Машинное обучение: вводная лекция

Воронцов Константин Вячеславович
vokov@forecsys.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov

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

МФТИ • 5 сентября 2020

Эпиграф №1 к курсу машинного обучения

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

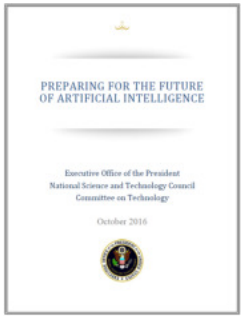
Клаус Мартин Шваб,
президент Всемирного экономического форума

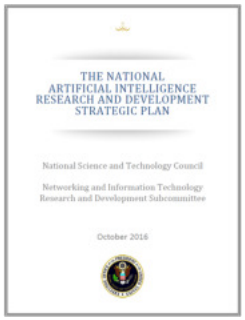
Мир наконец поверил в искусственный интеллект.
Машинное обучение — новый двигатель прогресса.
Машинное обучение — это технологии, которые меняют мир.

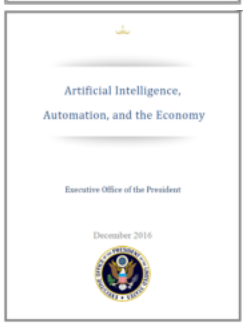
Эпиграф №2. Отчёты Белого Дома США, октябрь 2016

«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
- 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)
- Мониторинг сельского хозяйства
- Автономные системы вооружений







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

Разные названия нашей науки

- Статистический анализ данных (Statistical Data Analysis)
- Искусственный интеллект (Artificial Intelligence) — 1955
- Распознавание образов (Pattern Recognition)
- Машинное обучение (Machine Learning) — 1959
- Статистическое обучение (Statistical Learning)
- Интеллектуальный анализ данных (Data Mining) — 1989
- Knowledge Discovery in Databases — 1989
- Науки о данных (Data Science) — 1997
- Бизнес-аналитика (Business Intelligence, Business Analytics)
- Предсказательная аналитика (Predictive Analytics) — 2007
- Большие данные (Big Data) — 2008
- Аналитика больших данных (Big Data Analytics)

Содержание

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. **Методология машинного обучения**
   - Особенности данных
   - Межотраслевой стандарт CRISP-DM
   - Эксперименты на синтетических и реальных данных

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;
- в каком смысле «a приближает y»;
- как строить функцию a.

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 $Y = \mathbb{R}^m$.

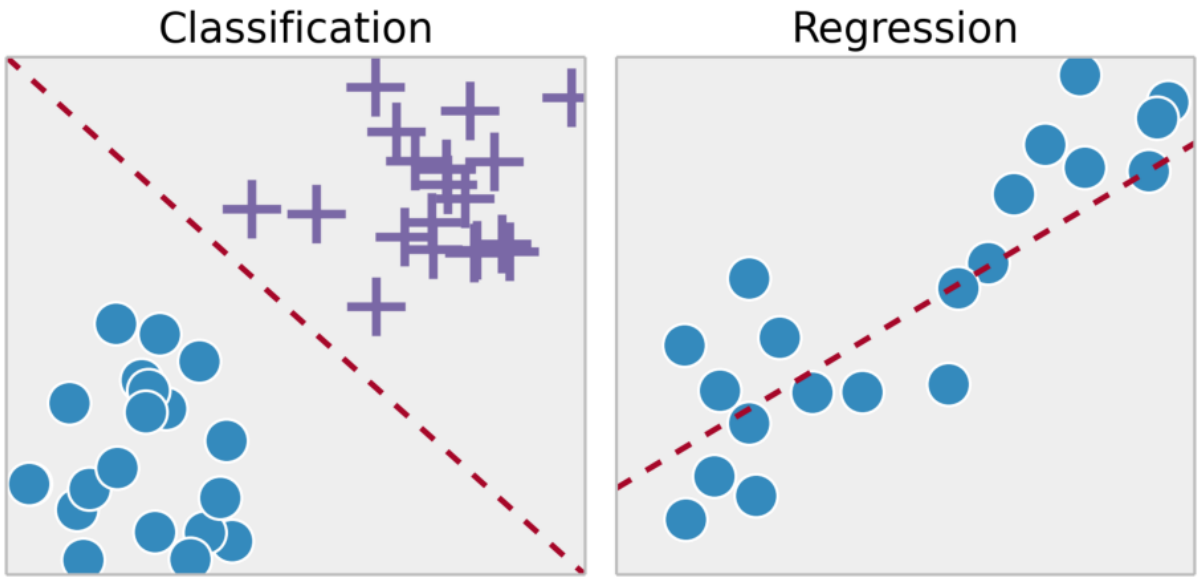
Задачи ранжирования (ranking, learning to rank):

- Y — конечное упорядоченное множество.

Статистическое (машинное) обучение с учителем

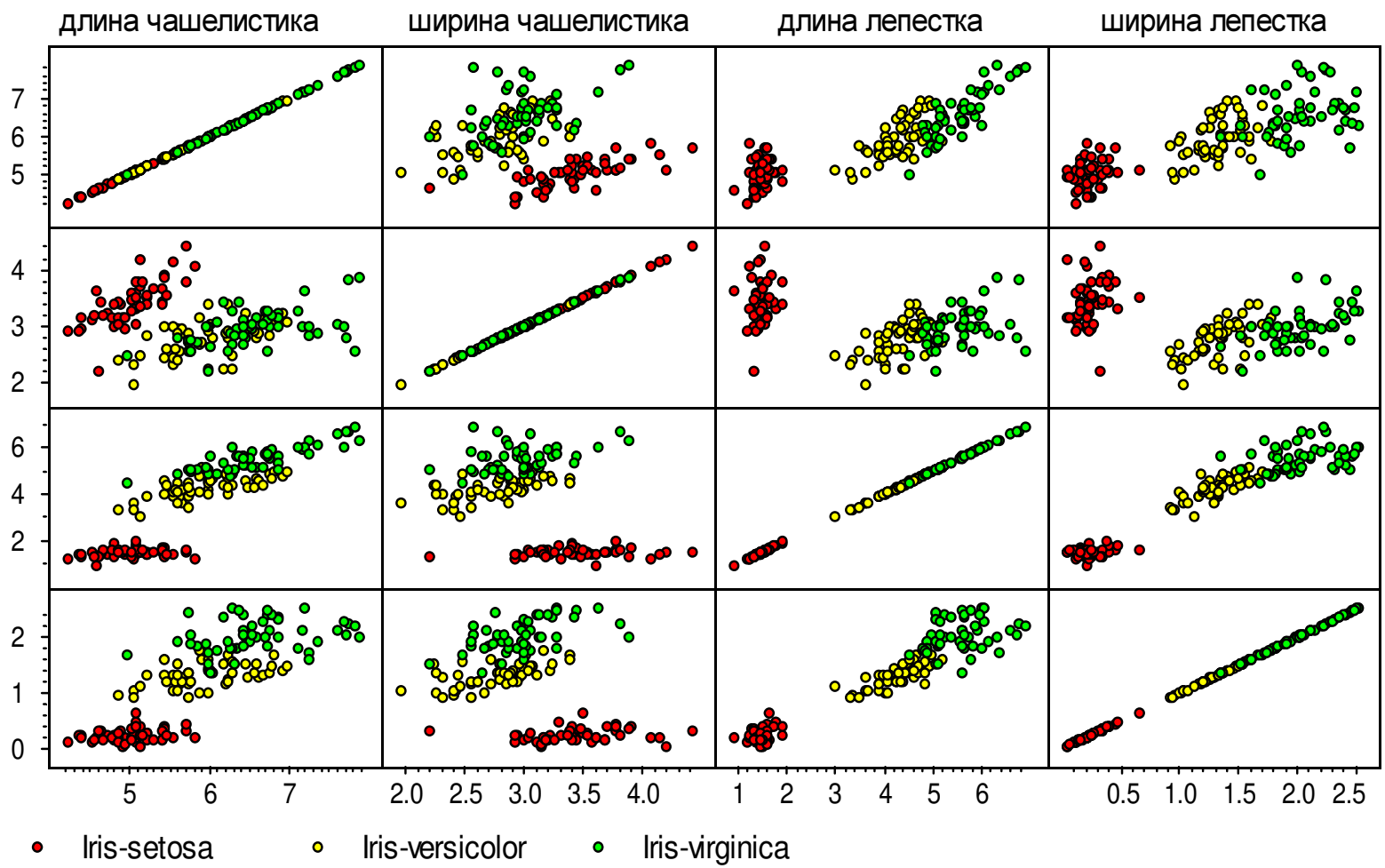
= обучение по прецедентам
= восстановление зависимостей по эмпирическим данным
= предсказательное моделирование
= проведение функции через заданные точки

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 алгоритмов (предсказательная модель)

Модель (predictive model) — параметрическое семейство функций

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

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

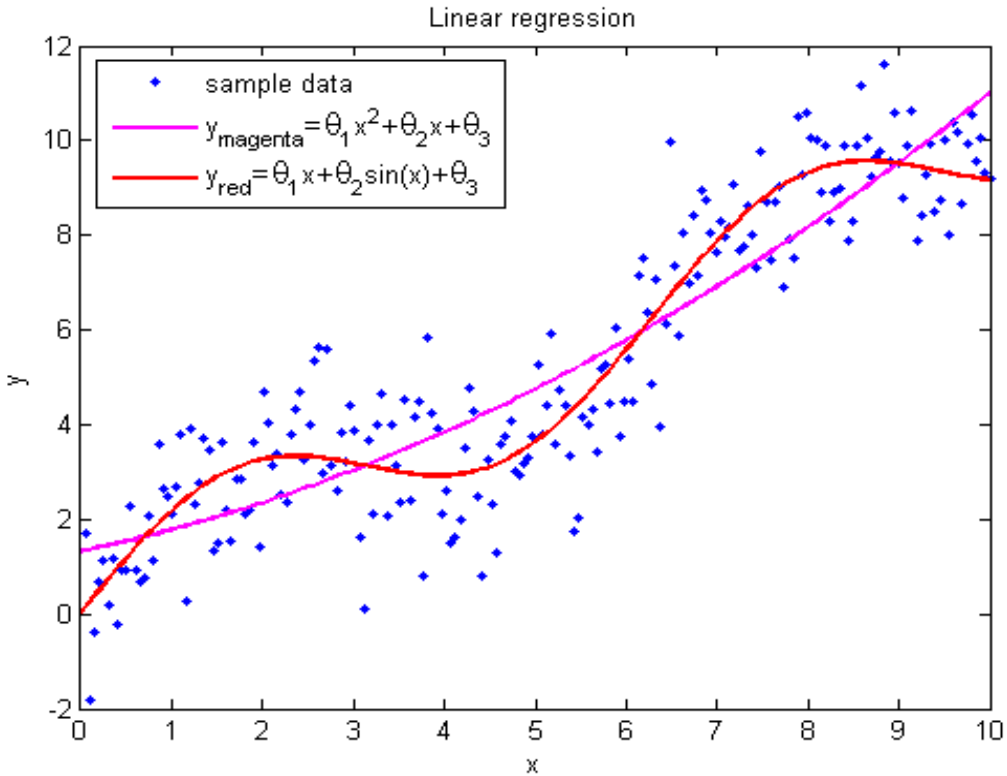
Пример.
Линейная модель с вектором параметров $\theta = (\theta_1, \ldots, \theta_n)$, $\Theta = \mathbb{R}^n$:

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

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

Метод обучения

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'_i выдаёт ответы $a(x'_i)$.

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

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:

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

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

Сведение задачи обучения к задаче оптимизации

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

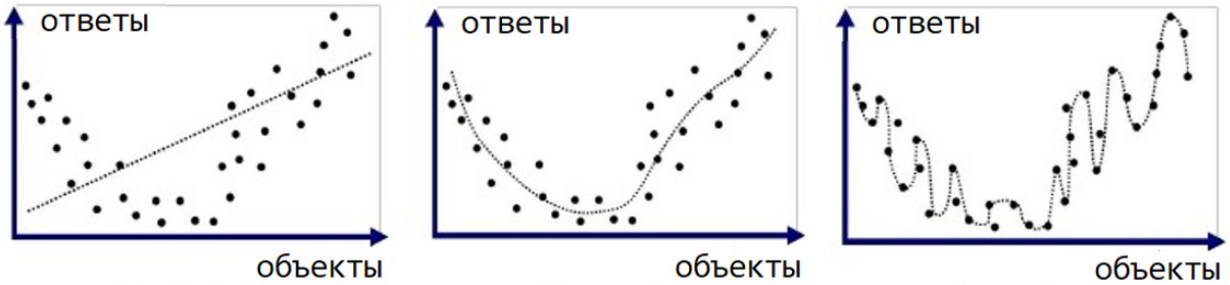
$$\mu(X^\ell) = \arg\min_{a \in A} Q(a, X^\ell).$$

Пример: задача регрессии, $Y = \mathbb{R}$;
n числовых признаков $f_j : X \to \mathbb{R}$, $j = 1, \ldots, n$;
линейная модель регрессии: $g(x_i, \theta) = \sum_{j=1}^{n} \theta_j f_j(x)$, $\theta \in \mathbb{R}^n$;
квадратичная функция потерь: $\mathscr{L}(a, x) = \big(a(x) - y(x)\big)^2$.

Частный случай ERM — *метод наименьших квадратов*:

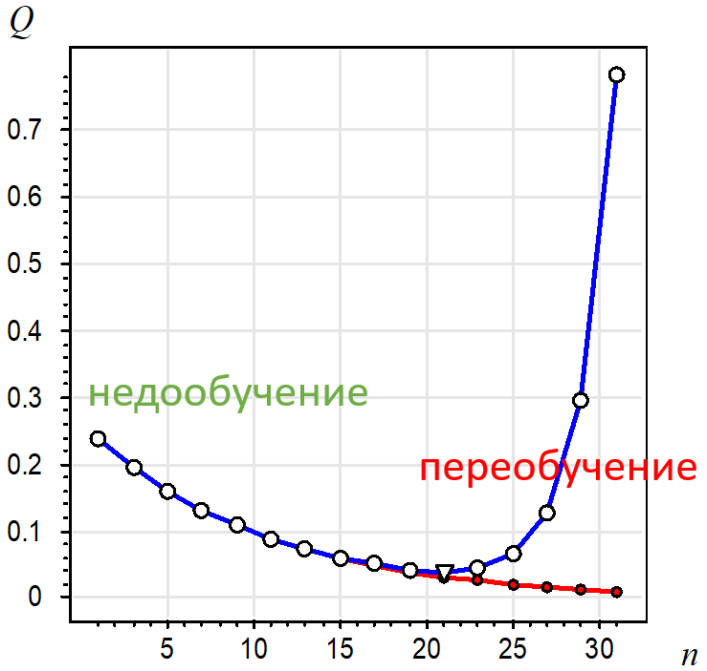
$$\mu(X^\ell) = \arg\min_{\theta} \sum_{i=1}^{\ell} \big(g(x_i, \theta) - y_i\big)^2.$$

Проблемы недообучения и переобучения



недообучение переобучение

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



Пример недообучения и переобучения

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:

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

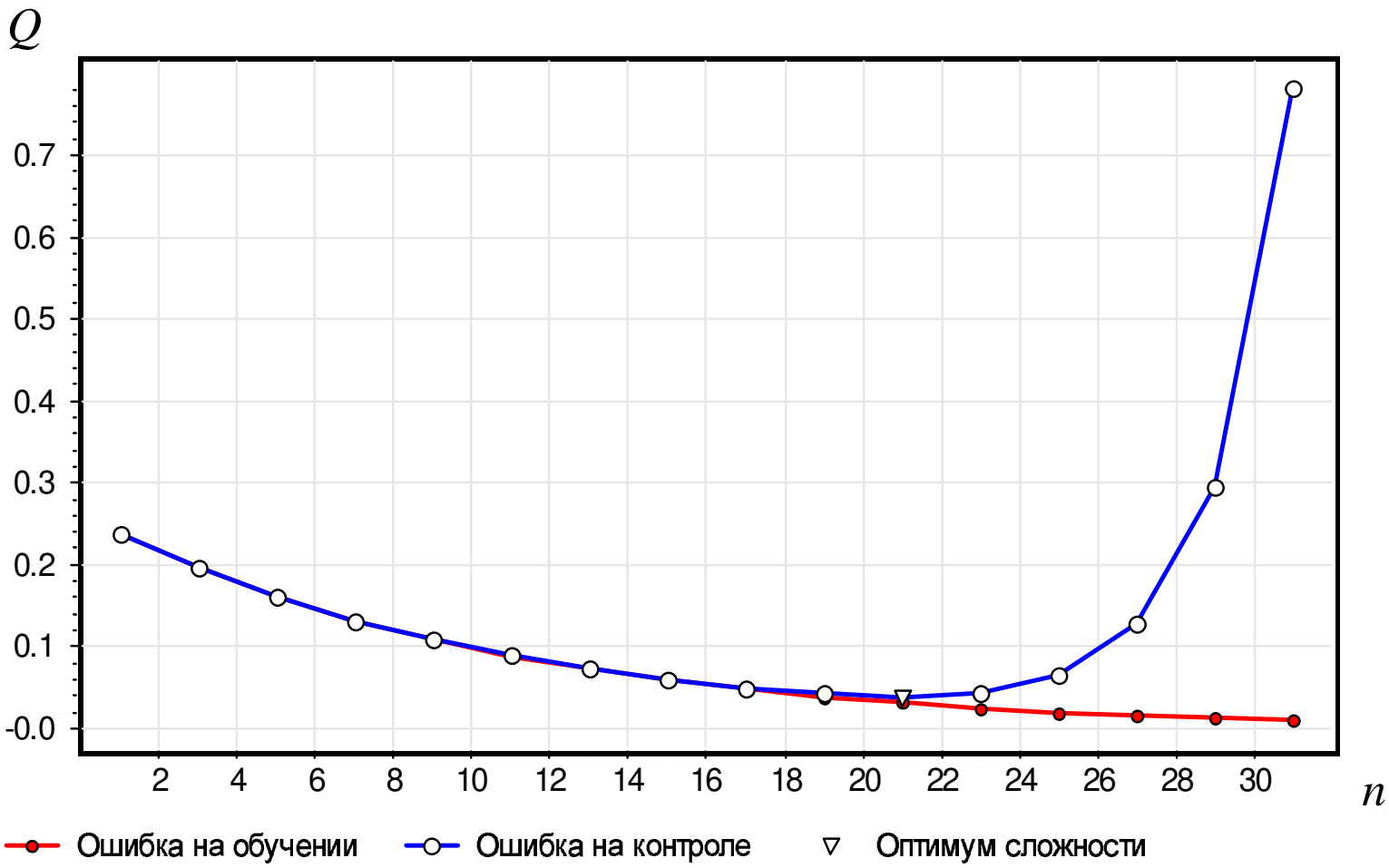
Обучающая выборка: $X^\ell = \left\{ x_i = 4\frac{i-1}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell \right\}$.

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

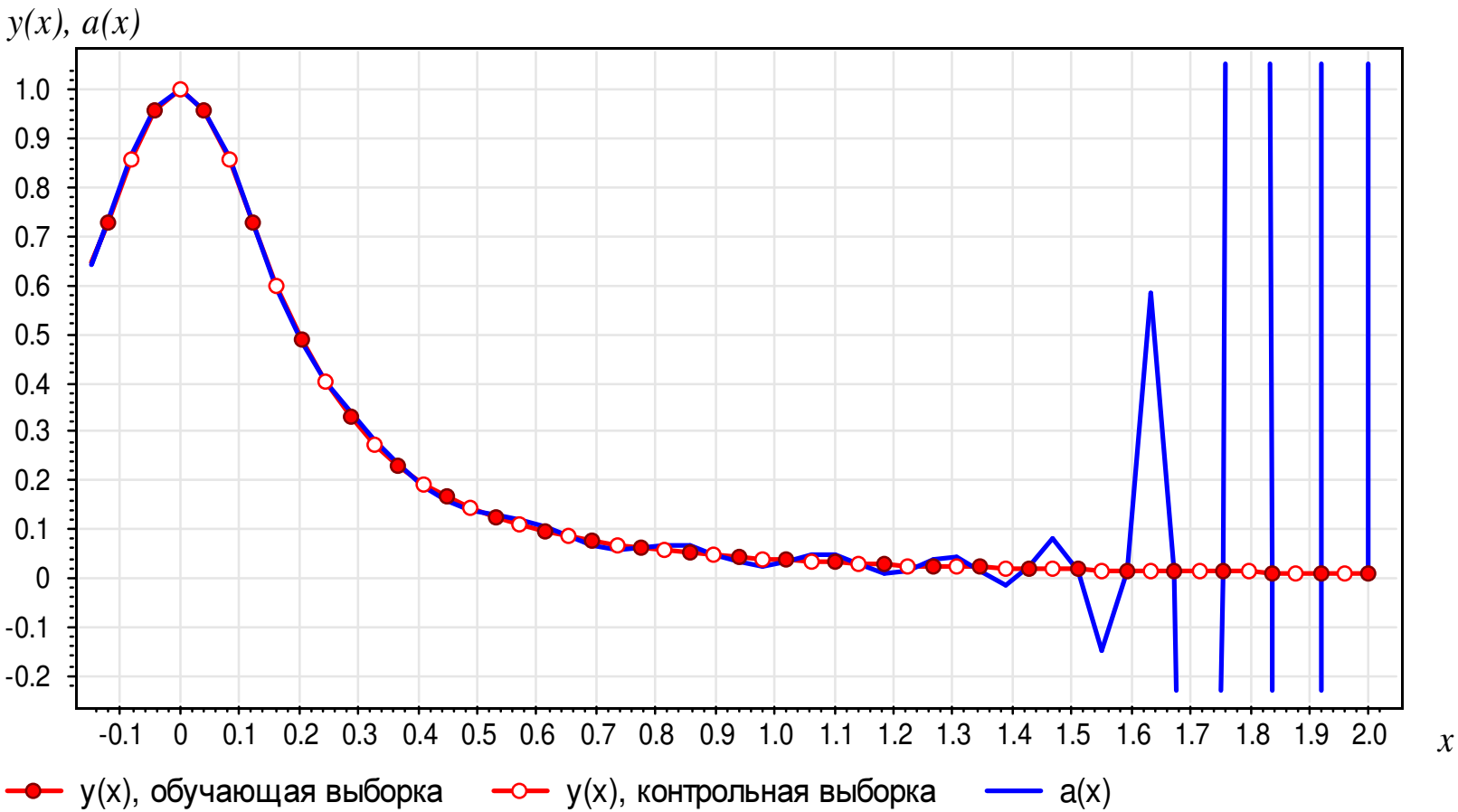
Пример переобучения: эксперимент при $\ell = 50,\ n = 1..31$

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



Пример переобучения: эксперимент при $\ell = 50$

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



Переобучение — одна из проблем машинного обучения

1. **Из-за чего возникает переобучение?**
   — избыточная сложность пространства параметров Θ,
   лишние степени свободы в модели $g(x, \theta)$ «тратятся»
   на чрезмерно точную подгонку под обучающую выборку;
   — переобучение есть всегда, когда есть выбор (a из A)
   по неполной информации (по конечной выборке X^ℓ).

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

3. **Избавиться от него нельзя. Как его минимизировать?**
   — минимизировать HoldOut, LOO или CV, но осторожно!
   — накладывать ограничения на θ (регуляризация);
   — минимизировать одну из теоретических оценок;

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

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

- Эмпирическая оценка вероятности переобучения:

$$Q_\varepsilon(\mu, X^L) = \frac{1}{|N|}\sum_{n \in N} \Big[Q\big(\mu(X_n^\ell), X_n^k\big) - Q\big(\mu(X_n^\ell), X_n^\ell\big) \geqslant \varepsilon \Big] \to \min$$

Задачи медицинской диагностики

Объект — пациент в определённый момент времени.

Классы: диагноз или способ лечения или исход заболевания.

Примеры признаков:

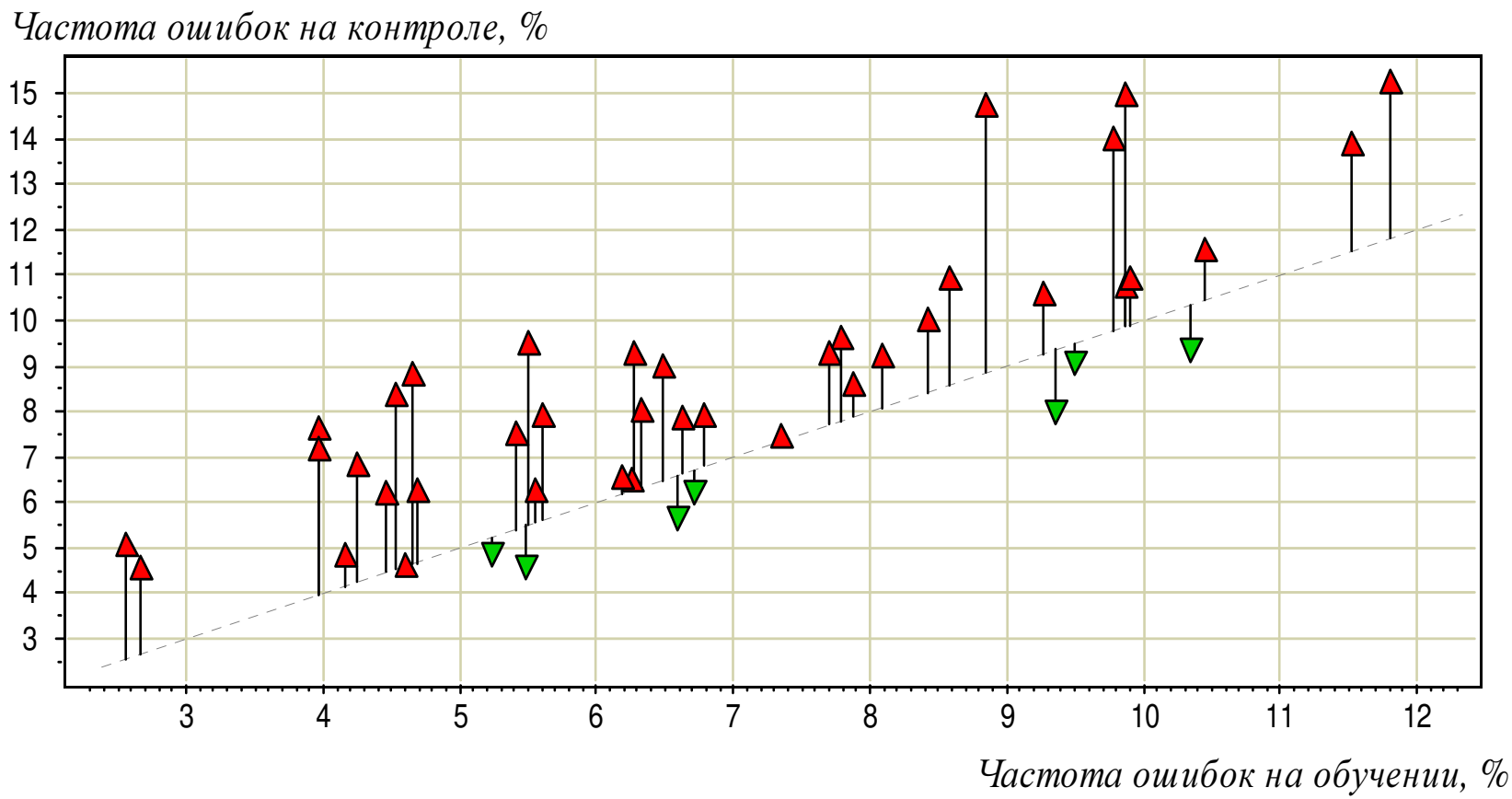
- **бинарные:** пол, головная боль, слабость, тошнота, и т. д.
- **порядковые:** тяжесть состояния, желтушность, и т. д.
- **количественные:** возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т. д.

Особенности задачи:

- обычно много «пропусков» в данных;
- 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.



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.

Задачи биометрической идентификации личности

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

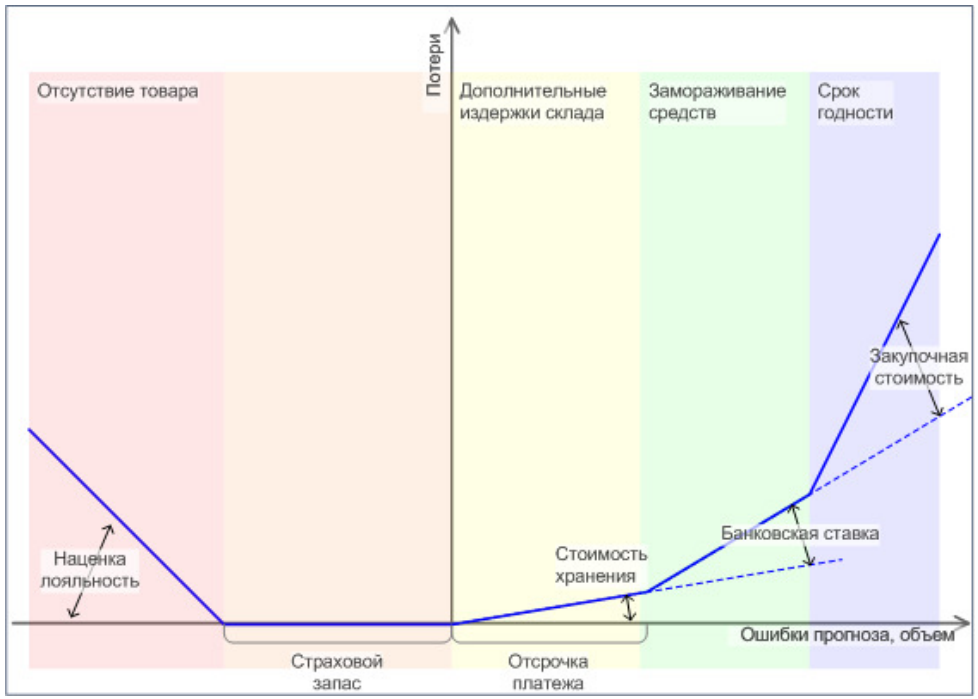
Объект — тройка $\langle$товар, магазин, день$\rangle$.

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.



Конкурс kaggle.com: TFI Restaurant Revenue Prediction

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 $\langle$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.

Задача тематического информационного поиска

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

Предсказать — оптимальный порядок чтения документов.

Примеры признаков:

- близость тематических представлений пары текстов,
- широта/узость тематики,
- широта/узость, уникальность, актуальность терминологии,
- когнитивная сложность текста.

Особенности задачи:

- темы латентные, их надо сначала выявить;
- разнообразие лингвистической предобработки текста;
- графическая поисковая выдача в виде «карты знаний».

Конкурс kaggle.com: Avito Context Ad Clicks Prediction

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;

Машинное обучение на данных сложной структуры

- **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

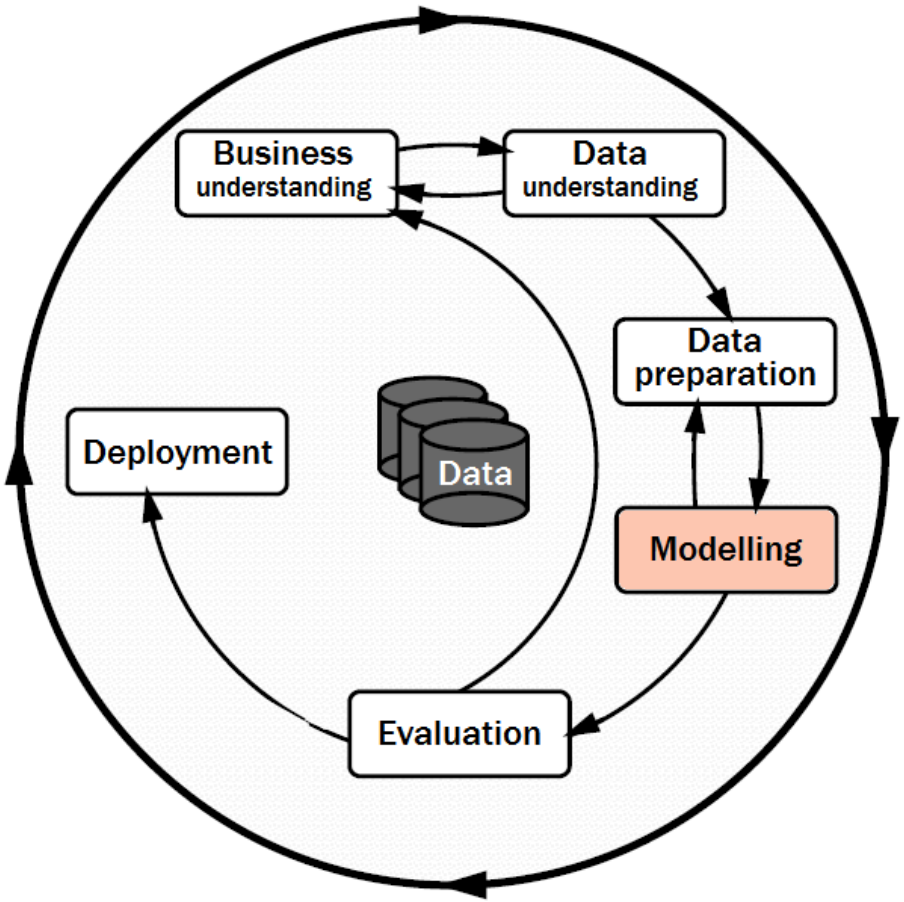
- разнородные (признаки измерены в разных шкалах)
- неполные (измерены не все, имеются пропуски)
- неточные (измерены с погрешностями)
- противоречивые (объекты одинаковые, ответы разные)
- избыточные (сверхбольшие, не помещаются в память)
- недостаточные (объектов меньше, чем признаков)
- неструктурированные (нет признаковых описаний)

Риски, связанные с постановкой задачи:

- «грязные» данные
  (заказчик не обеспечивает качество данных)
- неясные критерии качества модели
  (заказчик не определился с целями и бизнес-процессом)

Межотраслевой стандарт интеллектуального анализа данных

CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



Компании-инициаторы:

- SPSS
- Teradata
- Daimler AG
- NCR Corp.
- OHRA

Шаги процесса:

- понимание бизнеса
- понимание данных
- предобработка данных
- инженерия признаков
- разработка моделей
- настройка параметров
- оценивание качества
- внедрение

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
- отечественная платформа: http://DataRing.ru

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 (488 задач, 2020-02-11)

Эксперименты на синтетических данных

Используются для тестирования новых методов обучения.
Преимущество — мы знаем истинную $y(x)$ (ground truth)

Эксперименты на синтетических данных:

- цель — отладить метод, выявить границы применимости
- объекты x_i из придуманного распределения (часто 2D)
- ответы $y_i = y(x_i)$ для придуманной функции $y(x)$
- двумерные данные + визуализация выборки

Эксперименты на полу-синтетических данных:

- цель — протестировать помехоустойчивость модели
- объекты x_i из реальной задачи (+ шум)
- ответы $y_i = a(x_i)$ для полученного решения $a(x)$ (+ шум)

Резюме в конце лекции

- **Основные понятия машинного обучения:**
  объект, ответ, признак, алгоритм, модель алгоритмов, метод обучения, эмпирический риск, переобучение.

- **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;
  — внедрение и эксплуатация.

- **Прикладные задачи машинного обучения:**
  очень много, очень разных,
  во всех областях бизнеса, науки, производства.