Машинное обучение (Machine Learning)

Введение. Основные понятия

Уткин Л.В.

ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

Содержание

Презентация является компиляцией и заимствованием материалов из замечательных курсов и презентаций по машинному обучению:

К.В. Воронцова, А.Г. Дьяконова, Н.Ю. Золотых, С.И. Николенко, Andrew Moore, Lior Rokach, Rong Jin, Jessica Lin, Luis F. Teixeira, Alexander Statnikov и других.

Основные понятия

Что такое машинное обучение (machine learning)?

Машинное обучение *– это подраздел ИИ, включающий методы построения алгоритмов, способных обучаться.*

Машинное обучение *– подраздел ИИ, математическая дисциплина, использующая разделы математической статистики, численных методов оптимизации, теории вероятностей, дискретного анализа,* ***выделяющая знания из данных****. (из Википедии)*

Машинное обучение *изучает методы построения алгоритмов, которые могут обучаться из данных и делать прогноз на данных.*

Что такое машинное обучение (machine learning)?

Говорят, что компьютерная программа обучается на основе опыта E по отношению к некоторому классу задач T и меры качества P, если качество решения задач из T, измеренное на основе P, улучшается с приобретением опыта E. - *T.M.Mitchell Machine Learning. McGraw-Hill, 1997.*

Дедуктивное и индуктивное методы обучения

— способы обучения и в компьютерных системах:

1. **Дедуктивное**, или аналитическое, обучение (экспертные системы). »меются знания, сформулированные экспертом и как-то формализованные. Программа выводит из этих правил конкретные факты и новые правила.
2. **Индуктивное** обучение (статистическое обучение). На основе эмпирических данных программа строит общее правило. Эмпирические данные могут быть получены самой программой в предыдущие сеансы ее работы или просто предъявлены ей.
3. **Комбинированное** обучение.

От данных к знаниям

Сферы приложения

1. Компьютерное зрение (computer vision)
2. Распознавание речи (speech recognition)
3. Компьютерная лингвистика и обработка естественных языков (natural language processing)
4. Медицинская диагностика
5. Биоинформатика
6. Техническая диагностика
7. Финансовые приложения
8. Рубрикация, аннотирование и упрощение текстов
9. Информационный поиск
10. . . .

Смежные и близкие области

- Pattern Recognition (распознавание образов)
- Data Mining (интеллектуальный анализ данных, включая Big Data)
- Artificial Intelligence (искусственный интеллект)

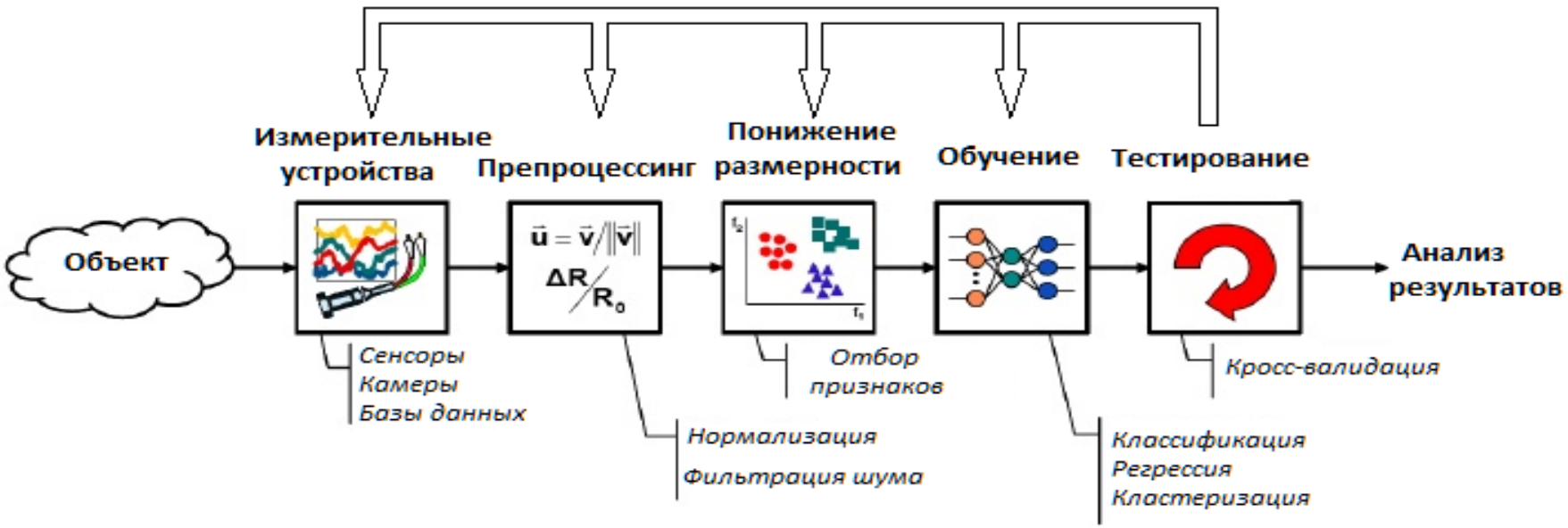
Разделы математики, используемые в машинном обучении

- Линейная алгебра
- Теория вероятностей и математическая статистика
- Методы оптимизации
- Численные методы
- Математический анализ
- Дискретная математика
- и др.

Классификация задач индуктивного обучения

- Обучение с учителем, или обучение по прецедетам (supervised learning):**классификация; восстановление регрессии**; структурное обучение;
- Обучение без учител€ (unsupervised learning): **кластеризация**; визуализаци€ данных; **понижение размерности**;
- Активное обучение (active learning).
- Обучение с подкреплением (reinforcement learning).

Схема всего процесса машинного обучения



Обучение по прецедентам или с учителем

Множество X —**объекты, примеры, образцы** (samples)
Множество Y — **ответы, отклики, «метки», классы** (responses)
Имеется некоторая зависимость $g : X \to Y$, позволяющая по $x \in X$ предсказать (или оценить вероятность появления) $y \in Y$.
Зависимость известна только на объектах из **обучающей выборки**:

$$T = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

Пара $(x_i, y_i) \in X \times Y$ - **прецедент**.
Задача обучения по прецедентам: научиться по новым объектам $x \in X$ предсказывать ответы $y \in Y$.

Пример обучающей выборки (классификаци)

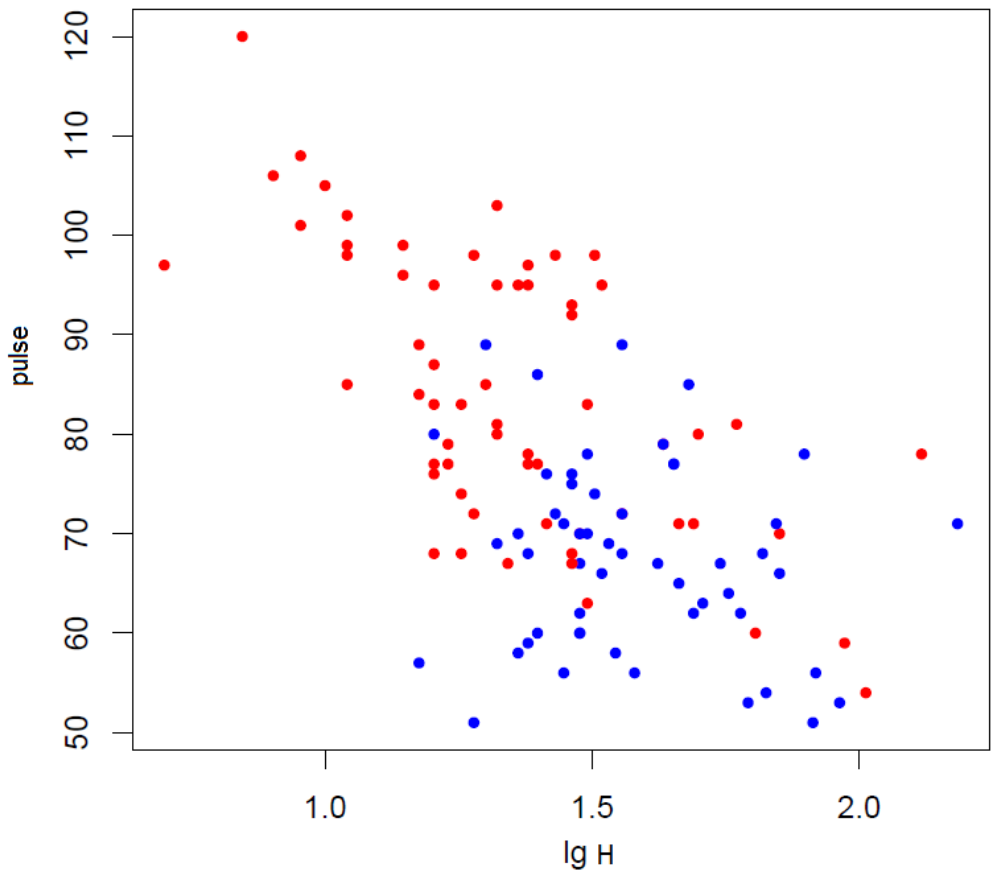
	пульс	гемоглобин	диагноз
x_1	70	140	здоров $(y = -1)$
x_2	60	160	здоров $(y = -1)$
x_3	94	120	миокардит $(y = 1)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_{114}	86	98	миокардит $(y = 1)$

Обучающая выборка:

$$((70, 140), -1), (60, 160), -1), (94, 120), 1)..., (86, 98), 1))$$

Задача обучения: новый пациент $x = (75, 128)$, $y =?$

Графическое представление обучающей выборки



Другой пример обучающей выборки (классификация)

	вес	рост	возраст	ср.дл.волос	пол
x_1	96	170	42	0	м $(y = -1)$
x_2	60	180	25	8	м $(y = -1)$
x_3	54	165	30	21	ж $(y = 1)$
x_4	83	178	47	18	ж $(y = 1)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_{100}	108	193	32	40	ж $(y = 1)$

Задача обучения: $x = (75, 184, 28, 10)$, $y = ?$

Обучающая выборка с категориальными данными

	вес	рост	возраст	ср.дл.волос	пол
x_1	96	170	42	короткие	м $(y = -1)$
x_2	60	180	25	короткие	м $(y = -1)$
x_3	54	165	30	длинные	ж $(y = 1)$
x_4	83	178	47	короткие	ж $(y = 1)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_{100}	108	193	32	длинные	ж $(y = 1)$

Задача обучения: $x = (75, 184, 28, \text{“короткие”})$, $y = ?$

Пример пропущенных данных (missing data)

	вес	рост	возраст	ср.дл.волос	пол
x_1	96	170	42	короткие	м $(y = -1)$
x_2	60	180	25	короткие	-
x_3	54	165	-	длинные	ж $(y = 1)$
x_4	-	178	47	короткие	ж $(y = 1)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_{100}	108	193	32	длинные	ж $(y = 1)$

Задача обучения: $x = (75, 184, 28, \text{“короткие”})$, $y = ?$

Пример ненужного признака

	вес	рост	возраст	ср.дл. волос	оценка по маш.обуч.	пол
x_1	96	170	42	короткие	5	м
x_2	60	180	25	короткие	3	-
x_3	54	165	-	длинные	5	ж
x_4	-	178	47	короткие	4	ж
...	...	...	...	...	...	...
x_{100}	108	193	32	длинные	3	ж

Задача обучения: $x = (75, 184, 28, \text{“короткие”}, 5)$, $y = ?$

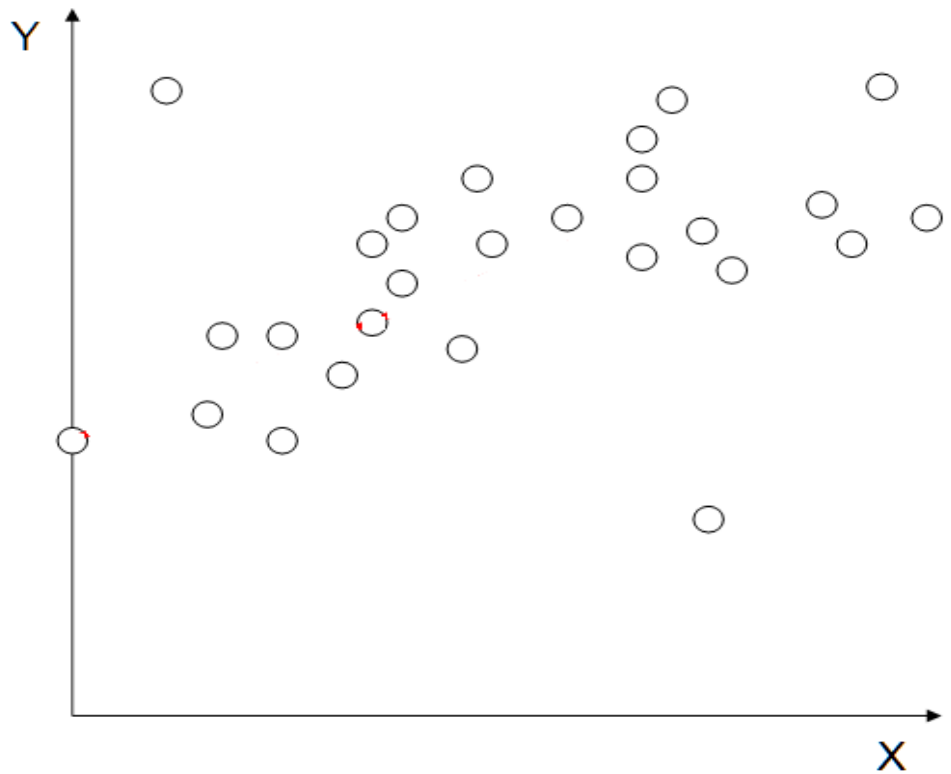
Пример регрессионных данных

	вес	рост	ср.дл. волос	пол	возраст (y)
x_1	96	170	короткие	м	42
x_2	60	180	короткие	м	25
x_3	54	165	длинные	ж	30
x_4	83	178	короткие	ж	47
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_{100}	108	193	длинные	ж	32

Задача обучения: определить возраст

$x = (75, 184, \text{“короткие”}, \text{“м”})$, $y = ?$

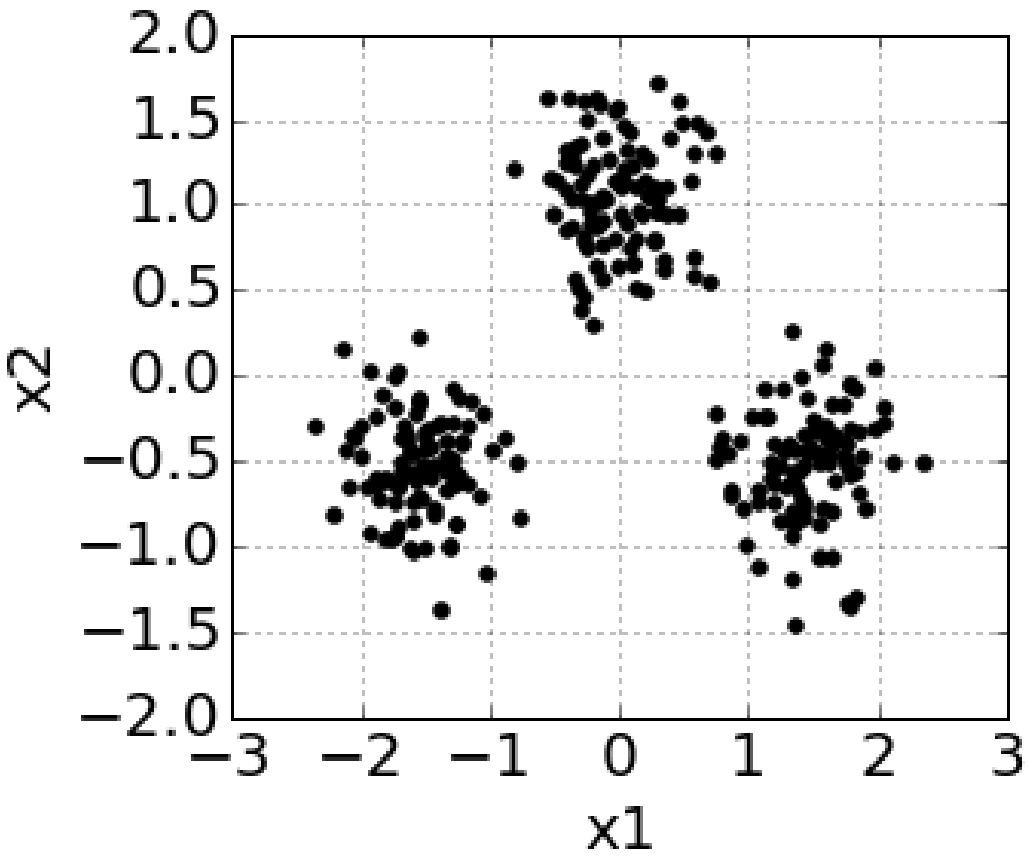
Графическое представление данных для регрессии



Обучение без учителя

- В этом случае нет “учителя” и “обучающая выборка” состоит только из объектов, т.е. Y отсутствует.
- Задача **кластеризации**: разбить объекты на группы (кластеры), так, чтобы в одном кластере оказались близкие друг к другу объекты, а в разных кластерах объекты были существенно различные.
- **Кластер** можно охарактеризовать как группу объектов, имеющих общие свойства.

Графическое представление данных для кластеризации



Пример задачи без учителя

	вес	рост	возраст	ср.дл.волос
x_1	96	170	42	короткие
x_2	60	180	25	короткие
x_3	54	165	30	длинные
x_4	83	178	47	короткие
...	...	...	...	...
x_{100}	108	193	32	длинные

Задача обучения: “отгадать” пол всех людей из обучающей выборки

Признаковые описания

Каждый объект характер изуется набором **признаков** (свойств, атрибутов, features) $f_j : X \to D_j$, $j = 1, ..., n$

Типы признаков:

- $D_j = \{0, 1\}$ бинарный признак;
- $D_j = \{1, 2, 3, ..., s\}$ номинальный (категориальный) признак (красный, зеленый, синий);
- D_j упорядочено - порядковый признак, например, вес:(малый, средний, большой).
- $D_j = \mathbb{R}$ количественный признак

Вектор $(f_1(x), f_2(x), ..., f_n(x))$ - признаковое описание объекта x.

Признаки в примерах определения пола

- **вес**: количественный
- **рост**: количественный
- **возраст**: количественный
- **ср.дл. волос**: бинарный или упорядочено - порядковый или количественный
- **оценка по маш.обуч.**: упорядочено - порядковый или категориальный

Описание меток классов

В зависимости от множества Y выделяют разные типы задачи обучени€:

1. Задачи **классификации** (classification):
   $Y = \{-1, +1\}$ классификация на 2 класса.
   $Y = \{1, ..., M\}$ на M непересекающихся классов.
   $Y = \{0, 1\}^M$ на M классов, которые могут пересекаться.
2. Задачи **восстановления регрессии** (regression):
   $Y = \mathbb{R}$.
3. «Задачи **ранжирования** (ranking, learning to rank): Y - конечное упорядоченное множество.

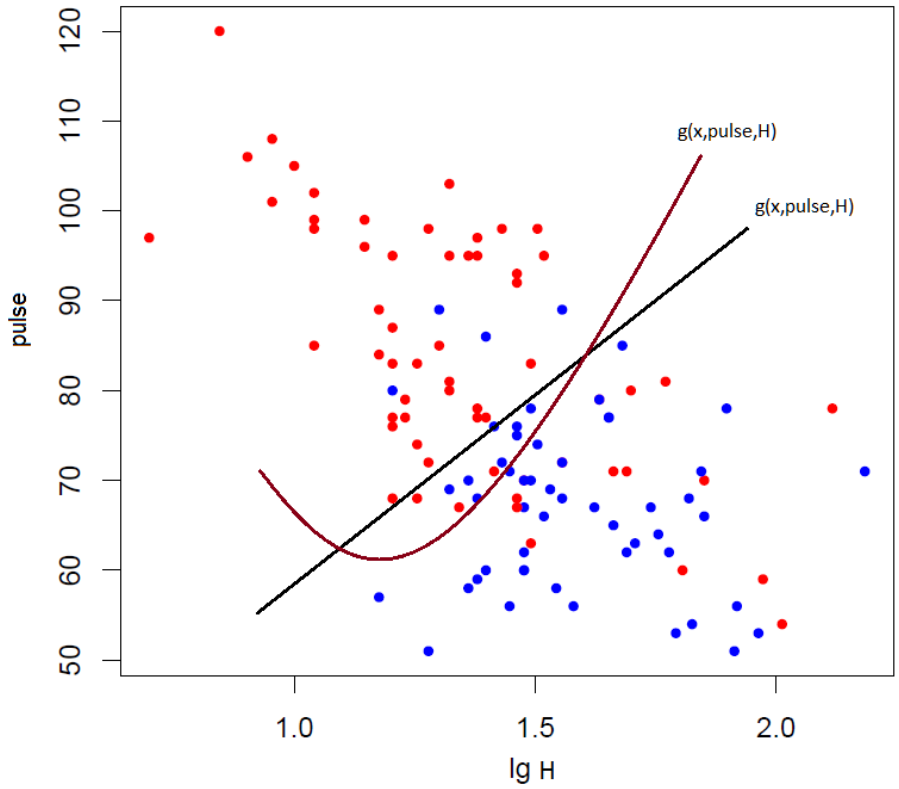
Модель алгоритма

Решить задачу машинного обучения *означает разработать алгоритм или модель алгоритма, зависящего от параметров и позволяющих определить значение метки класса (Y) для нового объекта (x).*

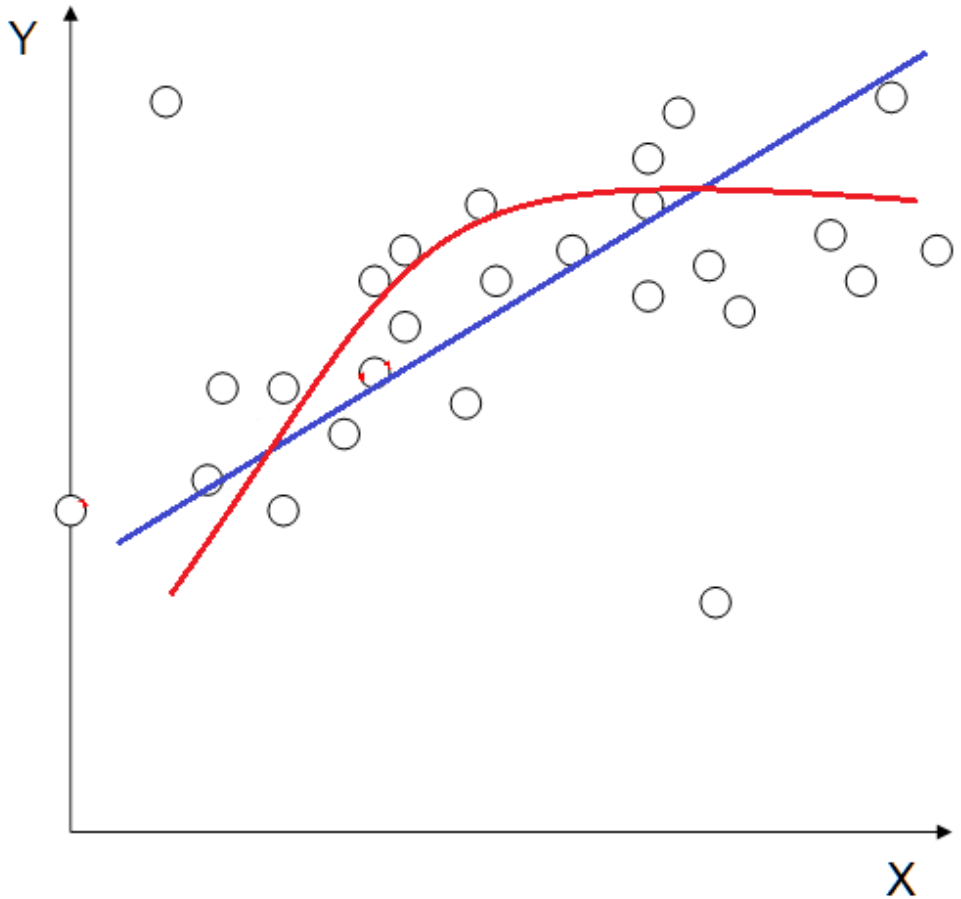
Модель алгоритма

- **Моделью алгоритма** a называется параметрическое семейство функций $g : X \to Y$ или $g(x, \theta)$, где $\theta \in \Theta$ параметры в пространстве параметров.
- **Пример**: В задачах с m признаками $f_j(x)$, $j = 1, ..., m$ используются линейные модели с $\theta = (\theta_1, ..., \theta_m)$: $g(x, \theta) = \sum_{j=1}^{m} \theta_j f_j(x)$
- Процесс подбора оптимальной функции g и оптимального параметра θ по обучающей выборке называют **настройкой** (fitting, tuning) или **обучением** (training) алгоритма a.

Модели алгоритмов классификации



Модели алгоритмов регрессии



Функционал качества

- **Функционал качества** может определяться как средняя ошибка ответов.
- **Функционал риска** или**качества** алгоритма a обучения есть

$$Q(a, X) = \int (\mathcal{L}(a, x) \cdot P(X, y))\, dXdy$$

- **Функция потерь** (loss function) - это неотрицательная функция $L(a, x)$, характеризующая величину ошибки алгоритма a на объекте x. Если $\mathcal{L}(a, x) = 0$, то ответ $a(x)$ называется **корректным**.
- $P(X, y)$ - совместная плотность вероятностей

Функции потерь

- Функции потерь для классификации:
  - $\mathcal{L}(a, x) = [a(x) \neq y(x)]$ - индикатор ошибки
  - $\mathcal{L}(a, x) = \max(0, 1 - y_i a(x))$ - петлевая функция (hinge-loss function)
- Функции потерь для регрессии:
  - $\mathcal{L}(a, x) = |a(x) - y(x)|$ - абсолютное значение ошибки
  - $\mathcal{L}(a, x) = (a(x) - y(x))^2$ - квадратичная ошибка
  - $\mathcal{L}(a, x) = \begin{cases} (y-a)^2/2, & \text{если } |y-a| \leq \delta \\ \delta(|y-a|) - \delta/2, & \text{если } y - a > \delta \end{cases}$ - функция потерь Хьюбера
- Функции потерь для кластеризации: $\mathcal{L}(a, x) = \sum_{i=1}^{n} \min_c \|x_i - a_c\|^2$

Эмпирический функционал качества

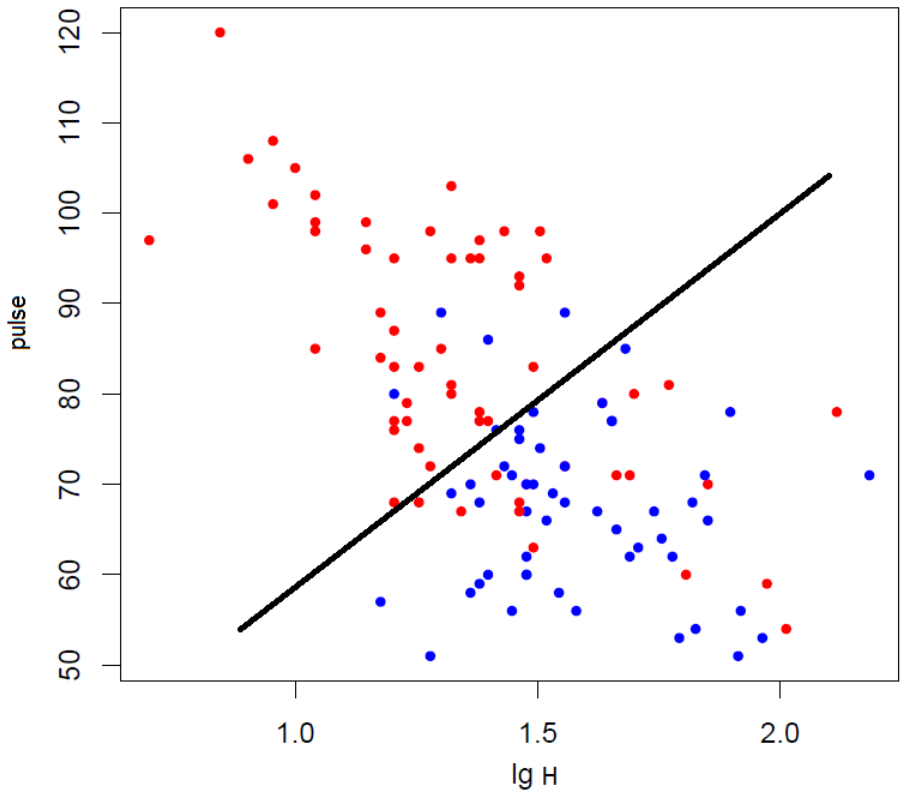
$$Q(a, X) = \int (\mathcal{L}(a, x) \cdot P(X, y))\, dXdy$$

- **Эмпирический функционал риска**или**качества** алгоритма a на выборке X есть

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

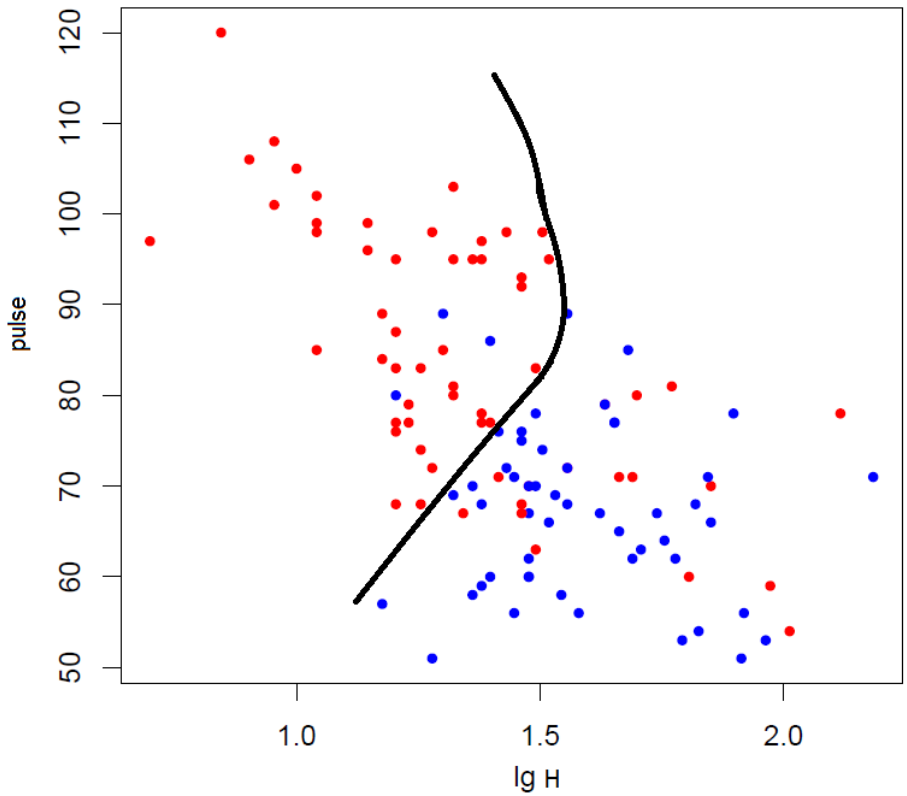
- Плотность $P(X, y)$ в функционале риска заменена на эмпирическое распределение (равномерное распределение) на элементах обучающей выборки.
- **Задача выбора "наилучшего" метода обучения** - это минимизаця функционала риска по множеству A или по множеству параметров Θ.

Эмпирический функционал качества



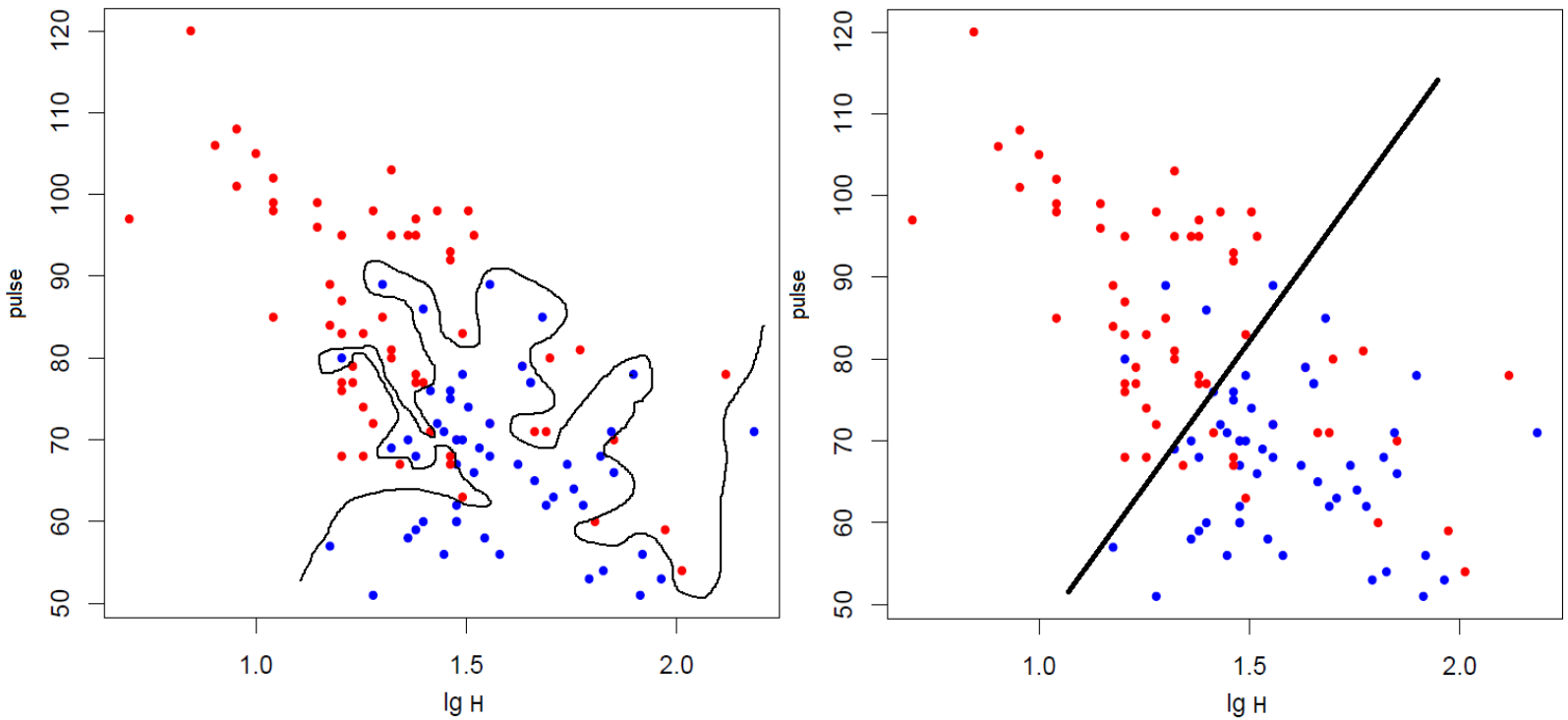
$$Q(a, X) = \frac{1}{n}\sum_{i=1}^{n}[a(x_i) \neq y_i] = \frac{1}{114}(5 + 15) = 0.175$$

Эмпирический функционал качества



$$Q(a, X) = \frac{1}{n}\sum_{i=1}^{n}[a(x_i) \neq y_i] = \frac{1}{114}(3 + 14) = 0.149$$

Переобучение и недообучение в классификации

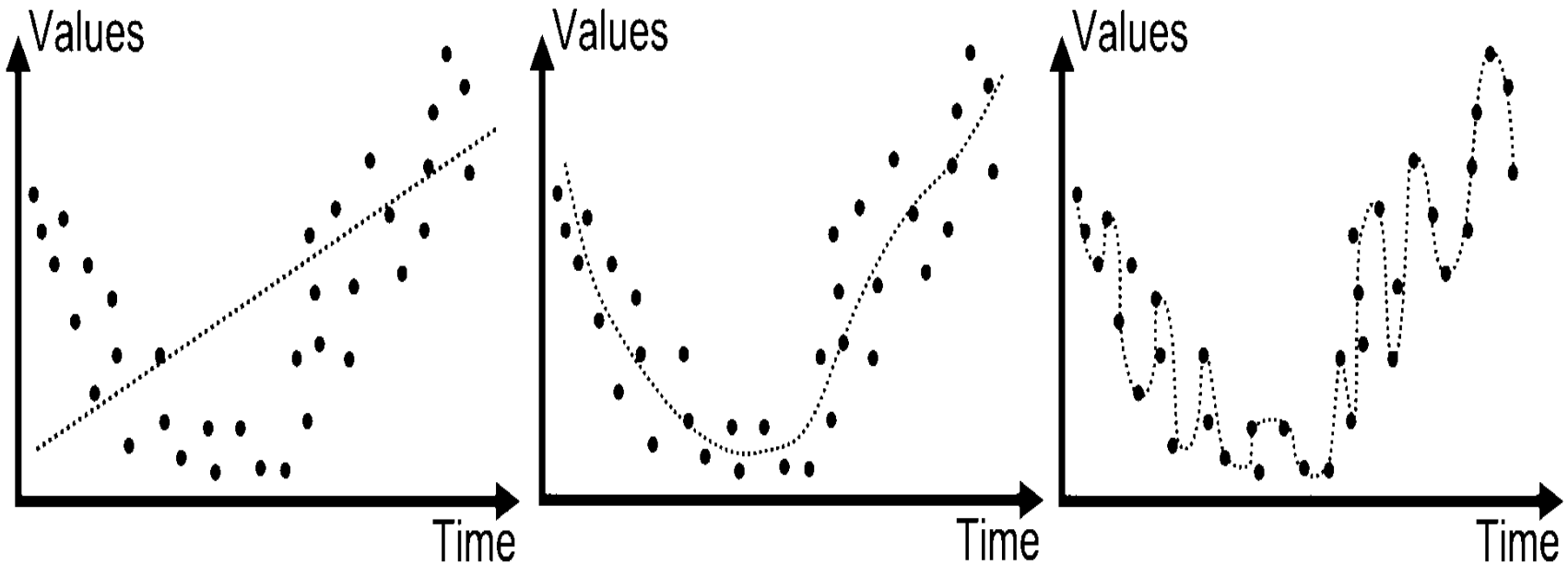


Проблема переобучения и недообучения

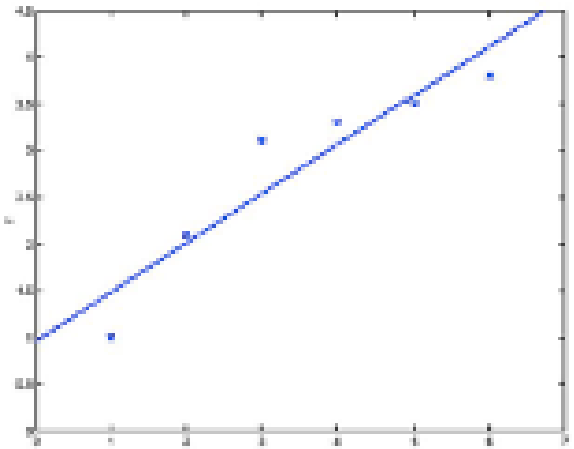
Переобучение (overfitting) — нежелательное явление, возникающее при решении задач обучения по прецедентам, когда вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке. Переобучение возникает при использовании избыточно сложных моделей.

Недообучение — нежелательное явление, возникающее при решении задач обучения по прецеден там, когда алгоритм обучения не обеспечивает достаточно малой величины средней ошибки на обучающей выборке. Недообучение возникает при использовании недостаточно сложных моделей.

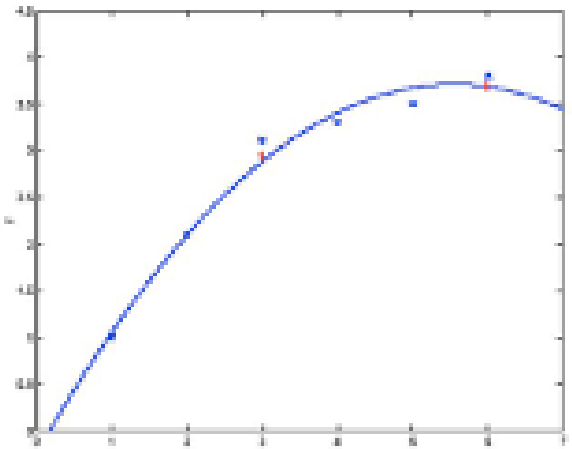
Переобучение и недообучение в регрессии



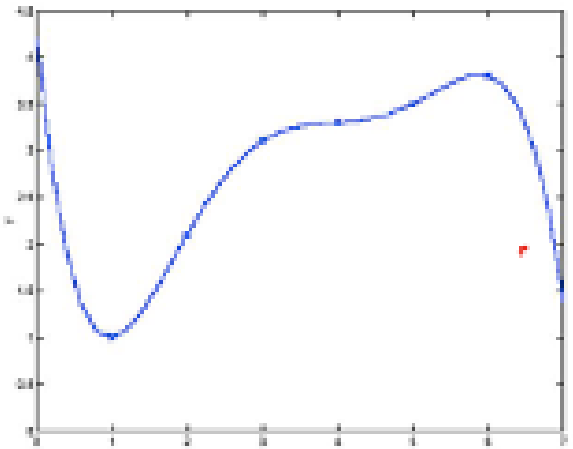
Переобучение и недообучение в регрессии



$$y = \theta_0 + \theta_1 x$$



$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$



$$y = \sum_{j=0}^{5} \theta_j x^j$$

Этапы решения задачи обучения

В задачах обучения по прецедентам всегда есть два этапа:

1. **Этап обучения** (training): по выборке X строится алгоритм a и определяется функция $g(x, \theta)$ с учетом функционала риска алгоритма a
2. **Этап применения или тестирования** (testing): насколько правильные или неправильные ответы $a(x)$ выдает алгоритм a для новых объектов x.

Обучающая и тестовая выборки

Случайно разделим все имеющиеся данные на:

- **обучающую** (train) выборку, которая используется для построения моделей
- **тестовую** (test) выборку, которая используется для оценки как модель ведет себя на новых данных