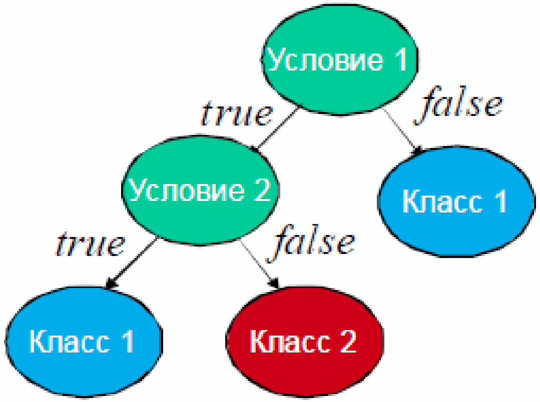
Введение в машинное обучение

Техническое зрение

Проблема распознавания

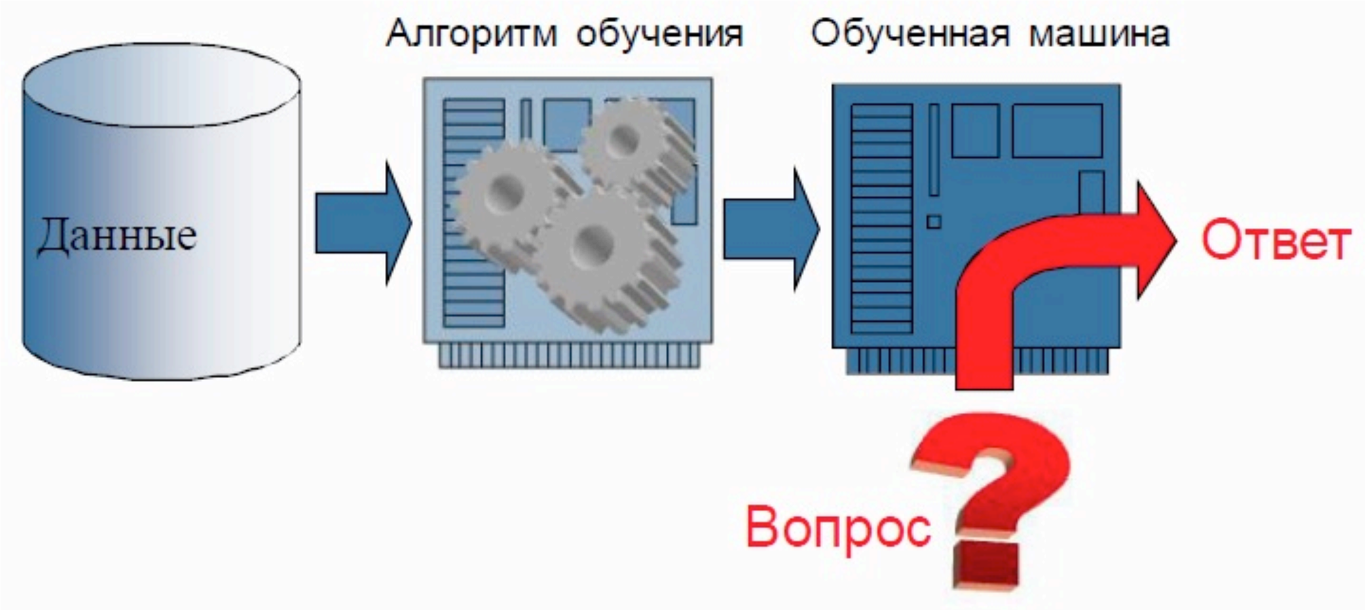
Проблемы распознавания

- **Основная проблема**: правила необходимо подбирать вручную.



- **Следствия:**
  1. Нужно использовать осмысленные и информативные признаки.
  2. Таких признаков крайне мало и их сложные комбинации невозможно обработать.

Идеальное решение

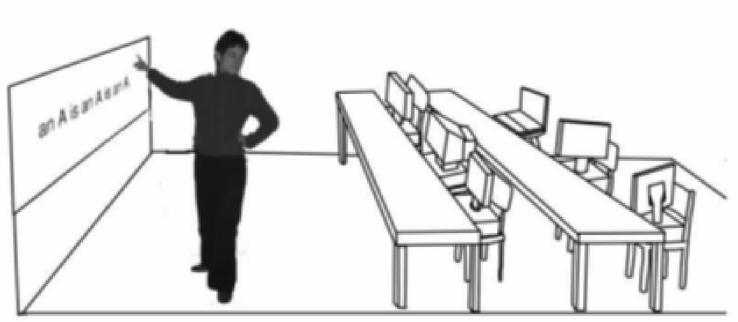


- Машина дает ответы на вопросы на основании уже обработанных данных.

Теория машинного обучения

Процесс обучения

- Обучение не равносильно заучиванию, заучивание для машины не проблема.
- Машина должна научиться делать выводы по набору обучающих данных.
- Машина должна корректно работать на основе новых данных, которые ей раньше не давали.

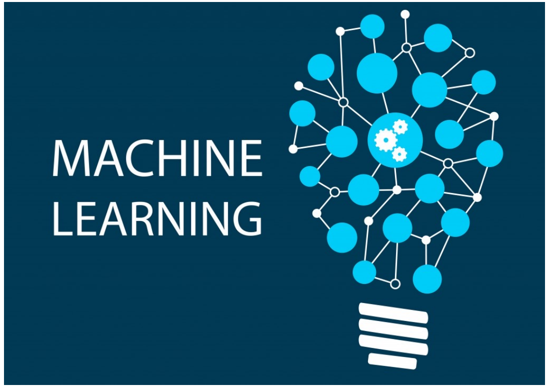


Определение

- «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» © Т.М. Митчелл, 1997.

Сферы применения

- компьютерное зрение,
- распознавание речи,
- компьютерная лингвистика и обработка естественных языков,
- медицинская диагностика,
- биоинформатика,
- техническая диагностика,
- финансовые приложения,
- поиск и рубрикация текстов,
- интеллектуальные игры,
- экспертные системы и др.



Классы машинного обучения

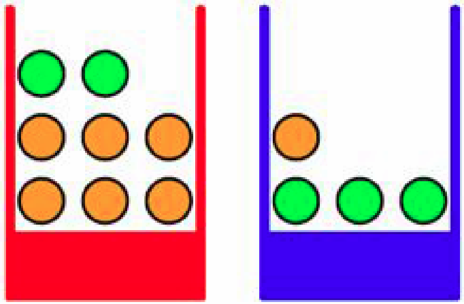
1. **Дедуктивное обучение** (от общего к частному).
   - Имеются формализованные данные. Требуется на основе них вывести правило, применимое к конкретному случаю.
   - Типовой пример: *экспертные системы.*
2. **Индуктивное обучение** (от частного к общему).
   - Имеются эмпирические данные. Требуется восстановить некоторую зависимость. Подразделяется на:
     a) Обучение с учителем;
     b) Обучение без учителя;
     c) Обучение с подкреплением (reinforcement learning);
     d) Активное обучение и пр.

Теория вероятностей и случайных процессов

- Что такое вероятность?
- В частотной интерпретации, вероятность – это *частота повторяемого события*.
- В Байесовской интерпретации вероятность – это *мера неопределенности исхода эксперимента*.

Пример

- **Опыт по извлечению фруктов из двух коробок**
  - **Эксперимент:**
    - Выбор коробки;
    - Извлечение фрукта;
    - Помещение фрукта назад.
  - **Две случайные величины:**
    - X – цвет коробки (красная или синяя);
    - Y – фрукт (апельсин или яблоко).

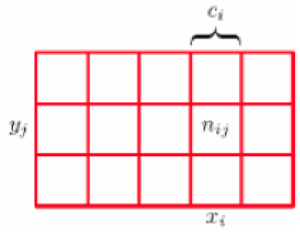


$$P(X = \text{красная}) = \frac{\text{Сколько раз выбрали красную коробку}}{\text{Сколько провели экспериментов}}$$

P – вероятность выбора красной коробки.

Пример

- Будем проводить эксперимент и заносить число исходов в таблицу (по горизонтали цвета коробки, по вертикали – фрукты).



- ***Вероятность пересечения событий***:

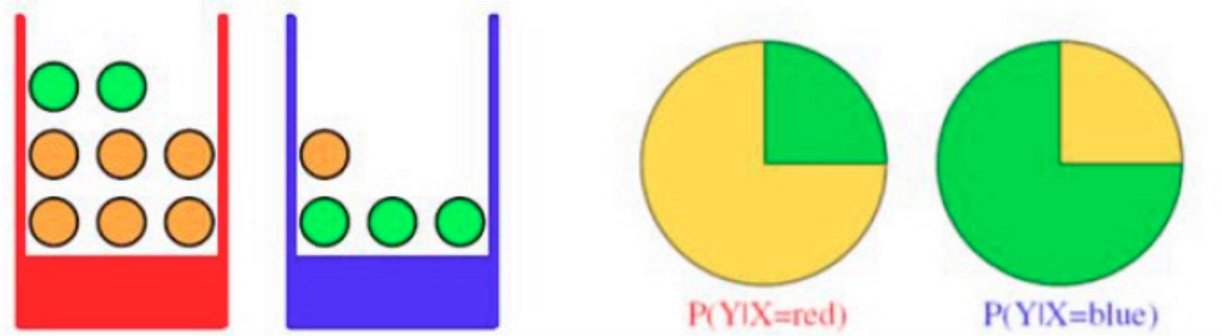
$$P\left(X = x_i, P = y_j\right) = \frac{n_{ij}}{N},$$

где N – число экспериментов, количество исходов n_{ij}.

- ***Условная вероятность:***

$$P\left(P = y_j | X = x_i,\right) = \frac{n_{ij}}{c_i}.$$

Пример



Условная вероятность:
75%, что апельсин в красной коробке,
25%, что в синей.

Формула Байеса

ИТМО

- Какова вероятность, что перед нами динозавр (x – наблюдение)?

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)} \text{ – Формула Байеса,}$$

где $P(x|y)$ – вероятность того, что динозавр выглядит именно так;

$P(y)$ – вероятность встретить динозавра;

$P(x)$ – вероятность увидеть такую сцену.

Теория вероятностей

- Правило суммы:

$$P(x) = \int_y P(x, y)dy \leftrightarrow P(y) = \int_x P(x, y)dx$$

- Правило произведения:

$$P(x, y) = P(y|x)P(x) = P(x|y)P(y)$$

- Если две случайные величины независимы:

$$P(x, y) = P(x)P(y)$$

Задачи машинного обучения

- Постановка задачи обучения с учителем:
  - $\mathbb{X}$ – множество *объектов* или *примеров*, *ситуаций*, *входов* (*samples*);
  - $\mathbb{Y}$ – множество *ответов* или *откликов*, *меток*, *выходов* (*responses*).
  - Имеется некоторая зависимость, позволяющая по $x \in \mathbb{X}$ предсказать $y \in \mathbb{Y}$.
  - Если зависимость *детерминированная*, то существует функция $f^*: \mathbb{X} \to \mathbb{Y}$.
  - Зависимость известна только на объектах *обучающей выборки* – некоторого конечного числа данных:

$$\left\{\left(x^{(i)}, y^{(i)}\right): x^{(i)} \in \mathbb{X}, y^{(i)} \in \mathbb{Y} (i = 1, \dots, N)\right\}$$

Задачи машинного обучения

- Упорядоченная пара «объект-ответ» $\left(x^{(i)}, y^{(i)}\right) \in (\mathbb{X} \times \mathbb{Y})$ называется *прецедентом*.
- **Задача обучения с учителем**: восстановление зависимости между входом и выходом по имеющейся обучающей выборке, т.е. необходимо построить функцию (*решающее правило*) $f: \mathbb{X} \rightarrow \mathbb{Y}$, по новым объектам $x \in \mathbb{X}$ предсказывающую ответ $f(x) \in \mathbb{Y}$:
$$y = f(x) \approx f^*(x).$$

Основные определения

- Функции f выбираются из параметрического семейства F, т.е. **из некоторого множества возможных моделей**.
- **Процесс нахождения функции** f называется *обучением (learning)*, а также *настройкой* или *подгонкой (fitting)* модели.
- **Алгоритм построения функции** f по заданной обучающей выборке называется *алгоритмом обучения*.
- **Некоторый класс алгоритмов** называется *методом обучения*.

Основные определения

- Алгоритмы обучения оперируют **описаниями объектов**: каждый элемент выборки описывается набором признаков $x = (x_1, x_2, \dots, x_d)$ (**вектор-признак**),

где $x_j \in Q_j, j = \overline{1, d}, \mathbb{X} = Q_1 \times Q_2 \times \cdots Q_d$.

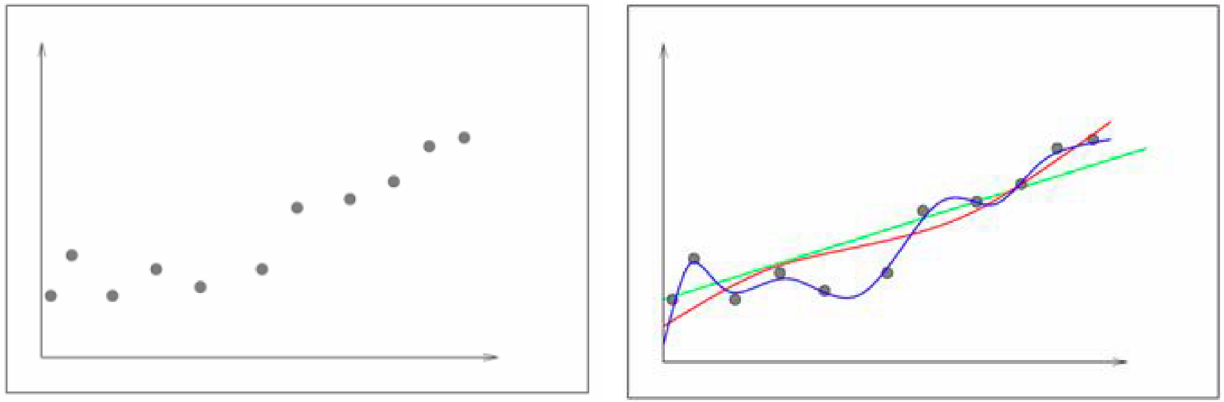
- Множество $\mathbb{X}$ называется *пространством признаков*.
- Необходимо сконструировать такую функцию $y = f(x)$ от вектора признаков $x = (x_1, x_2, \dots, x_d)$, которая бы выдавала ответ y для любого возможного наблюдения x.
- Компонента x_j называется j-м *признаком*, или *свойством* (*feature*), или *атрибутом объекта* x.

Основные определения

- Если $Q_j = \mathbb{R}$, то j-й признак называется *количественным* или *вещественным*.
- Если Q_j **конечно**, то j-й признак называется *номинальным*, или *категориальным*, или *фактором*.
  - Если $\dim Q_j = 2$, то признак называется *бинарным*.
  - Если Q_j упорядочено, то признак называется *порядковым*.

Задача восстановления регрессии

- Если $\mathbb{Y} = \mathbb{R}$, то это задача *восстановления регрессии*.
  - Решающее правило f называют *регрессией*.
- Если $\mathbb{Y}$ конечно ($\mathbb{Y} = \{1,2,\dots,K\}$), то это задача *классификации*.
  - Решающее правило f называют *классификатором*.



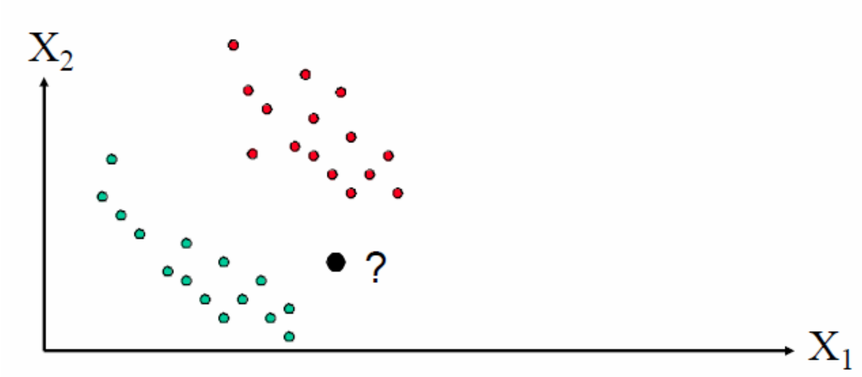
Пример восстановления регрессии, y – непрерывная величина.

Задача бинарной классификации

- Дана обучающая выборка:

$$X_m = \{(x_1, y_1), \dots, (x_m, y_m)\}, (x_i, y_i) \in R^m \times Y, Y = \{-1, +1\}.$$

- Объекты принадлежат одному из двух классов. Основной класс помечаем как «$+1$», второстепенный «фон» как «-1».
- Требуется для всех новых значений x определить класс «$+1$» или «-1».

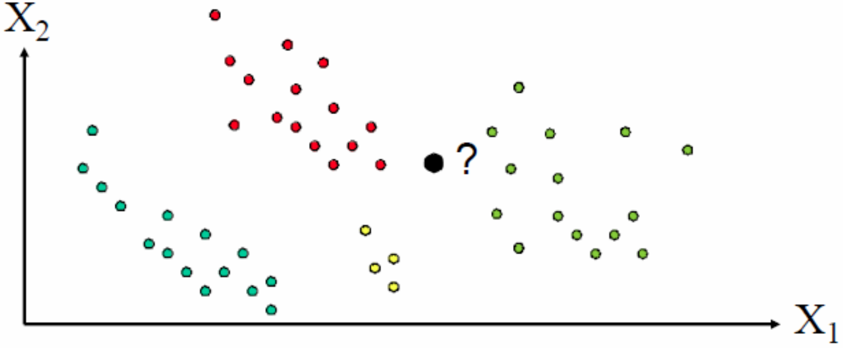


Задача многоклассовой классификации

- Дана обучающая выборка:

$$X_m = \{(x_1, y_1), \dots, (x_m, y_m)\}, (x_i, y_i) \in R^m \times Y, Y = \{1, \dots, K\}.$$

- Объекты принадлежат одному из K классов.
- Требуется для всех новых значений x определить класс и поставить метку от 1 до K.

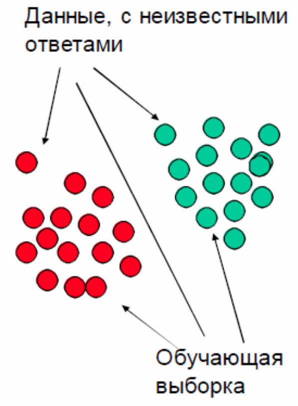


Замечания

1. Найденное решающее правило должно обладать *обобщающей способностью* (построенный классификатор или функция регрессии должны отражать общую зависимость выхода от входа, основываясь лишь на известных данных о прецедентах обучающей выборки).
2. Следует уделять внимание проблеме эффективной вычислимости функции f и к алгоритму обучения: настройка модели должна происходить за приемлемое время.

Задачи машинного обучения

- Интересно качество работы алгоритма на новых данных: необходимо связать имеющиеся данные с теми, которые будем обрабатывать в будущем.
- Для этого значения признаков будем считать случайными величинами.
- Будем считать, что данные, которые придется обрабатывать в будущем и имеющиеся данные, распределены одинаково.

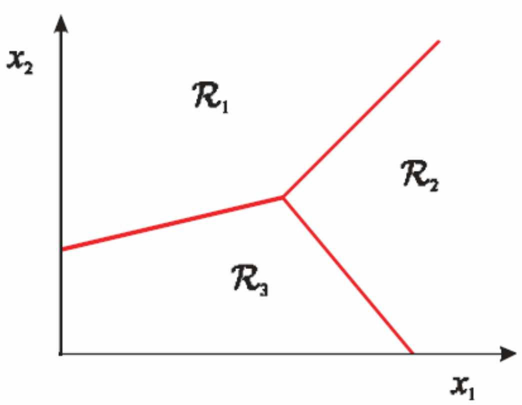


Парадигмы машинного обучения

1. **Воспроизводящий подход (generative approach)**
   a) Используем формулу Байеса $P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)}$;
   b) Моделируем каждый класс отдельно, оцениваем $P(x|y), P(y)$;
   c) Постановка задачи сходна классификации.
2. **Дискриминантный подход (discriminative approach)**
   a) Поскольку интересует $P(y|x)$, то её и будем оценивать;
   b) Постановка задачи сходна регрессии.

Парадигмы машинного обучения

2. **Дискриминантный подход**
   - Необходимо построить функцию $y = f(x)$ – *решающее правило* или *классификатор*.
   - Любое решающее правило делит пространство на *решающие регионы*, разделенные *решающими границами.*



Дискриминантный подход

- Будем выбирать функции f из параметрического семейства F, т.е. из некоторого множества возможных моделей.
- Введем некоторую *функцию потерь* (*функцию штрафа*) $L(y, f(x))$ от истинного значения выхода y и предсказанного $f(x)$:
  - В задаче восстановления регрессии *квадратичный штраф*:
$$L\big(y, f(x)\big) = \frac{1}{2}\,\big(y - f(x)\big)^2,$$
  - или *абсолютный штраф*:
$$L\big(y, f(x)\big) = |y - f(x)|.$$
  - В задаче классификации *ошибка предсказания*:
$$L\big(y, f(x)\big) = I\big(y \neq f(x)\big),$$

где $f(x)$ – предсказанный класс,

$I = \begin{cases} 1, \text{условие выполнено} \\ 0, \text{условие не выполнено} \end{cases}$ – индикаторная функция.

Риски

- Задача обучения состоит в том, чтобы найти набор параметров классификатора f, при котором потери для новых данных будут минимальны.
- Введем понятие *общий (средний) риск* – это математическое ожидание потерь:

$$R(f) = E\big(L(f(x), y)\big) = \int_{x,y} L(f(x), y) dP$$

- К сожалению, ввиду неизвестности распределения вероятности P совместной случайной величины (x, y) общий риск рассчитать невозможно.

Риски

- Введем понятие *эмпирический риск*. Пусть $X = \{x_1, \dots, x_m\}, Y = \{y_1, \dots, y_m\}$ – обучающая выборка. Эмпирический риск или *ошибка тренировки*:

$$R_{emp}(f, X) = \frac{1}{m}\sum_{i=1}^{m} L(y_i, f(x_i))$$

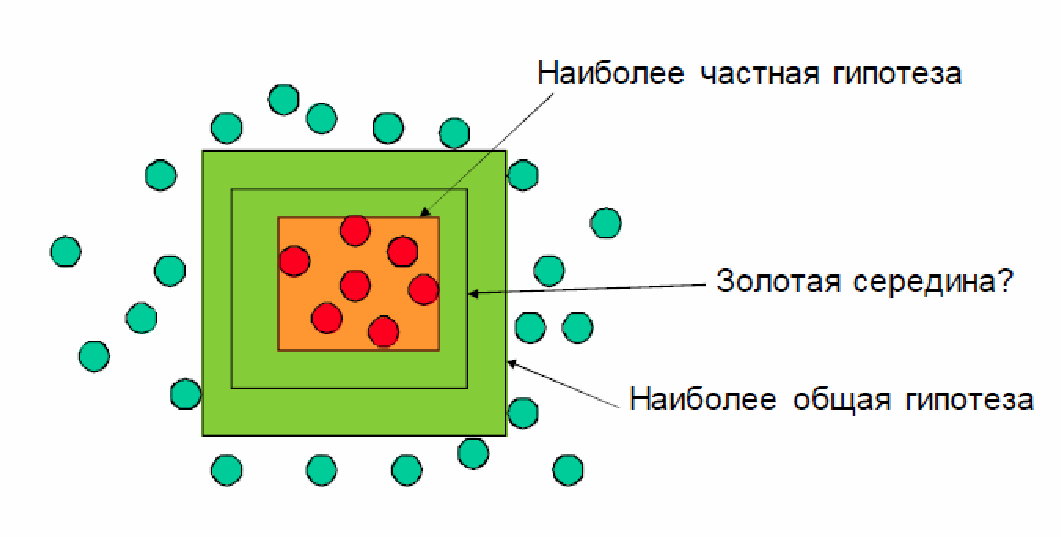
- Для минимизации эмпирического риска необходимо найти функцию f в соответствие с условием:

$$f = \arg\min_{f \in F} R_{emp}(f, X)$$

- Условие называется: *принцип минимизации эмпирического риска.*

Замечание

- Гипотез, имеющих нулевой эмпирический риск, может существовать неограниченное количество:



Задача обучения с учителем

- Задача свелась к отысканию функции f из допустимого множества F, удовлетворяющей условию:

$$f = \arg\min_{f \in F} R_{emp}(f, X),$$

F и L фиксированы и известны.

- Класс моделей F параметризован, т.е. есть имеется его описание вида $F = \{f(x) = f(x, \theta) : \theta \in \Theta\}$, где Θ – некоторое известное множество.
- Процесс настройки модели:
  - алгоритмом обучения выбираются значения набора параметров Θ, обеспечивающих выполнение условия f, т.е. минимизации ошибки на прецедентах обучающей выборки.

Переобучение

- Рассмотренное условие не подходит для оценки обобщающей способности алгоритма.
- Все имеющиеся данные разбивают на *обучающую* и *тестовую* выборки:
    - Обучение производится с использованием обучающей выборки,
    - Оценка качества предсказания на основе данных тестовой выборки.
- Значения $R(f)$ и $R_{emp}(f, X)$ могут различаться значительно.
- Явление, когда $R_{emp}(f, X)$ мало, а $R(f)$ чересчур велико, называется *переобучением*.

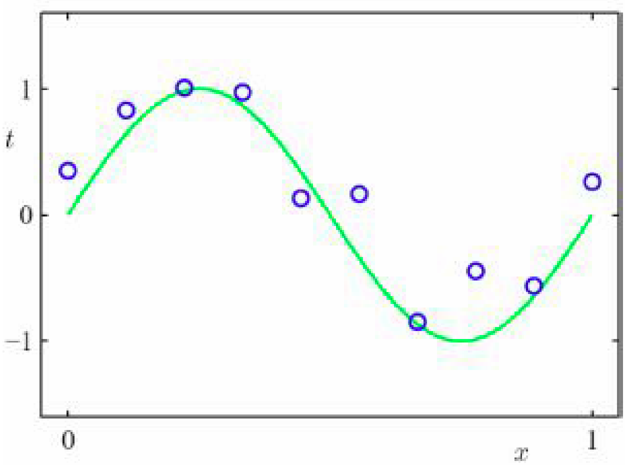
Переобучение

- Пусть имеется задача регрессии:

$$t = \sin(2\pi x) + \epsilon,$$

где ϵ – нормально распределенный шум, однако мы этого не знаем.

- Пусть имеется обучающая выборка и требуется восстановить зависимость:



Переобучение

- Будем выбирать целевую зависимость среди полиномов порядка M (параметризованного множества):

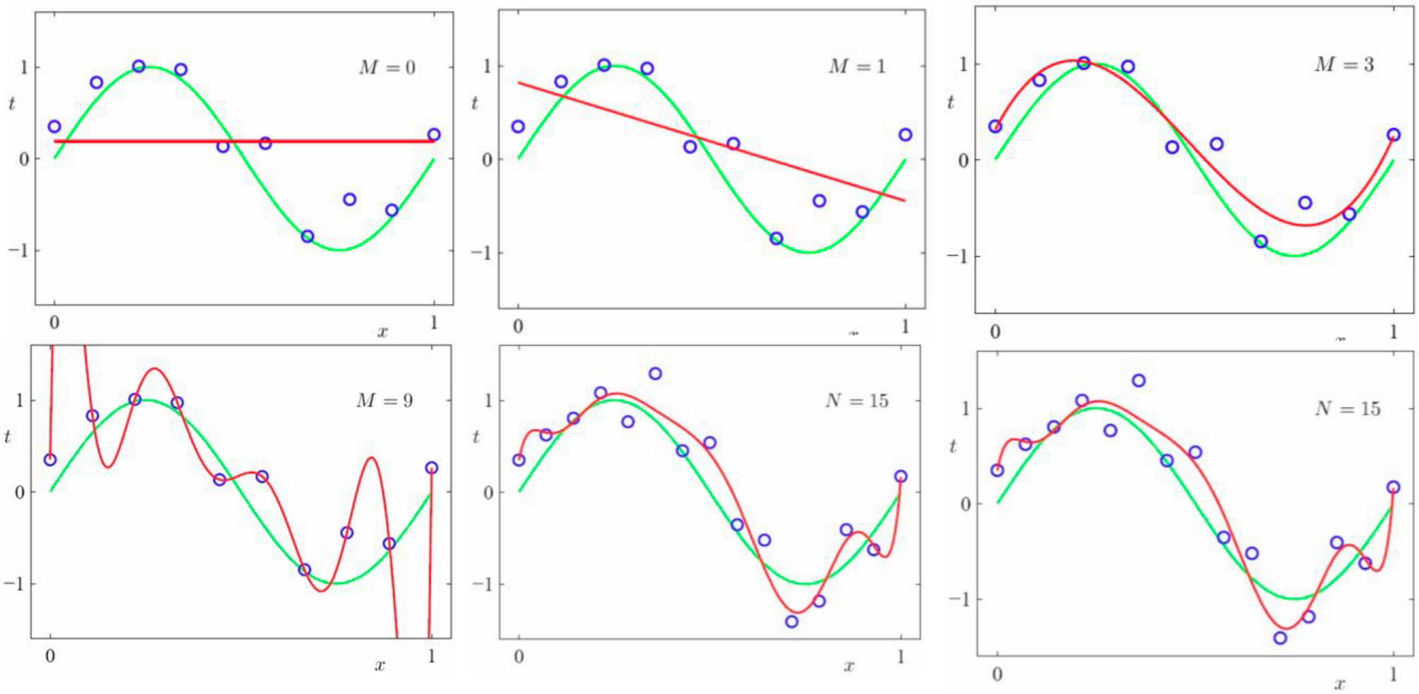
$$y(x, w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = w^{\mathrm{T}} \phi_M(x).$$

- Введем функцию потерь:

$$L\big((x, t), y\big) = \frac{1}{2}(y(x, w) - t)^2.$$

- Среди множества полиномов будем выбирать тот, который приносит наименьшие суммарные потери на обучающей выборке.

Переобучение



Переобучение

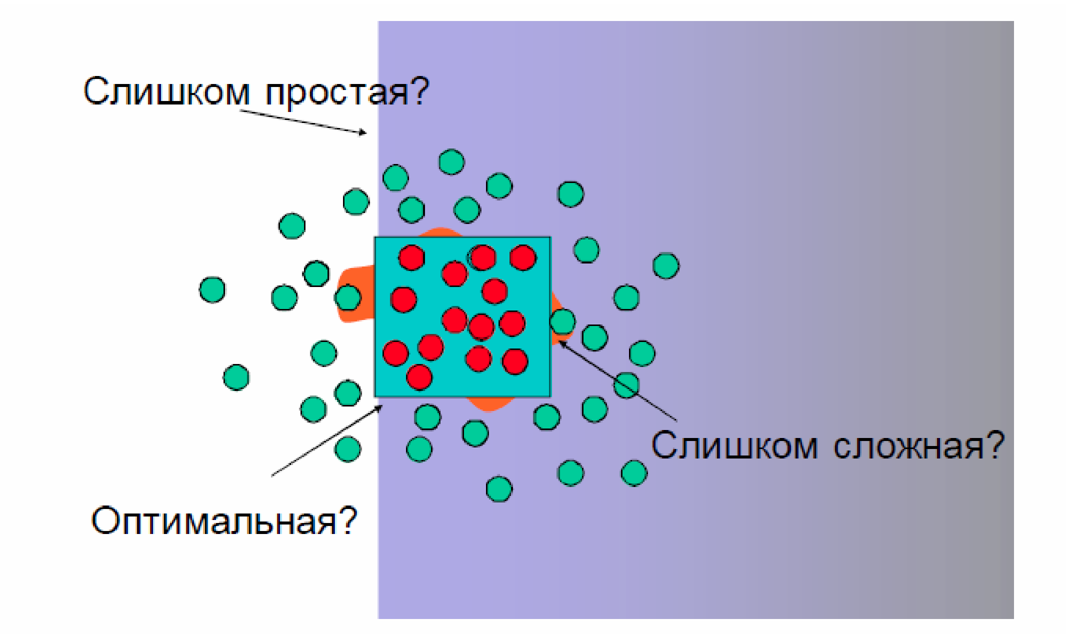
- Причина: гипотеза хорошо описывает свойства не объектов в целом, а только лишь объектов из обучающей выборки:
  - Слишком много степеней свободы параметров модели алгоритма (сложная модель);
  - Зашумленность данных;
  - Плохая обучающая выборка.

Теория Вапника-Червоненкиса (VC)

1. Оценка сложности параметрического семейства функций;
2. Оценка качества алгоритма через эмпирический риск и сложность модели.

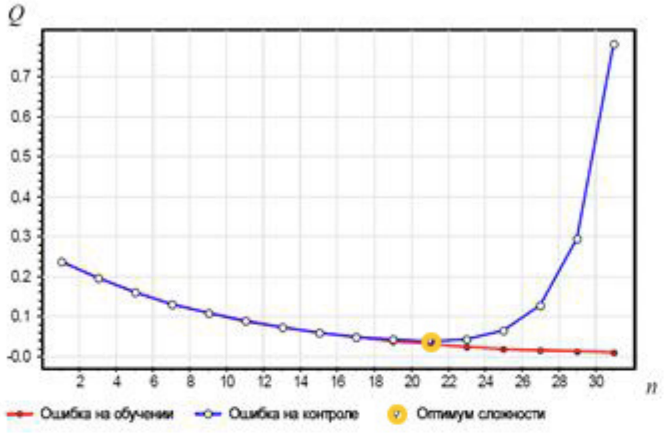
- Основная идея: выбор наиболее простой модели из достаточно точных.
- Пусть имеется последовательность вложенных параметрических семейств возрастающей сложности: $F_1 \subset F_2 \subset \cdots \subset F_h = F$.
- Необходимо выбрать семейство с минимальной сложностью, обеспечивающее нужную точность.

Иллюстрация



Практический вывод

- Требуется баланс между сложностью модели, обеспечивающей низкий эмпирический риск и простотой, обеспечивающей способность к обобщению.



Красная линия – ошибка на обучении,
синяя линия – ошибка на контроле,
отмеченная точка – оптимум сложности.

Оценка общего риска

- Минимизация общего риска является основной целью.
- Однако, его нельзя вычислить, поскольку требуются вычисления на неограниченном множестве:

$$R(f, X) = P_{X_m}(f(x) \neq y) = \int_x P(x)[f(x) \neq y]dx$$

Удерживание

- Оценим общий риск ошибкой на некотором конечном подмножестве X^c не пересекающемся с обучающей выборкой:

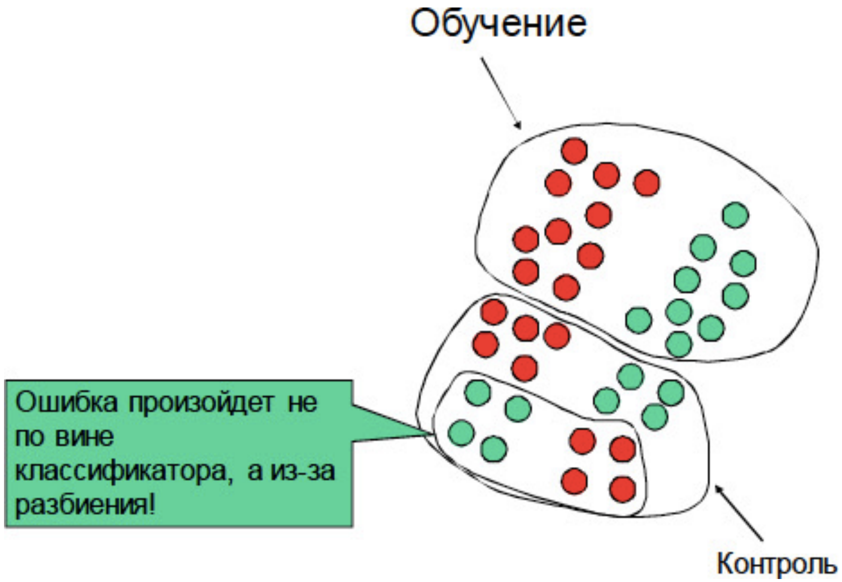
$$R(f, X) \sim P(f(x) \neq y | X^c) = \frac{1}{c} \sum_{j=1}^{c} [f(x_j) \neq y_j]$$

- Пусть имеется набор данных $X^k = \{x_1, \dots, x_k\}$ с известными ответами.
- Разобьем $X^l \cup X^c = X^k$: $X^l \cap X^c = \emptyset$.
- Будем использовать для обучения X^l, а для контроля X^c:

$$P(f(x) \neq y) \approx P(f(x) \neq y | X^c)$$

Удерживание

- Характеристики:
  1. Быстро и просто рассчитывается.
  2. Некоторые сложные прецеденты могут полностью попасть в только одну из выборок и тогда оценка ошибки будет смещенной.



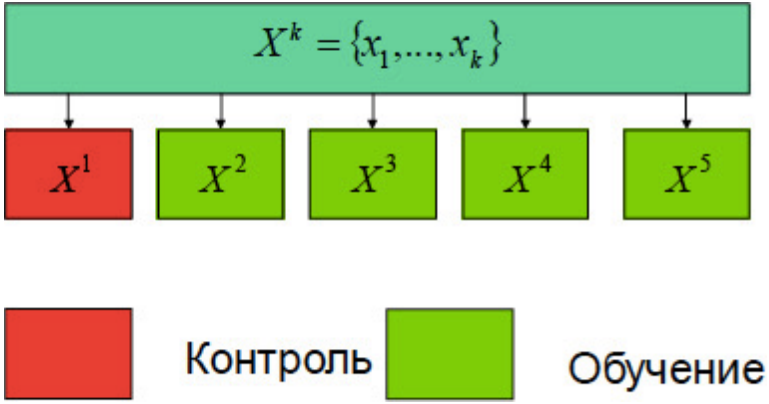
Скользящий контроль

- Разделим выборку на d непересекающихся частей и будем поочередно использовать одно из них для контроля, а остальные для тренировки.
- Разбиваем: $\{X^i\}^d : X^i \cap X^j = \emptyset, i \neq j, \bigcup_{i=1}^{d} X^i = X^k$.
- Вычислим приближенный риск:

$$P(f(X^k) = y^*) \approx \frac{1}{d}\sum_{i=1}^{d} P(f(X^i) \neq y^* | \bigcup_{i \neq j} X^i).$$

Скользящий контроль

- Результат считаем, как среднюю ошибку по всем итерациям.



Свойства скользящего контроля

- В пределе приближенный риск будет равен общему риску.
- Каждый прецедент будет присутствовать в контрольной выборке.
- Обучающие выборки будут сильно перекрываться (чем больше сегментов, тем больше перекрытие).
- Если одна группа «сложных прецедентов» попала полностью в один сегмент, то оценка будет смещенной.

Перекрестный скользящий контроль

- CV – Cross Validation
- 5-2 cross-validation:
  - Разделим выборку случайным образом пополам.
  - Обучим алгоритм на одной половине, протестируем на другой и наоборот.
  - Повторим этот эксперимент пять раз и усредним результат.
- Свойство: каждый из прецедентов будут участвовать в контрольных выборках на каждом из 5 этапов.

Вопросы?

s.shavetov@itmo.ru