Уткн с вставкой внцва

Машинное обучение (Machine Learning)

Введение. Основные понятия

ЩЕТИНИН Е.Ю.

2020

Содержание

1. Что такое машинное обучение?
2. Постановки задач:
   - Обучение по прецедентам
   - Обучение без учителя
3. Примеры практических задач
4. О курсе

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). Имеются 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

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. . . .

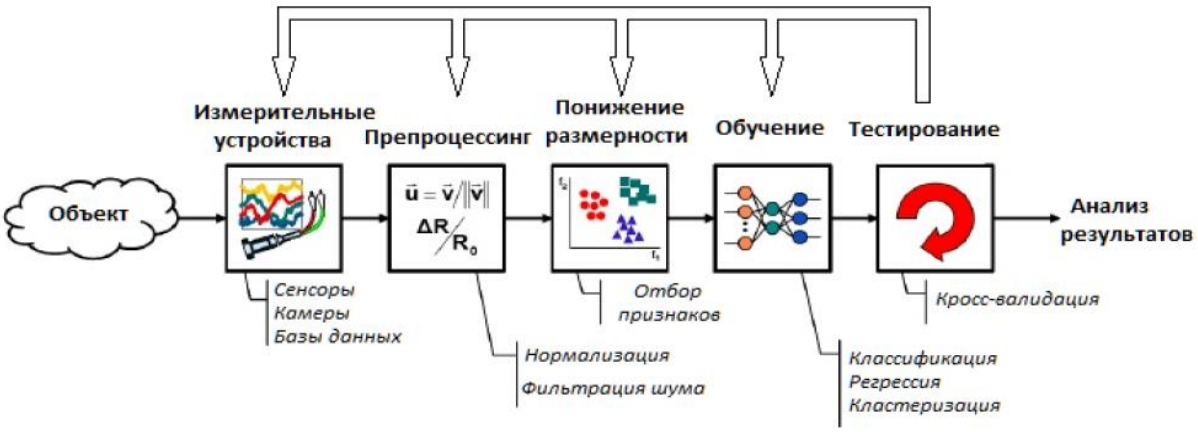
Разделы математики, используемые в машинном обучении

- 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 прецедентам (supervised learning): **классификация; восстановление регрессии**; структурное обучение
- Обучение без учителя (unsupervised learning): **кластеризация**; визуализация 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 (классификация)

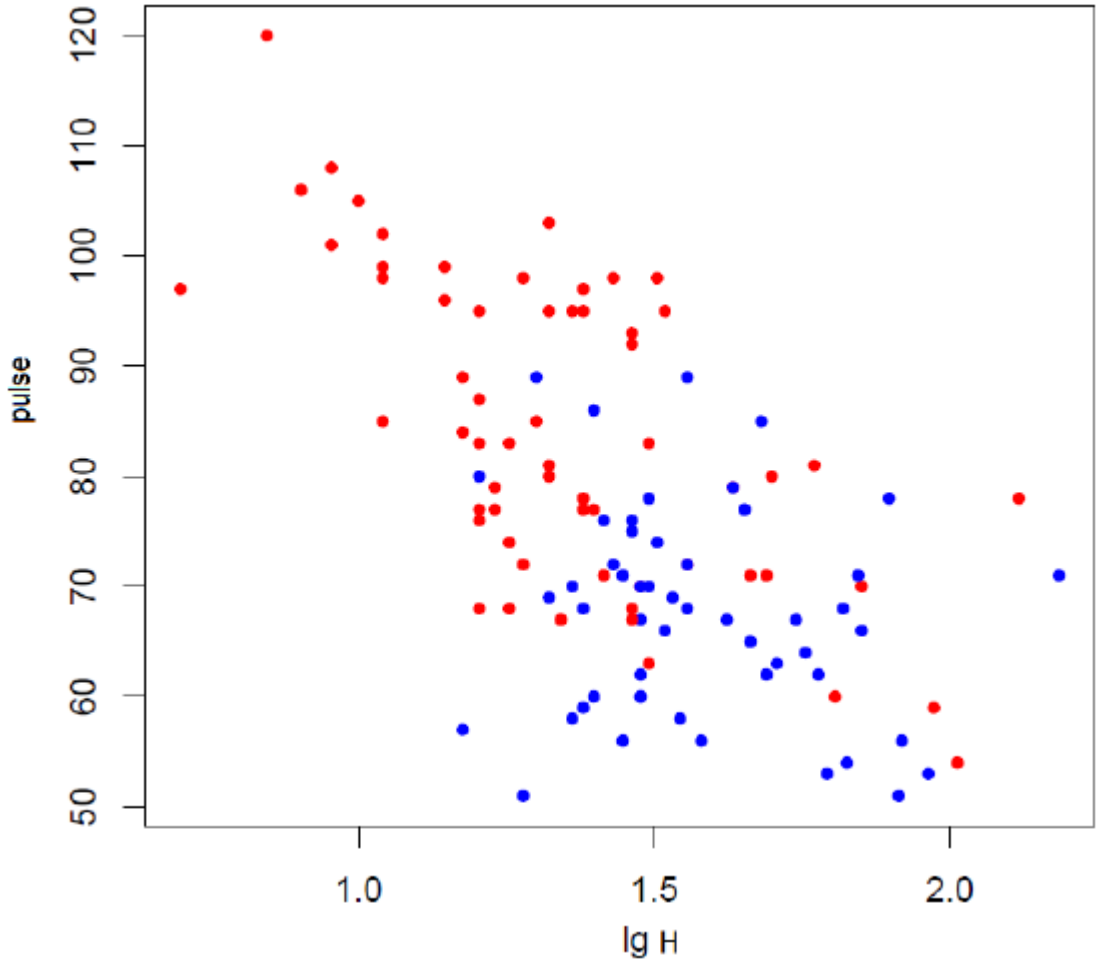
	пульс	гемоглобин	диагноз
x_1	70	140	здоров $(y = -1)$
x_2	60	160	здоров $(y = -1)$
x_3	94	120	миокардит $(y = 1)$
$\dots$	$\dots$	$\dots$	$\dots$
x_{114}	86	98	миокардит $(y = 1)$

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

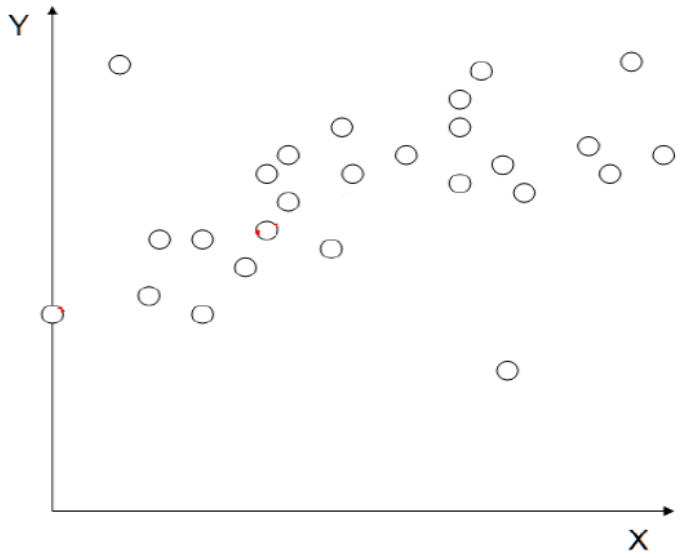
Задача обучения: $x = (75, 184, 28, \text{“короткие”}, 5)$, $y = ?$

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
...	...	...	...	...	...
x_{100}	108	193	длинные	ж	32

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

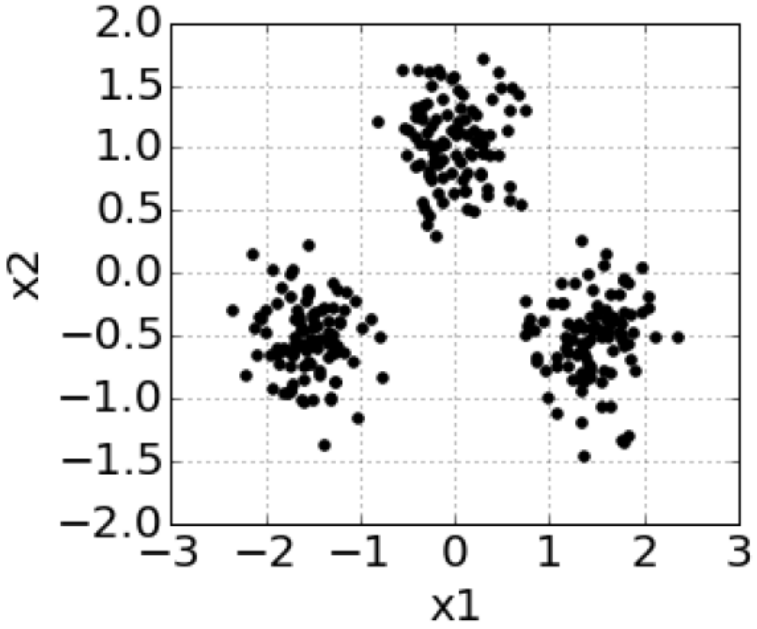
Графическое представление данных для регрессии



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 характеризуется 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 обучения:

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.

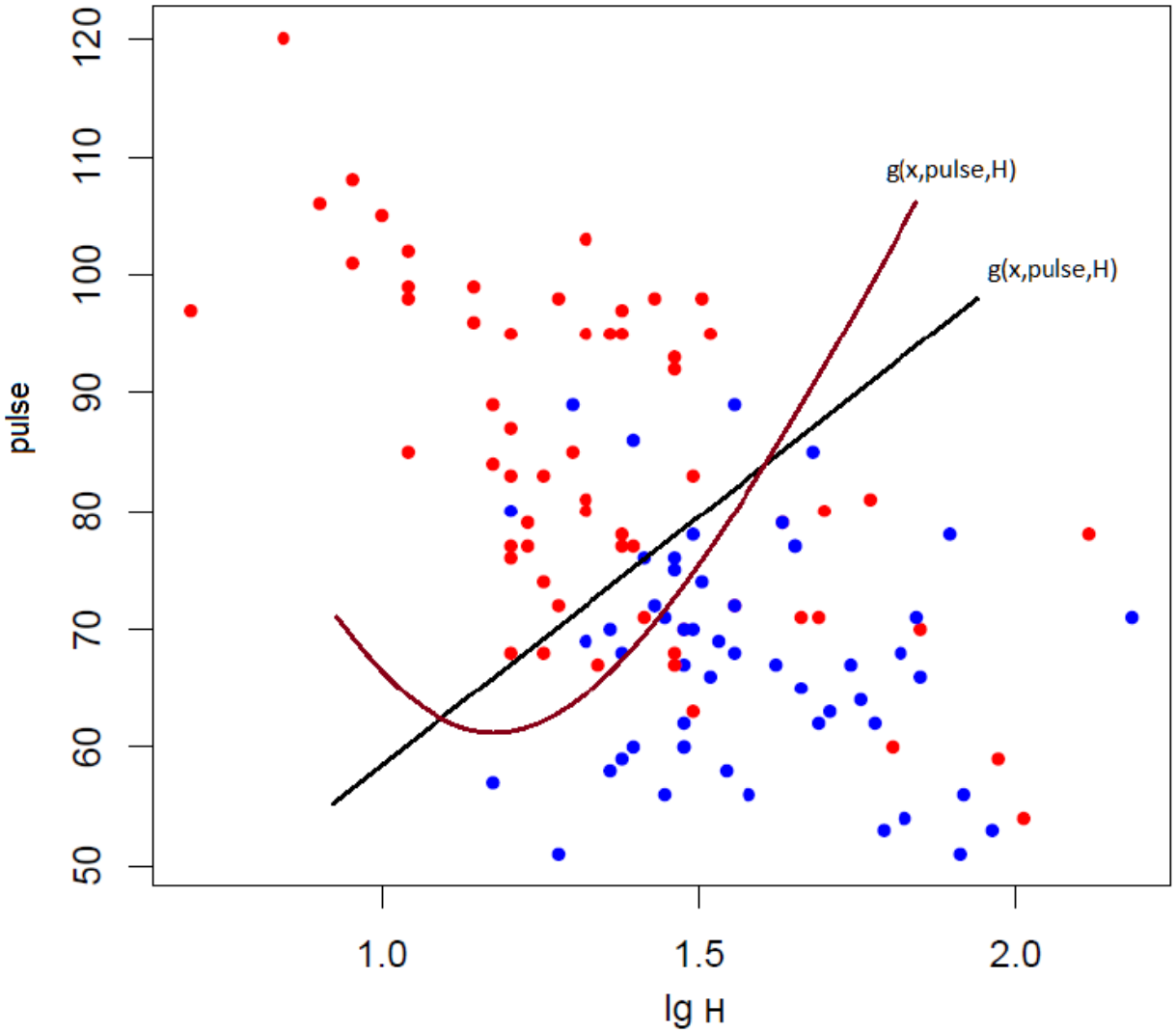
Модель алгоритма

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

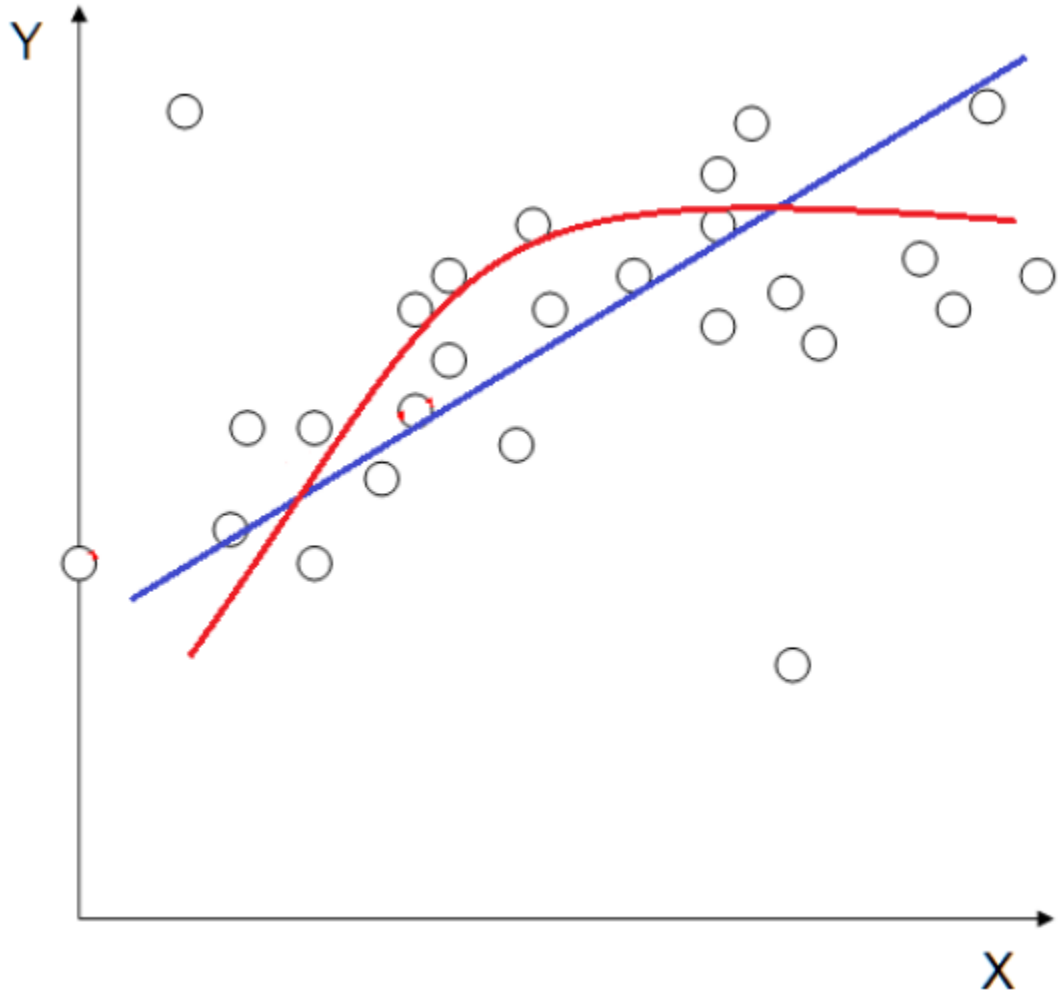
Модель алгоритма

- **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

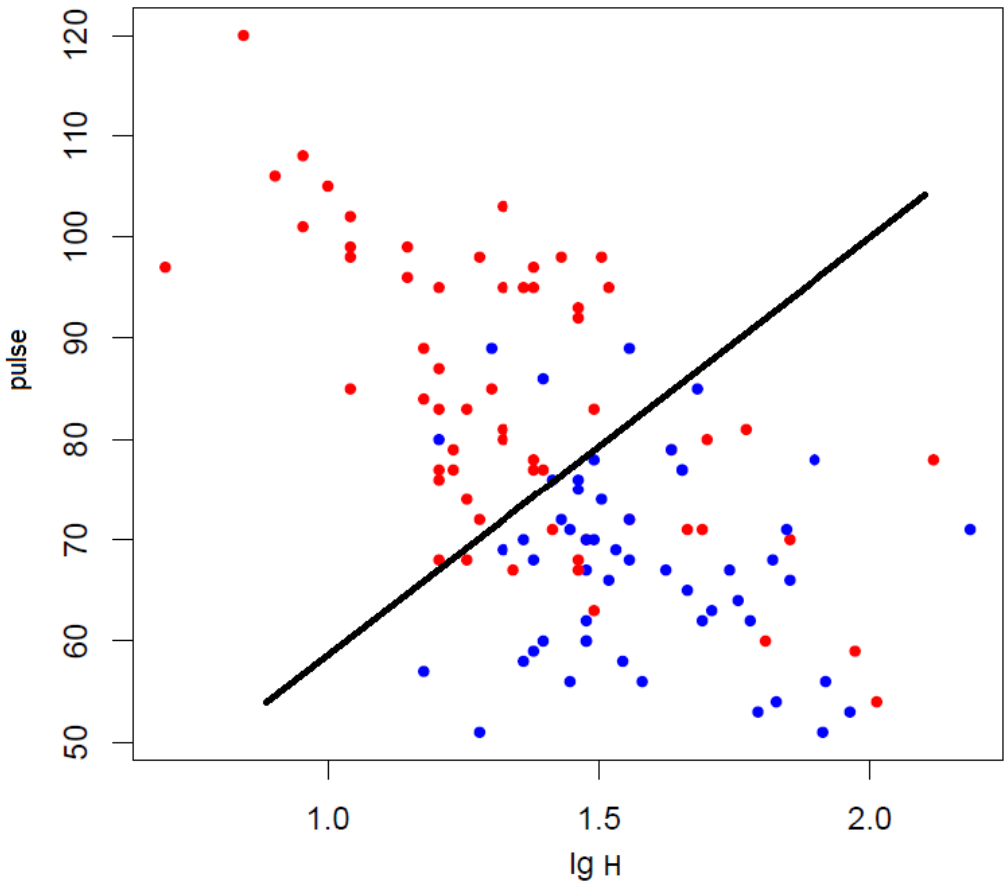
$$Q(a, X) = \int (\mathcal{L}(a, x) \cdot P(X, y))\, dXdy$$

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

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

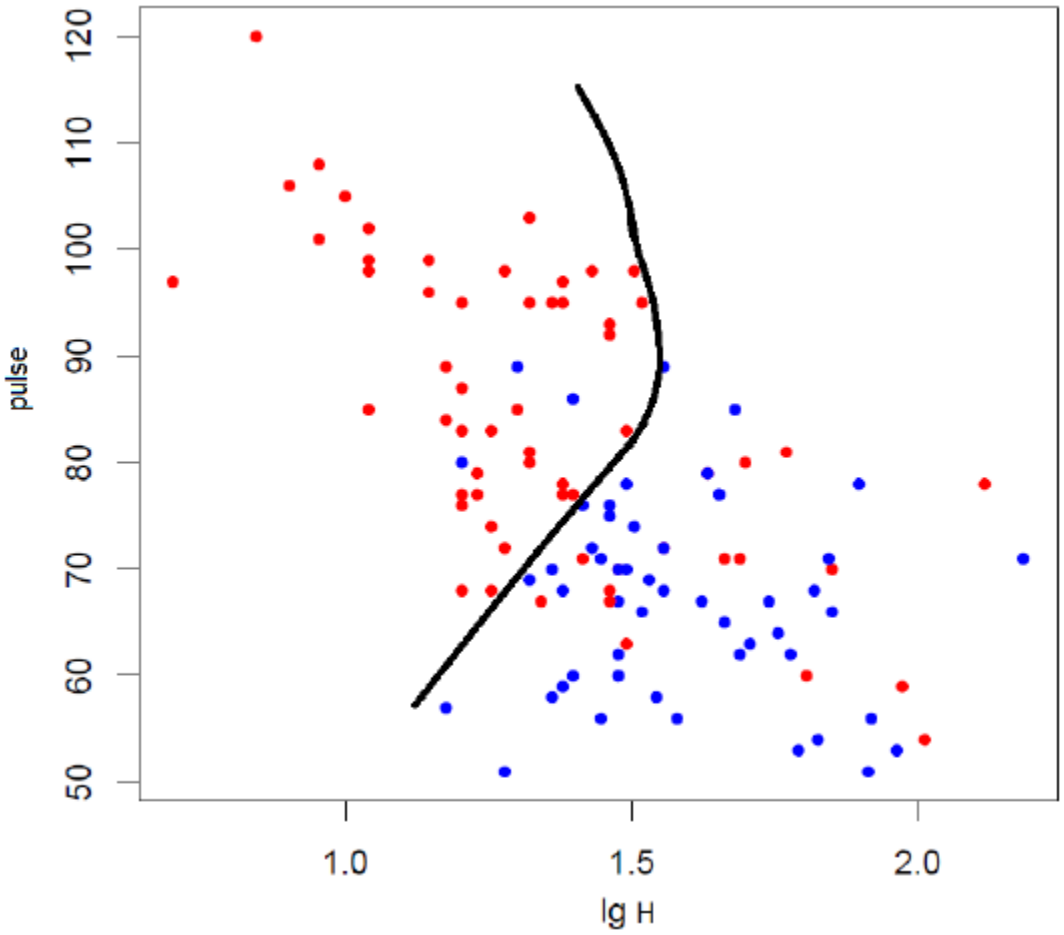
- Плотность $P(X, y)$ в функционале риска заменена на эмпирическое распределение (равномерное распределение) на элементах обучающей выборки.
- **Задача выбора "наилучшего" метода обучения** - это минимизация функционала риска по множеству A или по множеству параметров Θ.

Эмпирический функционал качества



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

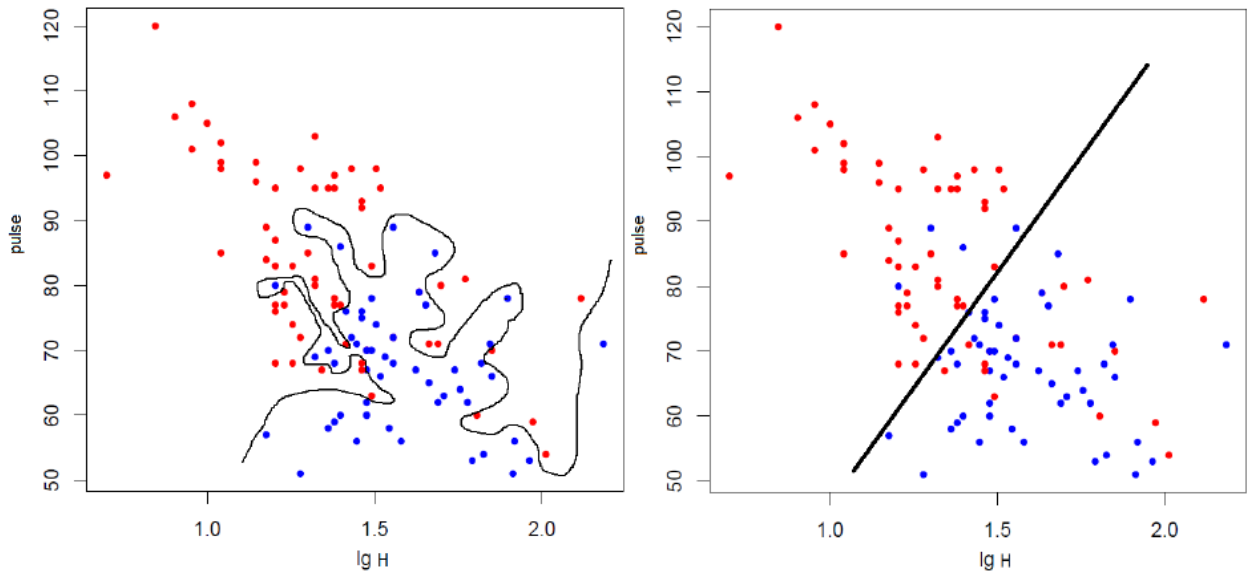
Эмпирический функционал качества



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

Оценка качества обучения

Переобучение и недообучение в классификации

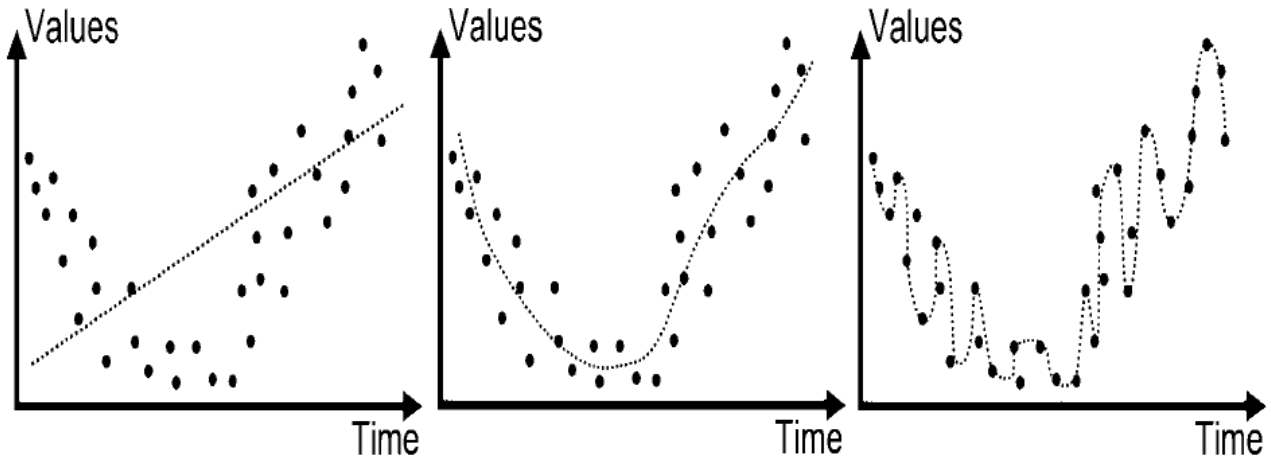


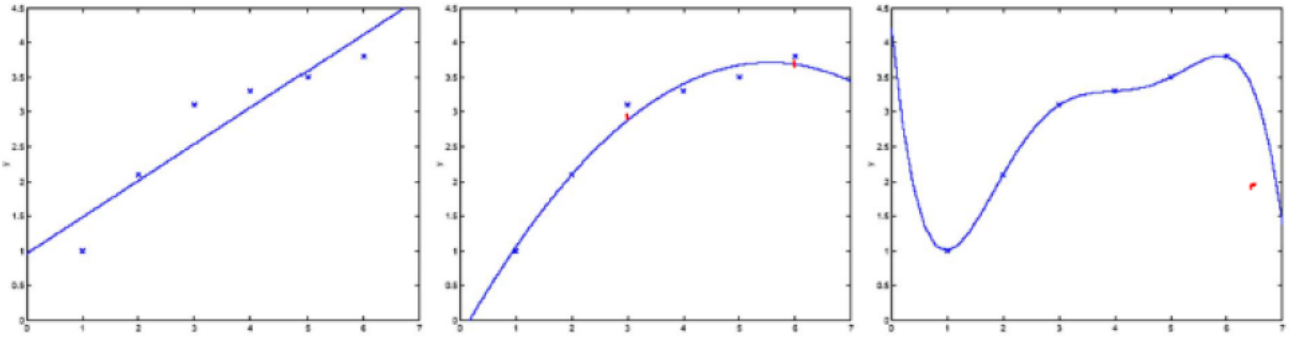
Проблема переобучения и недообучения

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 прецедентам, 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





$y = \theta_0 + \theta_1 x$

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

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

Этапы решения задачи обучения

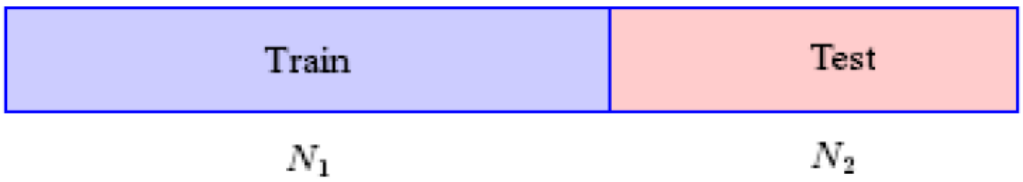
В задачах обучения по прецедентам всегда есть два этапа:

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.

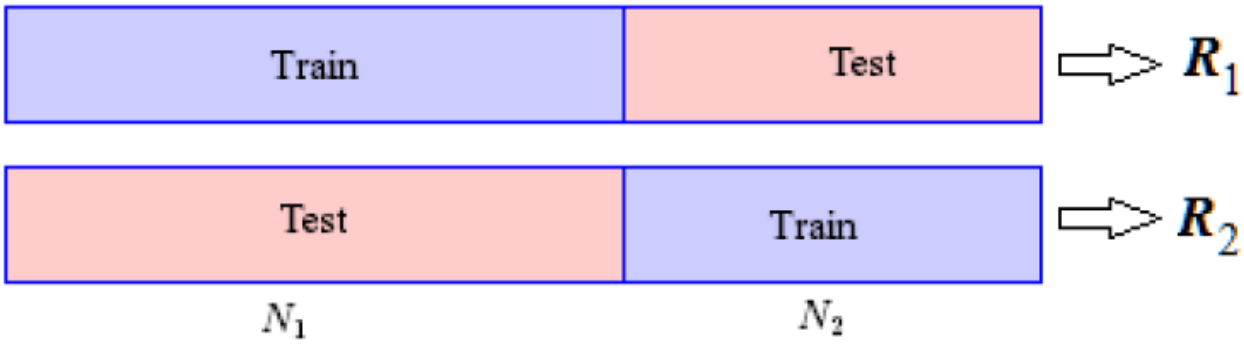
Обучающая и тестовая выборки

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



Метод перекрестного (скользящего) контроля

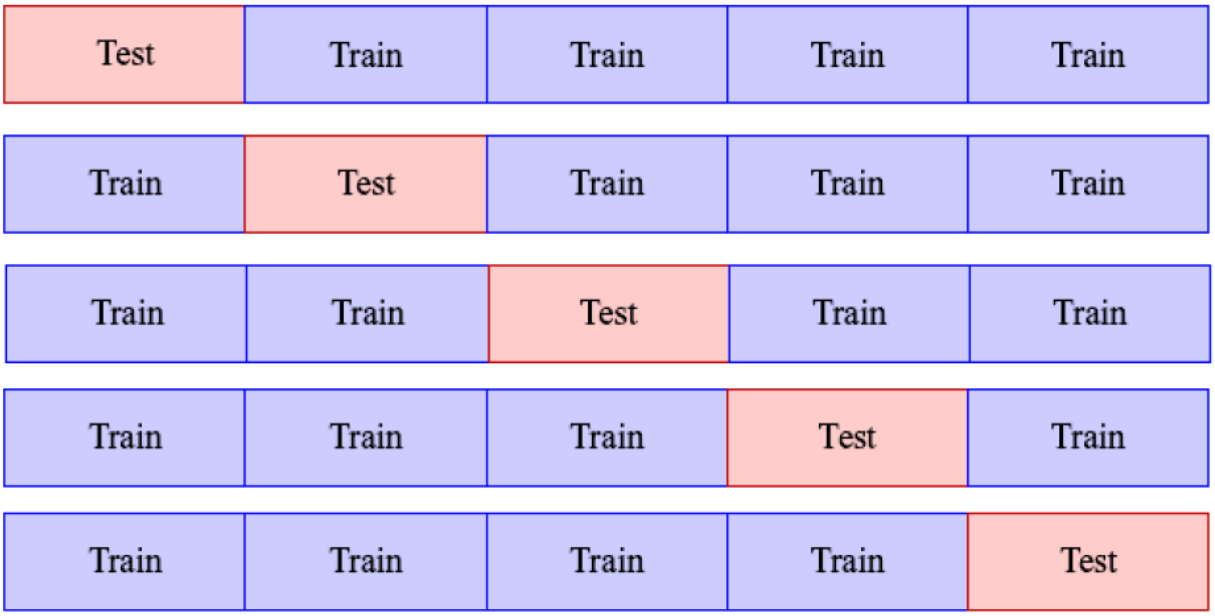


- Модель обучим на обучающей (train) выборке, а оценку ошибки R произведем на тестовой (test) выборке. Получим оценку R_1.
- Поменяем выборки ролями. Получим оценку R_2.
- Итоговая оценка качества - среднее взвешенное оценок R_1 и R_2.

Обобщение этой процедуры называется методом ***перекрестного (скользящего) контроля*** *(cross-validation).*

Метод перекрестного контроля в общем виде

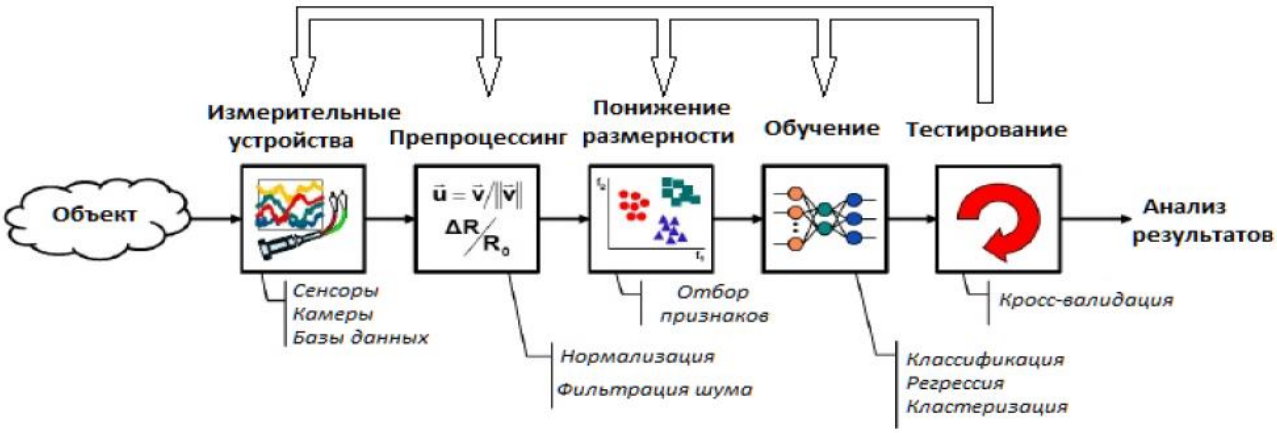
1. Случайным образом разобьем исходную выборку на M непересекающихся примерно равных по размеру частей.
2. Последовательно каждую из этих частей рассмотрим в качестве тестовой выборки, а объединение остальных частей — в качестве обучающей выборки.
3. Таким образом построим M моделей и соответственно M оценок для ошибки предсказания.
4. В качестве окончательной оценки ошибки возьмем их среднее взвешенное.



Частный случай - один отделяемый элемент

- $M = N$ - метод перекрестного контроля с **одним отделяемым элементом** или число частей равно числу элементов выборки
- (**leave-one-out cross-validation, LOO**)
- LOO — самый точный, но требует много времени

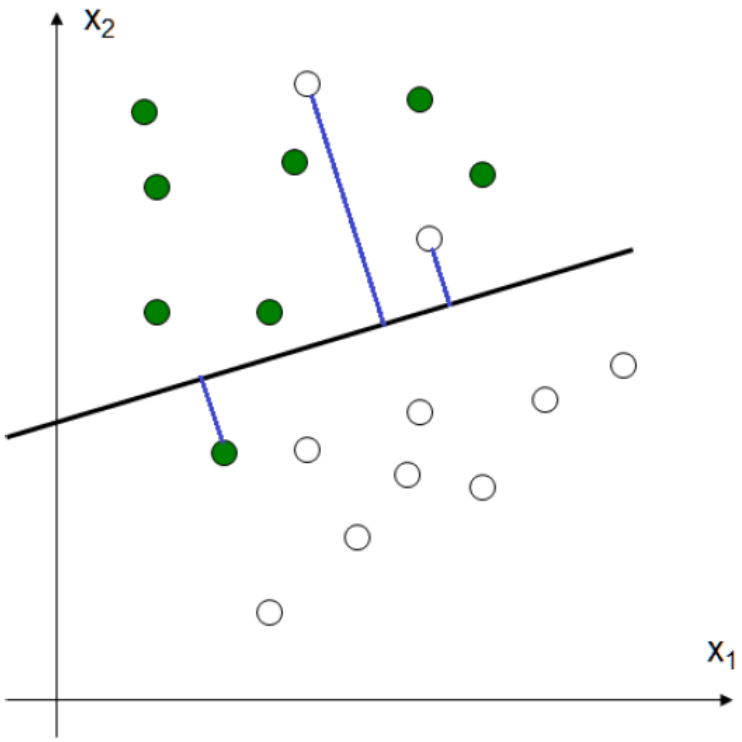
Этапы задачи обучения



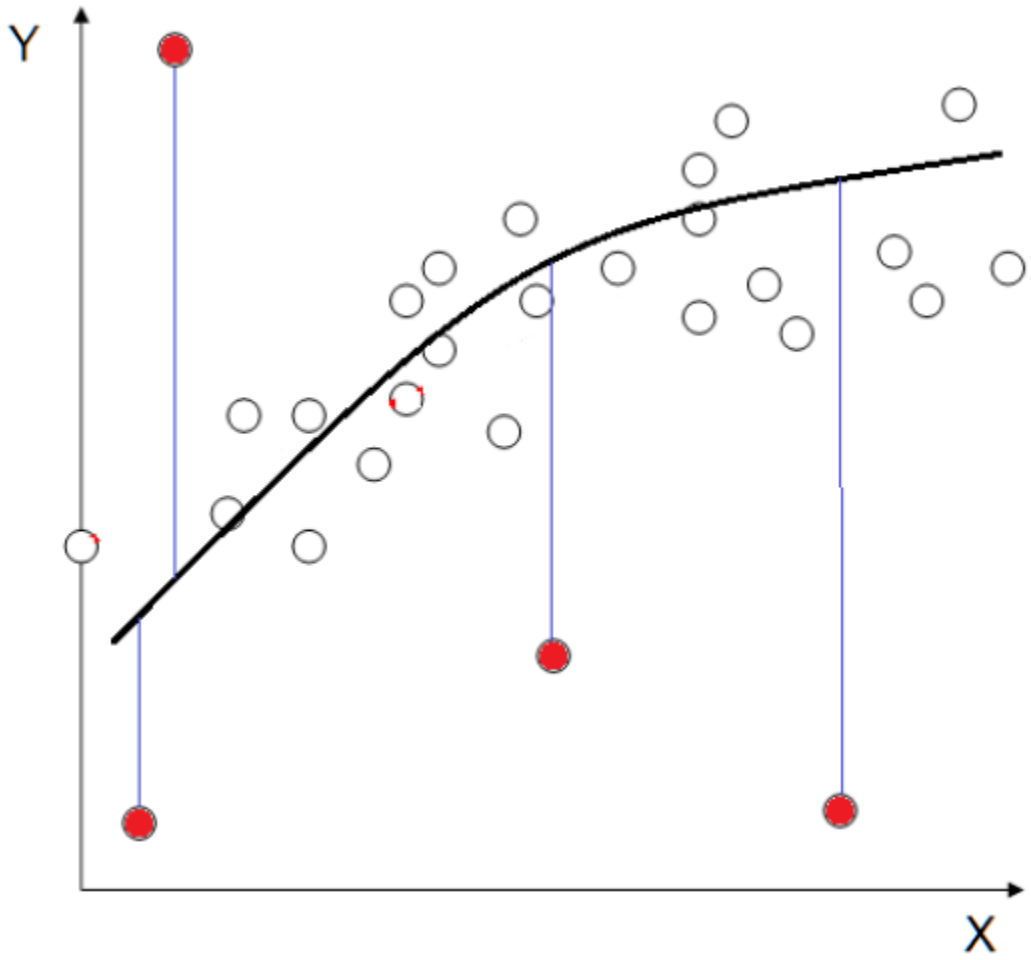
Препроцессинг

1. Задача **фильтрации выбросов** (outliers detection) — обнаружение в обучающей выборке небольшого числа нетипичных объектов.
   - В некоторых приложениях их поиск является самоцелью (например, обнаружение мошенничества).
   - Следствие ошибок в данных или неточности модели, то есть шум.
   - Используются робастные методы и одноклассовая классификация.
2. Задача заполнения **пропущенных значений** (missing values) — замена недостающих значений признаков их прогнозными значениями.

Фильтрации выбросов в классификации



Фильтрация выбросов в регрессии



Задача фильтрации выбросов

- Пусть $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ - обучающая выборка $\mathcal{Y} = \{1, 2, ..., c\}$ - множество классов
- **Отступ**: $M(x_i, y_i) = g_{y_i}(x_i) - \max_{y \in \mathcal{Y} \setminus \{y_i\}} g_y(x_i)$
  - отступ отрицательный означает, что объект x_i был неправильно классифицирован
  - величина отступа показывает, насколько классификатор уверен, что объект x_i может быть отнесен к истинному классу y_i

Удаление шумов

Шумы - это объекты, сильно выбивающиеся из закономерности, определяемой алгоритмом обучения, т.е. их можно определить как

$$\{x_i : M(x_i, y_i) < -\delta\}$$

для достаточно большого $\delta > 0$.

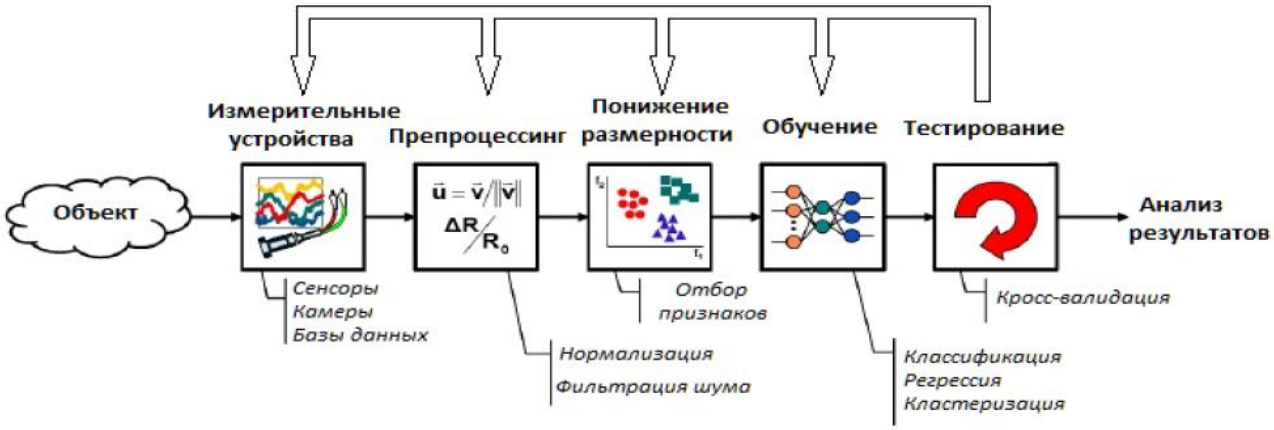
Алгоритм фильтрации шумов

1. для каждого (x_i, y_i) в обучающей выборке T вычислить $M(x_i, y_i)$
2. вернуть отфильтрованную обучающую выборку $T^* = \{(x_i, y_i) : M(x_i, y_i) \geq -\delta\}$

Задача заполнения пропущенных значений

	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 $(y = 1)$
...	...	...	...	...	...
x_{100}	108	193	32	длинные	ж $(y = 1)$

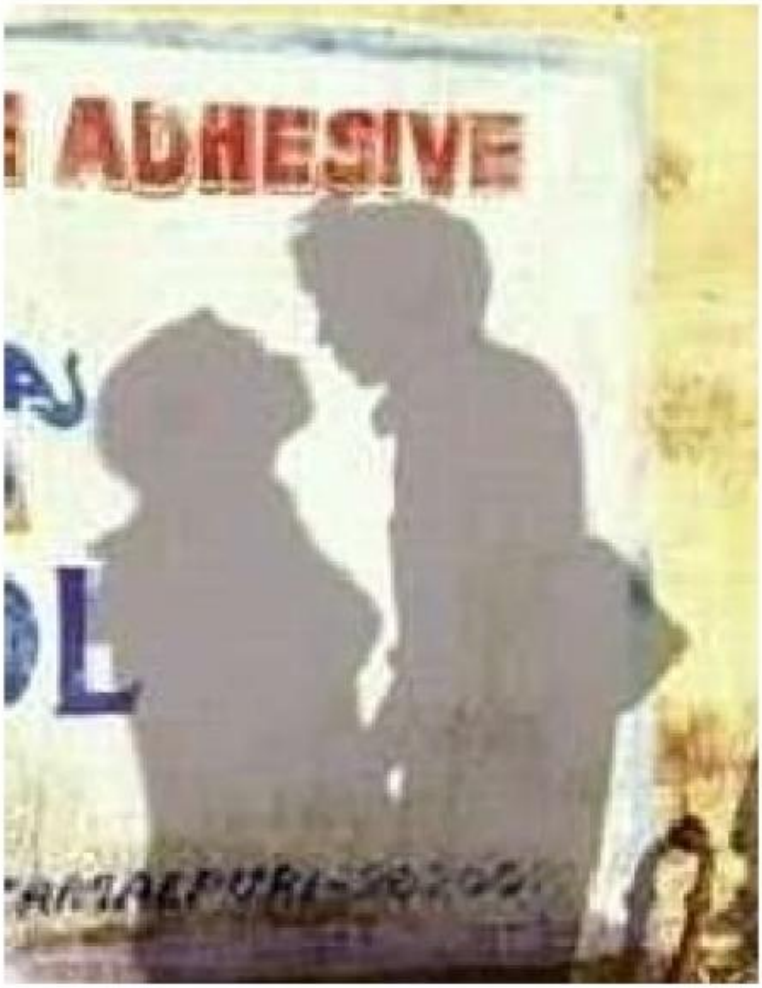
Схема всего процесса машинного обучения



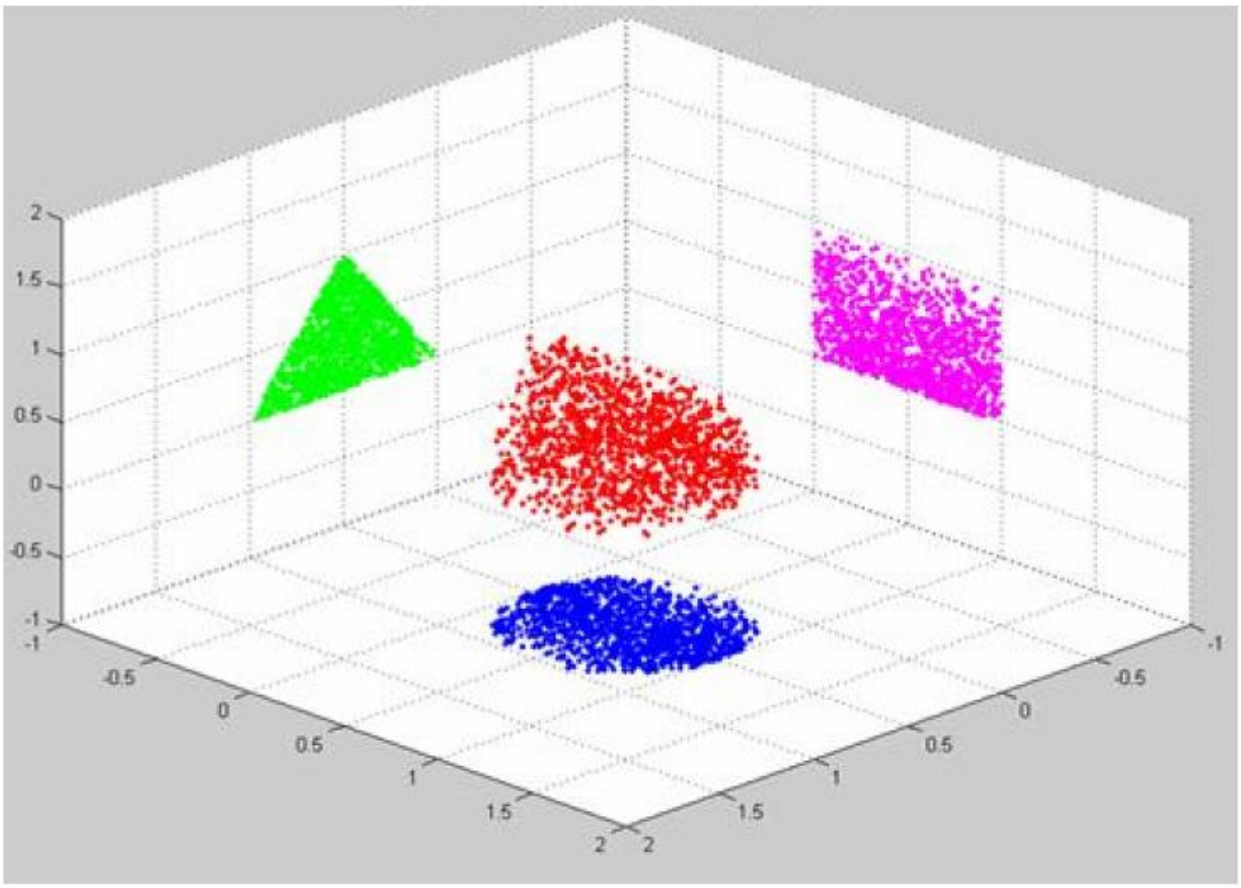
Отбор признаков и сокращение размерности (1)

- Задача **сокращения размерности** (dimensionality reduction) заключается в том, чтобы по исходным признакам с помощью некоторых функций преобразования перейти к наименьшему числу новых признаков, не потеряв при этом никакой существенной информации об объектах выборки.
- В классе линейных преобразований наиболее известным примером является **метод главных компонент**.

Отбор признаков и сокращение размерности (2)



Отбор признаков и сокращение размерности (3)



Примеры прикладных задач

Задачи медицинской диагностики

Объект - пациент в определенный момент времени.
Классы: диагноз или способ лечения или исход заболевания.
Примеры признаков:

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

Особенности задачи:

- обычно много "пропусков" в данных;
- нужен интерпретируемый алгоритм классификации;
- нужна оценка вероятности (риска | успеха | исхода).

Имеются данные о 114 лицах с заболеванием сердца: у 61 — проблемы, у 53 — нет проблем.
Для каждого пациента известны показатели:

- *pulse* — пульс,
- H — содержание гемоглобина в крови.

Можно ли научиться предсказывать (допуская небольшие ошибки) наличие проблем по *pulse* и H у новых пациентов?

Задача прогнозирования стоимости недвижимости

Объект - квартира в Санкт-Петербурге.
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 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.

Задача ранжирования поисковой выдачи

Объект - пара <запрос, документ>.
Классы: релевантен или не релевантен (разметка делается людьми асессорами).
Примеры признаков:

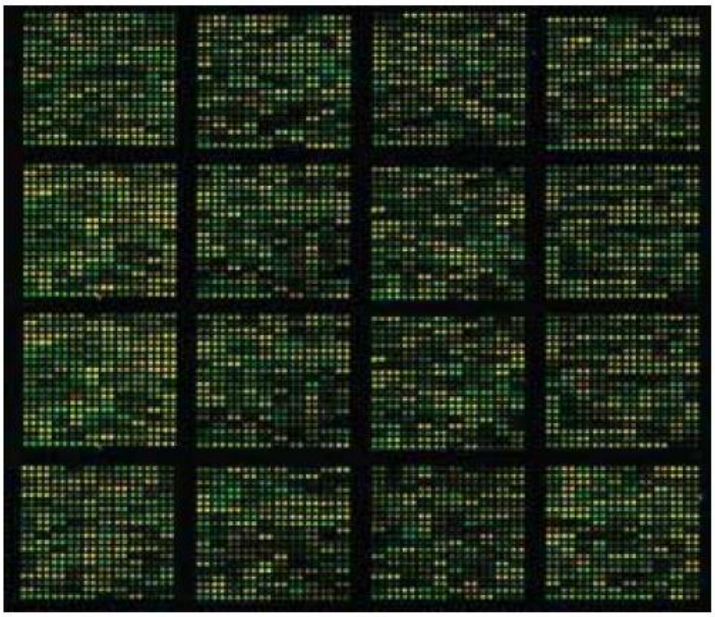
- *количественные*: частота слов запроса в документе, число ссылок на документ, число кликов на документ: всего, по данному запросу, и т. д.

Особенности задачи:

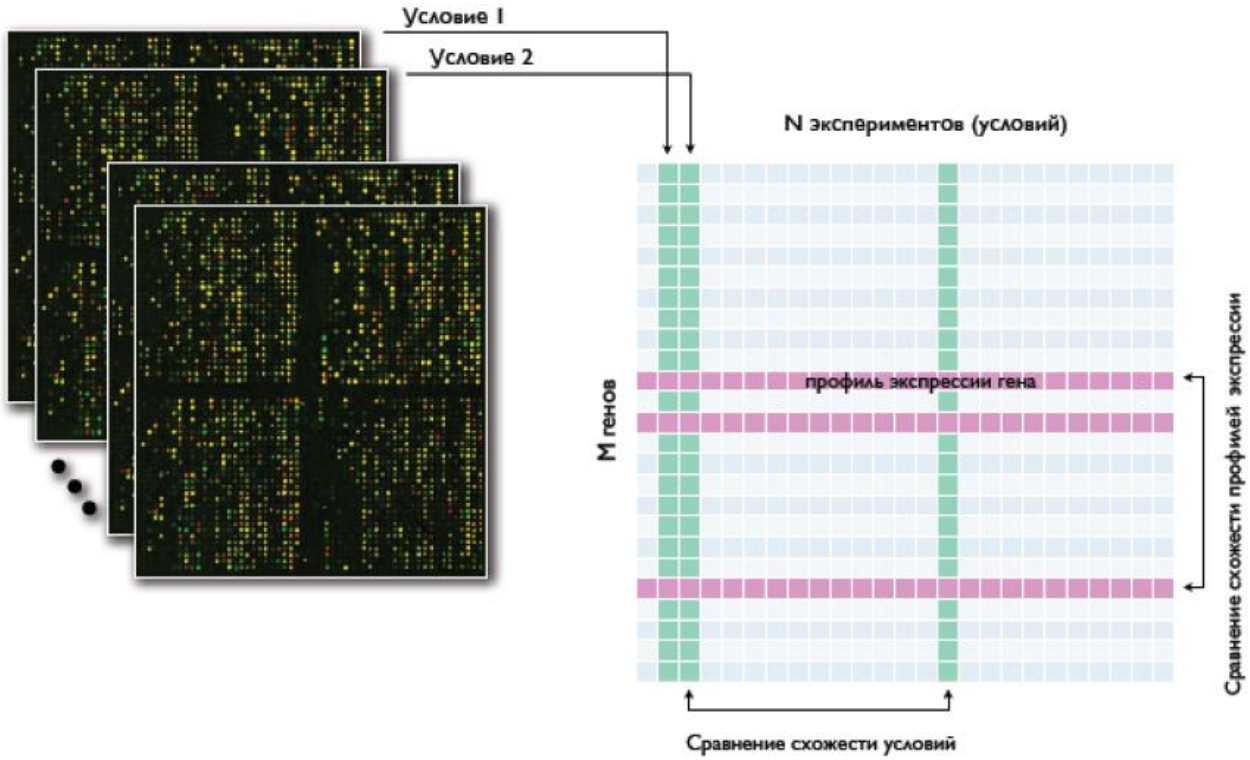
- оптимизируется не число ошибок, а качество ранжирования;
- сверхбольшие выборки;
- проблема конструирования признаков по сырым данным.

Анализ данных по экспрессии генов

ДНК-микрочипы - двумерный массив ДНК-зондов для тысяч нуклеотидных последовательностей, позволяющий измерять экспрессию генов при разных условиях



Анализ данных по экспрессии генов



Анализ данных по экспрессии генов

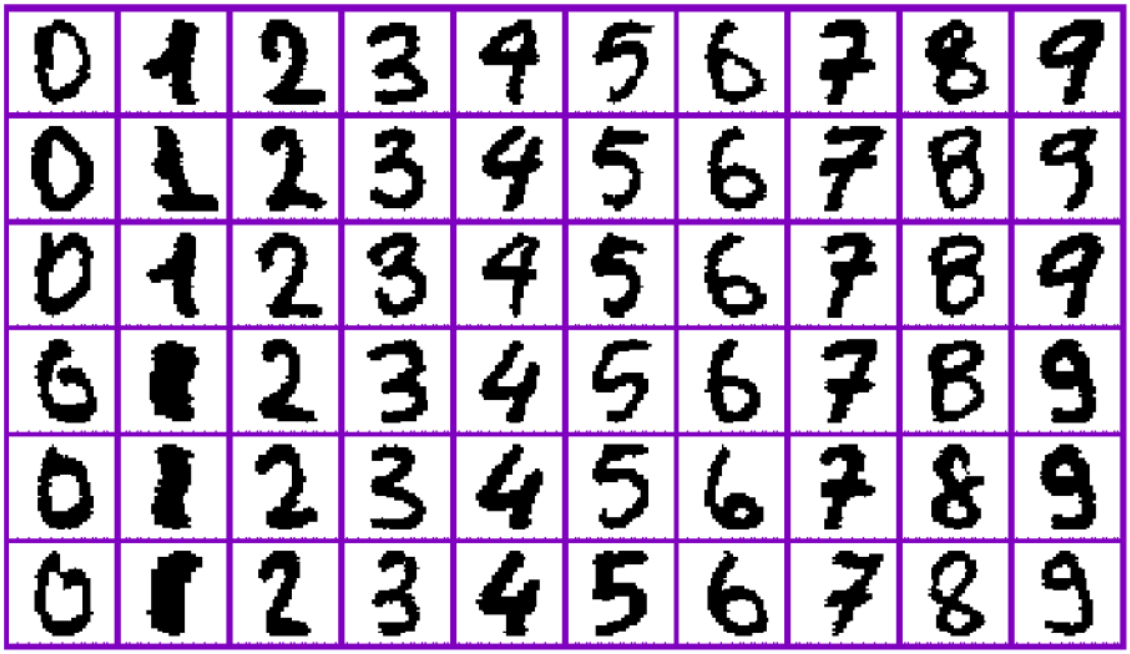
- **Кластеризация**: Группы генов выполняющие схожие функции имеют схожие профили экспрессии.
  - Задача: Поиск функциональных групп генов.
- **Классификация**: Клетка может находится в разных состояних (здоровая/раковая), различающихся уровнями экспрессии генов.
  - Задача: Определение состояния клетки на основе данных о профилях экспрессии генов.

Распознавание рукописных символов (цифр)

Объект - рукописный символ (цифра).
Классы: 0,1,...,9
Примеры признаков:

- *бинарные*: код (признаковое описание) - битовая матрица размера 32×32.

1 — пиксел черный, 0 — пиксел белый.



Страховая компания (кластеризация)

- Информация об автомобилях и их владельцах: марка автомобиля; стоимость автомобиля; возраст водителя; стаж водителя; возраст автомобиля
- Цель - разбиение автомобилей и их владельцев на классы, каждый из которых соответствует определенной рисковой группе.
- Наблюдения, попавшие в одну группу, характеризуются одинаковой вероятностью наступления страхового случая, которая впоследствии оценивается страховщиком.

ЗАДАЧИ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Машинный перевод

Распознавание речи

Распознавание эмоций

Резюме

- **Основные понятия машинного обучения**:
  - объект, ответ, метка 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.

Резюме - О курсе

Различные алгоритмы и подходы к решению задач машинного обучения:

- Линейная регрессия
- Метод ближайших соседей
- Байесовский подход
- Машина опорных векторов
- Нейронные сети
- Деревья решений
- Бустинг (AdaBoost, Random Forest) и бэггинг
- Обучение без учителя, кластеризация

Ресурсы

- Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2015