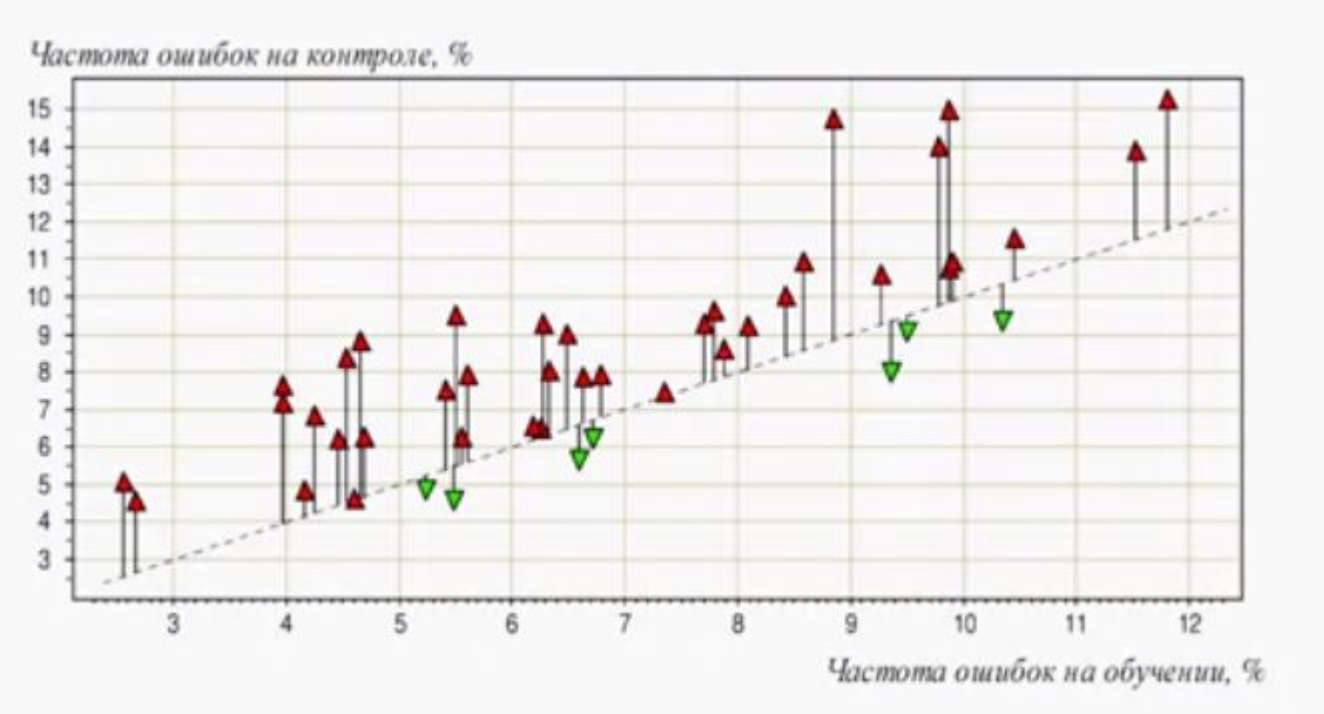
Модель, слишком сильно настроенная на особенности обучающей выборки (на шум в наблюдениях), а не на реальную закономерность $\mathcal{X} \to \mathcal{Y}$.

Примеры недообученных и переобученных моделей

истинная закономерность
оцененная закономерность полиномом степени M
объекты обучающей выборки



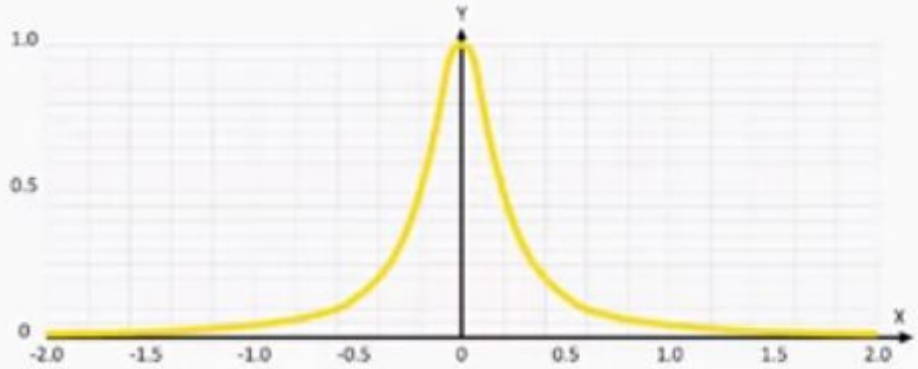
Пример. Переобучение в задаче медицинской диагностики



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

Пример. Переобучение полиномиальной регрессии

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



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

Модель полиномиальной регрессии

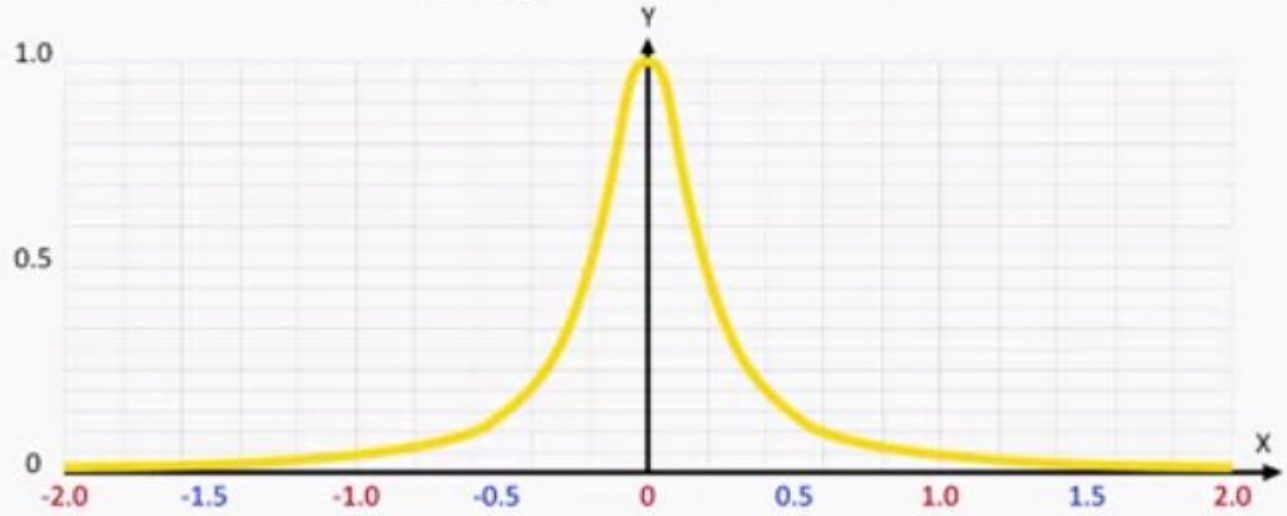
$$a(x, \theta) = \theta_0 + \theta_1 x + \cdots + \theta_n x^n \quad \text{— полином степени } n.$$

Обучение методом наименьших квадратов:

$$Q(a, X^\ell) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \cdots + \theta_n x_i^n - y_i)^2 \to \min_{\theta_0, \ldots, \theta_n}.$$

Пример. Переобучение полиномиальной регрессии

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



Обучающая выборка:

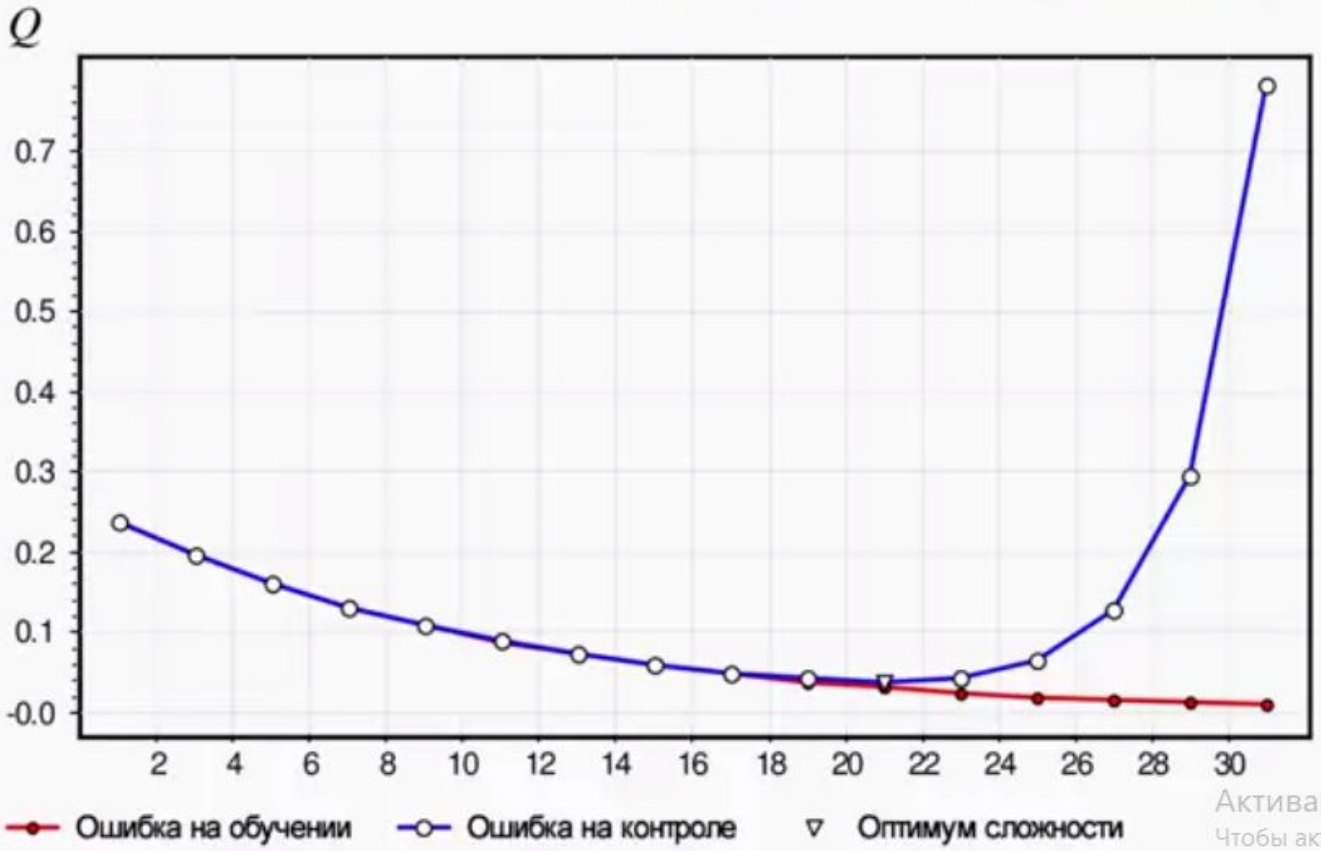
$$X^\ell = \left\{ x_i = 4\tfrac{i-1}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell \right\}.$$

Контрольная выборка:

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

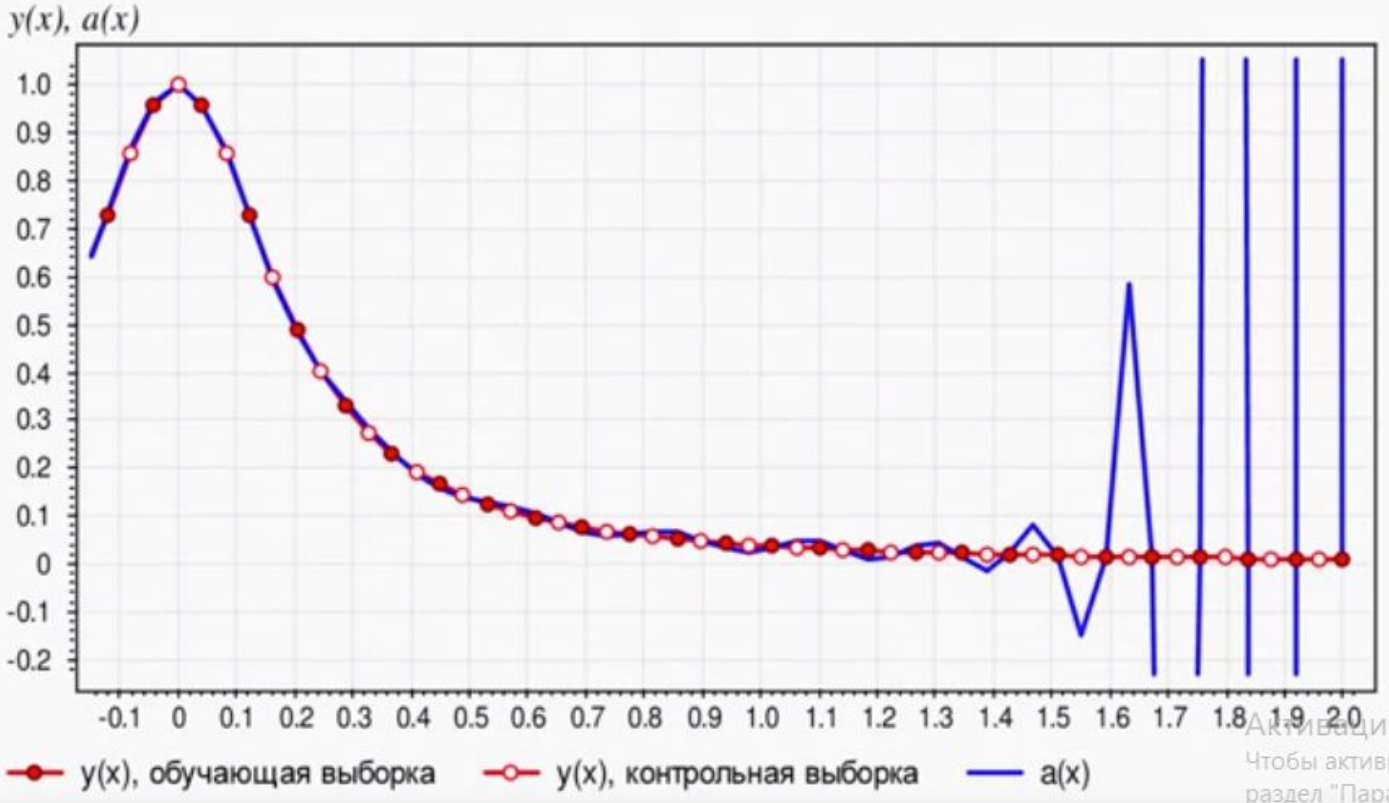
Пример переобучения: эксперимент при l = 50, n = 1,...,31

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



Пример переобучения: эксперимент при l = 50, n = 38

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



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) по N разбиениям, $X^L = X_n^\ell \sqcup X_n^k$, $L = \ell + k$:

$$\mathrm{CV}(\mu, X^L) = \frac{1}{N}\sum_{n=1}^{N} Q(\mu(X_n^\ell), X_n^k) \to \min$$

Эксперименты на реальных данных

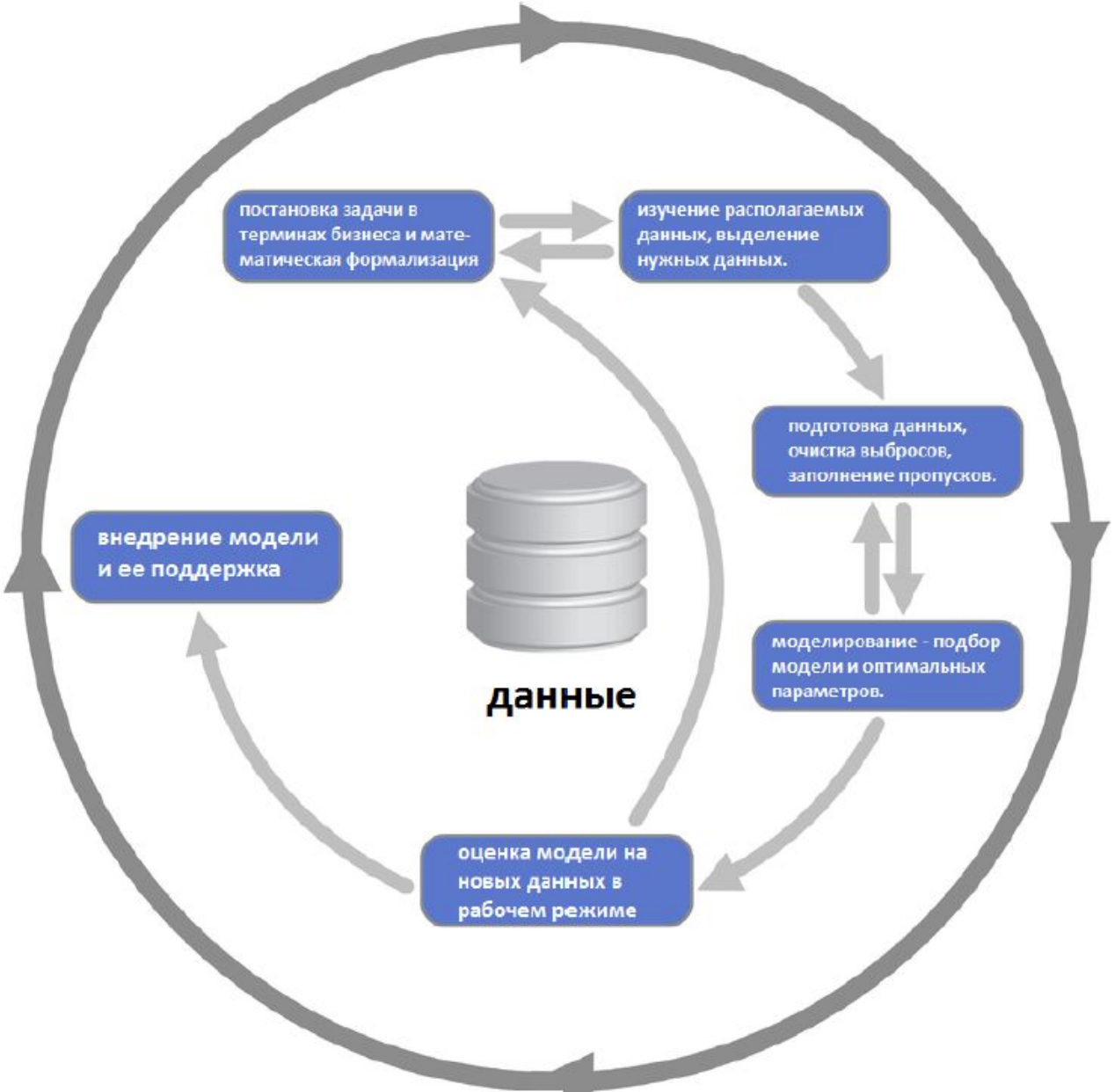
Эксперименты на конкретной прикладной задаче:

- цель — решить задачу как можно лучше
- важно понимание задачи и данных
- важно придумывать информативные признаки
- конкурсы по анализу данных: http://www.kaggle.com

Эксперименты на наборах прикладных задач:

- цель — протестировать метод в разнообразных условиях
- нет необходимости (и времени) разбираться в сути задач :(
- признаки, как правило, уже кем-то придуманы
- репозиторий UC Irvine Machine Learning Repository
  http://archive.ics.uci.edu/ml (308 задач, 09-02-2015)

Методология CrispDM



- **Прикладные задачи машинного обучения**
  встречаются во всех областях бизнеса, науки, производства

- **Особенности данных в прикладных задачах:**
  — 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);
  — нетривиальные критерии качества.

Этапы решения задач машинного обучения:

- понимание задачи и данных;
- предобработка данных и изобретение признаков;
- построение модели;
- сведение обучения к оптимизации;
- решение проблем оптимизации и переобучения;
- оценивание качества решения;
- внедрение и эксплуатация.

Спасибо за внимание!