

# Машинное обучение: вводная лекция

К. В. Воронцов, А.А. Романенко  
vokov@forecsys.ru  
alexromsput@gmail.com

12 февраля 2016

## 1 Основные понятия и обозначения

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

## 2 Примеры прикладных задач

- Задачи классификации
- Задачи регрессии
- Задачи ранжирования

## 3 Методология машинного обучения

- Эксперименты на модельных и реальных данных
- Межотраслевой стандарт CRISP-DM
- Резюме

## Задача обучения по прецедентам

$X$  — множество *объектов*;

$Y$  — множество *ответов*;

$y: X \rightarrow Y$  — 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;
- как строится функция  $a$ ;
- что значит « $a$  приближает  $y$  на всём  $X$ ».

## Объекты и признаки

$f_j: X \rightarrow D_j, j = 1, \dots, n$  — 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» (features data)

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

## Ответы и типы задач

### Задачи классификации (classification):

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

### Задачи восстановления регрессии (regression):

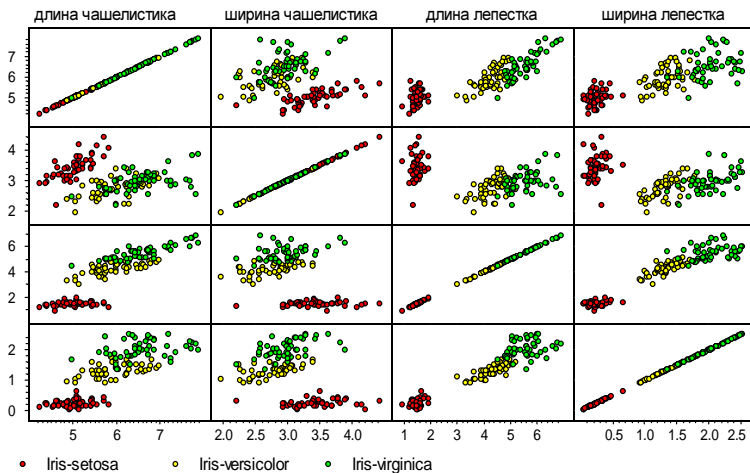
- $Y = \mathbb{R}$  или  $Y = \mathbb{R}^m$ .

### Задачи ранжирования (ranking, learning to rank):

- $Y$  — конечное упорядоченное множество.

## Пример: задача классификации цветков ириса [Фишер, 1936]

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



## Модель алгоритмов

*Модель алгоритмов* — параметрическое семейство отображений

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

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

**Пример.**

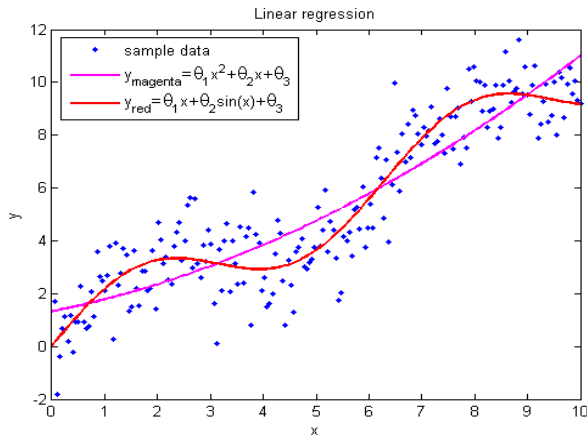
*Линейная модель* с вектором параметров  $\theta = (\theta_1, \dots, \theta_n)$ ,  $\Theta = \mathbb{R}^n$ :

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

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

## 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



## Метод обучения

*Метод обучения* (learning algorithm) — это отображение вида

$$\mu: (X \times Y)^\ell \rightarrow A,$$

которое произвольной выборке  $X^\ell = (x_i, y_i)_{i=1}^\ell$   
ставит в соответствие некоторый алгоритм  $a \in A$ .

В задачах обучения по прецедентам всегда есть два этапа:

- *Этап обучения* (training):  
метод  $\mu$  по выборке  $X^\ell$  строит алгоритм  $a = \mu(X^\ell)$ .
- *Этап применения* (testing):  
алгоритм  $a$  для новых объектов  $x$  выдаёт ответы  $a(x)$ .

## Функционалы качества

$\mathcal{L}(a, x)$  — *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  $X^\ell$ :

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

## Сведение задачи обучения к задаче оптимизации

*Метод минимизации эмпирического риска:*

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

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

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

**Проблема обобщающей способности:**

- 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*  $X^k = (x'_i, y'_i)_{i=1}^k$ ,  $y'_i = y(x_i)$ ?

## Пример переобучения

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

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

Модель полиномиальной регрессии

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

Обучение методом наименьших квадратов:

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

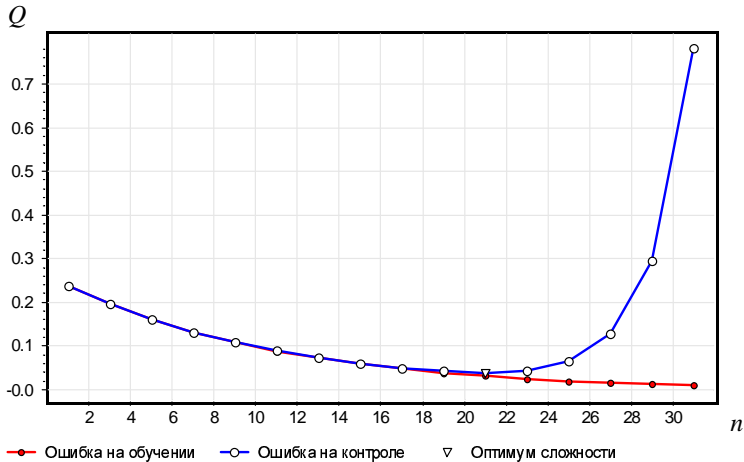
Обучающая выборка:  $X^\ell = \{x_i = 4\frac{i-1}{\ell-1} - 2 \mid i = 1, \dots, \ell\}$ .

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  $n$ ?

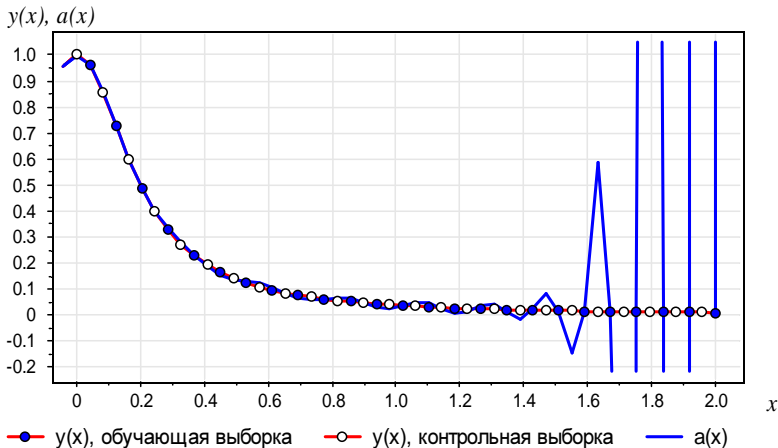
## Пример переобучения: эксперимент при $\ell = 50$ , $n = 1.31$

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



## Пример переобучения: эксперимент при $\ell = 50$

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



## Формализации понятия «обобщающая способность»

- Эмпирическая оценка на отложенных данных (hold-out):

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

- Эмпирическая оценка скользящего контроля (cross-validation):

$$\text{CV}(\mu, X^{\ell+k}) = \frac{1}{|N|} \sum_{n \in N} Q(\mu(X_n^\ell), X_n^k) \rightarrow \min;$$

- Теоретическая оценка ожидаемой потери  
(вероятность ошибки, если функция потерь бинарная):

$$\text{EQ}(\mu(X^\ell), X^k) \rightarrow \min;$$

- Теоретическая оценка вероятности переобучения:

$$Q_\varepsilon(\mu) = P \left[ Q(\mu(X^\ell), X^k) - Q(\mu(X^\ell), X^\ell) \geq \varepsilon \right] \rightarrow \min;$$

## Переобучение — одна из проблем машинного обучения

### ❶ Из-за чего возникает переобучение?

- избыточная сложность пространства параметров  $\Theta$ , лишние степени свободы в модели  $g(x, \theta)$  «тратятся» на чрезмерно точную подгонку под обучающую выборку.
- переобучение есть всегда, когда есть оптимизация параметров по конечной (заведомо неполной) выборке.

### ❷ Как обнаружить переобучение?

- эмпирически, с помощью скользящего контроля.

### ❸ Избавиться от него нельзя. Как его минимизировать?

- минимизировать одну из теоретических оценок;
- накладывать ограничения на  $\theta$  (регуляризация);
- минимизировать HoldOut или CV, но осторожно!



## Задачи медицинской диагностики

**Объект** — пациент в определённый момент времени.

**Классы:** способы лечения или исходы 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.

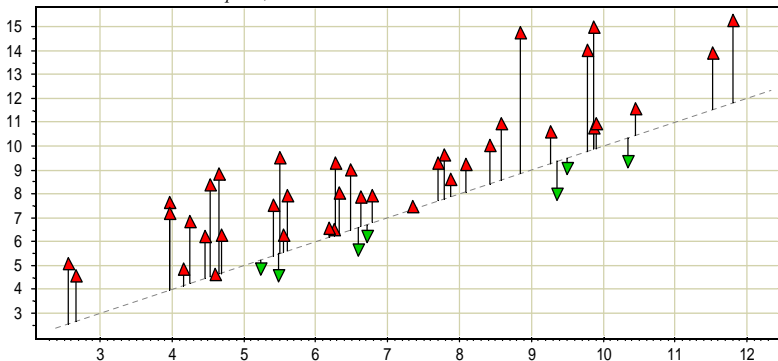
**Особенности задачи:**

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

## Пример переобучения. Реальная задача классификации

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

*Частота ошибок на контроле, %*



*Частота ошибок на обучении, %*

## 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 объёмов продаж

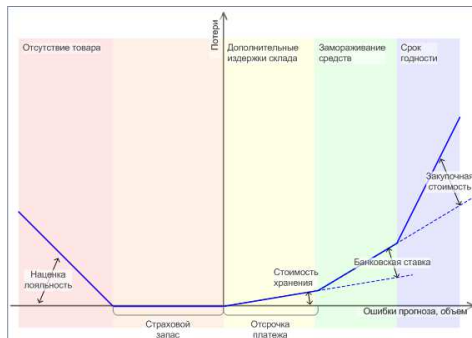
Объект — тройка  $\langle \text{товар, магазин, день} \rangle$ .

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.



## Задача ранжирования поисковой выдачи

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

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

**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:**

- оптимизируется не число ошибок, а качество ранжирования;
- сверхбольшие выборки;
- проблема конструирования признаков по сырым данным.



## Эксперименты на реальных данных

### Эксперименты на конкретной прикладной задаче:

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

### Эксперименты на наборах прикладных задач:

- цель — протестировать метод «в реальных условиях»
- нет необходимости (и времени) разбираться в сути задач : (
- признаки уже кем-то придуманы
- репозиторий UC Irvine Machine Learning Repository  
<http://archive.ics.uci.edu/ml/>

## Эксперименты на модельных данных

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

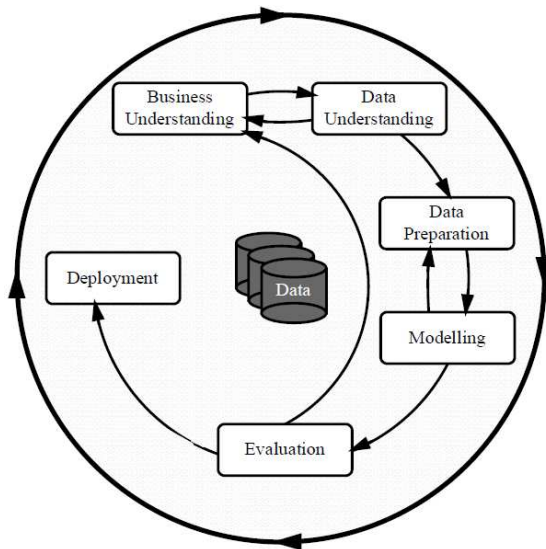
### Эксперименты на модельных данных:

- цель — отладить метод, выявить границы применимости
- объекты — из придуманного распределения
- ответы — заданная функция  $y(x)$
- двумерные данные + визуализация выборки

### Эксперименты на полумодельных данных:

- цель — протестировать помехоустойчивость модели
- объекты — из реальной задачи + шум
- ответы — заданная функция  $y(x)$  + шум

## CRISP-DM: CRoss Industry Standard Process for Data Mining



- **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 переобучения и эффективности;
  - оценивание качества;
  - внедрение и эксплуатация.
- **Прикладные задачи машинного обучения:**  
очень много, очень разных,  
во всех областях бизнеса, науки, производства.