

MACHINE LEARNING

1. MACHINE LEARNING (INTRO)

1.1. What is machine learning?

Искусственный интеллект (ИИ, AI) — наука и технология создания интеллектуальных компьютерных программ.

Два подхода к обучению ИИ:

- *дедуктивное* (аналитическое) — знания формулируются экспертом. Программа выводит из них факты и новые правила.

Пример — **экспертные системы**.

- *индуктивное* (статистическое) — программа обучается на собственном опыте. Из эмпирических данных выводит общие правила. Под **машинным обучением** будет пониматься только этот вид обучения ИИ.

Машинное обучение (ML) — подраздел ИИ, изучающий методы построения алгоритмов, способных обучаться.

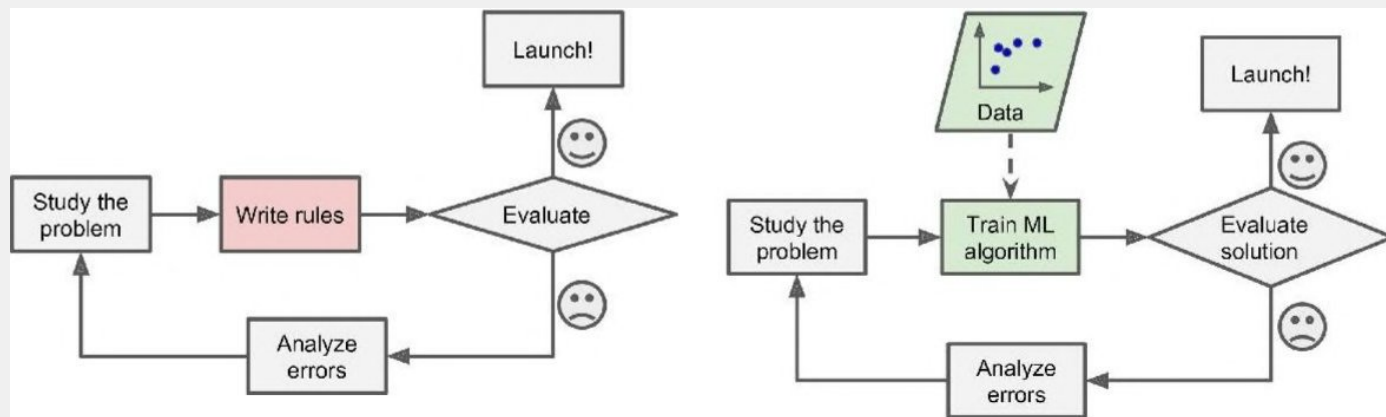


Рис. 1. Традиционный подход к решению задачи и подход на основе ML

В отличие от «классических» алгоритмов обученная машина способна показать поведение, которое в ней *не было* явно запрограммировано.

1.2. Терминология.

Модель = алгоритм(данные) = функция с варьируемыми параметрами, способная обучаться для решения некоторой задачи:

$$y = f_w(x).$$

Параметры (веса) модели — внутренние переменные w модели, которые оптимизируются в процессе обучения модели (от них зависит функция ошибок);

Гиперпараметры модели — величины, которые характеризуют модель, однако не входят в функцию ошибок. Как правило, постоянны во время обучения модели. Гиперпараметры настраиваются до начала обучения модели.

Функция ошибок (функция потерь, loss) — функция от весов модели, которая показывает насколько сильно ошибается модель.

Обучение (fitting) — процесс минимизации функции ошибок, т.е. *настройка* весов модели таким образом, чтобы ошибка была минимальна (на каком-то фиксированном наборе данных, называемом **train dataset**).

Валидация — проверка качества модели на наборе данных (**validation dataset**), который не участвовал в процессе обучения.

1.3. Виды машинного обучения.

(1) *с учителем* (обучение по прецедентам):

- **классификация** — прогноз метки класса
- **регрессия** — прогноз вещ. числа (интерполяция либо аппроксимация).
- **ранжирование** — сортировка по релевантность текста к поисковому запросу

(2) *без учителя*:

- **неконтролируемое преобразование**: сокращ. размерности, поиск компонент
- **кластеризация** — разбиение на группы схожих элементов (когда не известен правильный ответ)

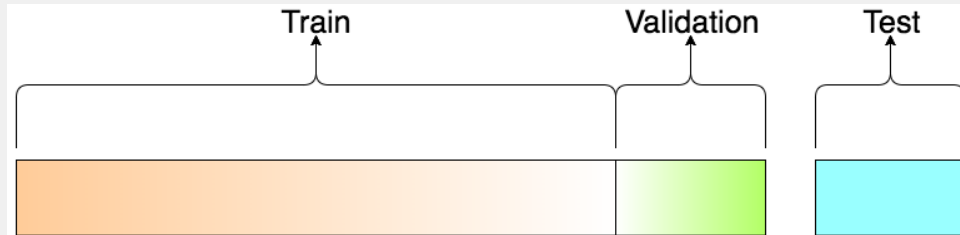
(3) *обучение с подкреплением* (reinforcement learning) — обучение алгоритмов реагировать на внешнее окружение:

- марковский процесс принятия решений

1.4. Подготовка данных и обучение модели.

Деление выборки (dataset) на 3 непересекающиеся части:

- *обучающая* (**training**) — используется для настройки (оптимизации) параметров) модели (например, весов нейросети).
- *валидационная* (**validation**) — используется для выбора наилучшей модели среди обучаемых (оптимизация гиперпараметров).
- *тестовая* (**testing**) — для окончательной проверки, которая производится только один раз в конце исследования.



1.5. Обучение с учителем.

Пусть

X — множество объектов (**samples**),

D — множество «правильных» ответов (меток) (**labels**),

$g : X \rightarrow D$ — неизвестная истинная зависимость (**true, target**).

Обучающая выборка — это множество объектов, для которых известен ответ:

$$X_0 = \{(x_i, d_i)\} \subset X.$$

Пара вида $(x_i, d_i) \in (X, D)$ называется *прецедент*. Т.е. прецедент — объект $x \in X$, для которого известен (истинный) ответ $d \in D$.

Задача обучения с учителем состоит в том, чтобы по обучающей выборке X_0 восстановить зависимость g , т.е. чтобы научиться получать ответ для **любого** объекта $x \in X$.

— это обобщение задачи *аппроксимации* функции на произвольные виды объектов (изображения, видео, тексты, различные нечисловыми или неточные данные).

Пусть $y(x)$ — ответ модели f на объекте $x \in X$.

Ошибка модели на x — это разность $\varepsilon(x) = d(x) - y(x)$.

Пусть $Y = \{f(x) \mid x \in X\}$ — множество ответов модели. Получаемая таким образом зависимость $f : X \rightarrow Y$ называется *решающей функцией* (decision function).

Таким образом, задача ML — построить решающую функцию f , которая бы приближала целевую (истинную) зависимость g .

Задача регрессии — задача ML для случая $D = \mathbb{R}$.

Например, задача прогнозирования (forecasting).

Задача классификации — для случая $D = \{1, 2, \dots, m\}$, где $m \in \mathbb{N}$.

Подмножество объектов $K_d = \{x \in X \mid g(x) = d\}$ называется *классом* с меткой $d \in D$.

Таким образом, задача классификации — ответить на вопрос «какому классу принадлежит объект $x \in X$ »

В задаче распознавании образов классы называются *образами* (pattern).

Функция потерь

Функция потерь (loss), *функция ошибок*, *функция стоимости* — неотрицательная функция от параметров модели f , которая характеризует величину ошибки алгоритма на заданном объекте x :

$$L_f = L_f(\varepsilon(x)), \quad \varepsilon(x) = d(x) - y(x).$$

Например,

$$L_f(x) = (d(x) - y(x))^2.$$

Функционал качества (ф-л средних потерь) алгоритма f на выборке X — это усредненное значение функции потерь на заданном подмножестве объектов:

$$Q_f(X) = \text{avg}(L_f(X)) = \frac{1}{|X|} \sum_{x \in X} L_f(x).$$

Соответственно различают: avg train loss, avg valid loss, avg test loss.

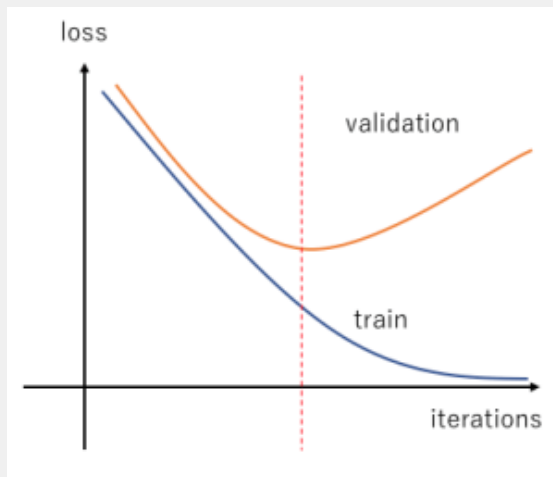
1.6. Вопросы использования моделей.

Две основные стадии:

- (1) *обучение* (fitting) — построение по обучающей выборке X_0 с помощью некоторого метода μ машинного обучения алгоритма (модели) $f_w = \mu(X_0)$.
- (2) *применение* (inference) — использование обученной модели f_w для предсказания результата на новых данных: $y = f_w(x)$.

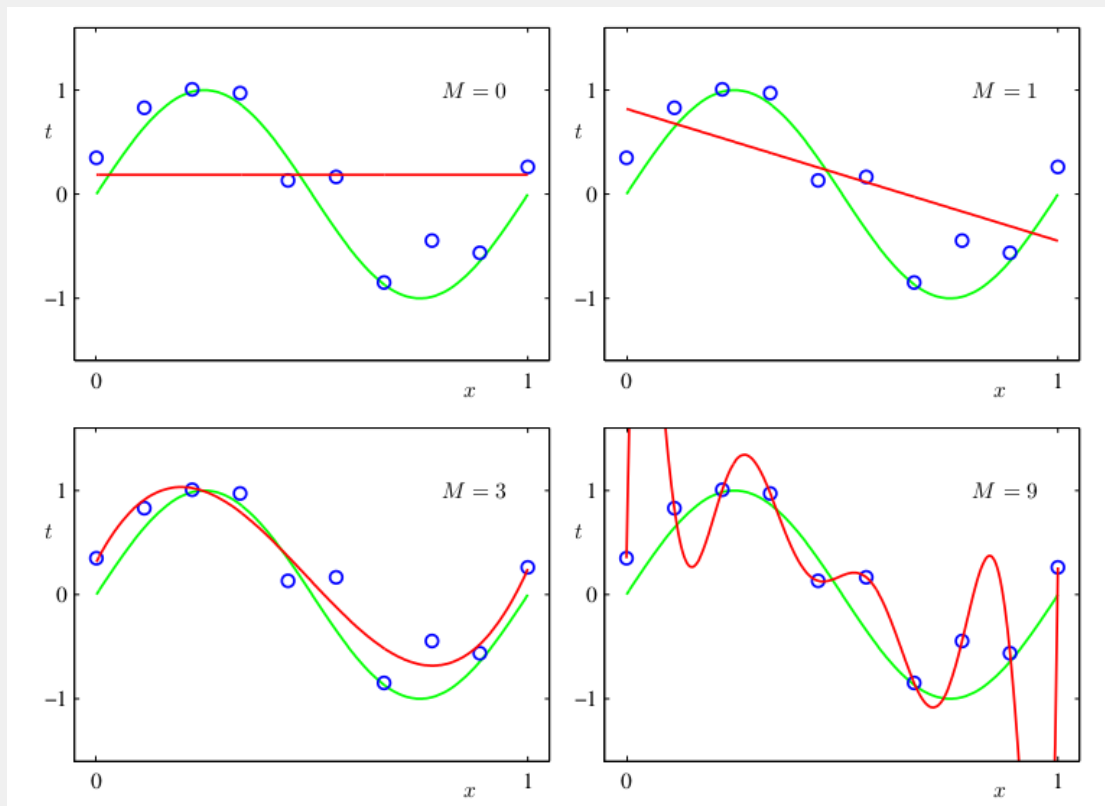
1.7. Overfitting.

Переобучение (overfitting) — ситуация, когда средняя ошибка модели на валидационной (и тестовой) выборке оказывается существенно выше, чем ошибка на обучающей выборке:



Переобученная модель хорошо запоминает образы, на которых обучается, но не обладает достаточной обобщающей способностью.

На примере аппроксимации функции полиномом (M — степень полинома):



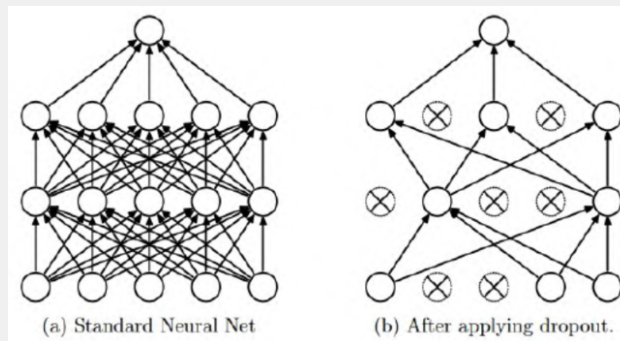
1.8. Методы борьбы с переобучением.

- *Аугментация* (augmentation) — увеличение размера обучающей выборке за счет применения к ней различных искажений и деформаций: вращение, аффинные преобразования, цветовое, добавление шума.
- *Регуляризация* (regularization) — добавление к функции потерь слагаемого в виде штрафа за сложность модели:

$$L1: \text{MSE} += \lambda ||W||_1,$$

$$L2: \text{MSE} += \lambda ||W||_2^2.$$

- *Dropout* — случайное «удаление» части нейронов в скрытых полносвязных слоях.



2. REFERENCE

<http://www.machinelearning.ru>

<https://github.com/demidovakatya/vvedenie-mashinnoe-obuchenie>

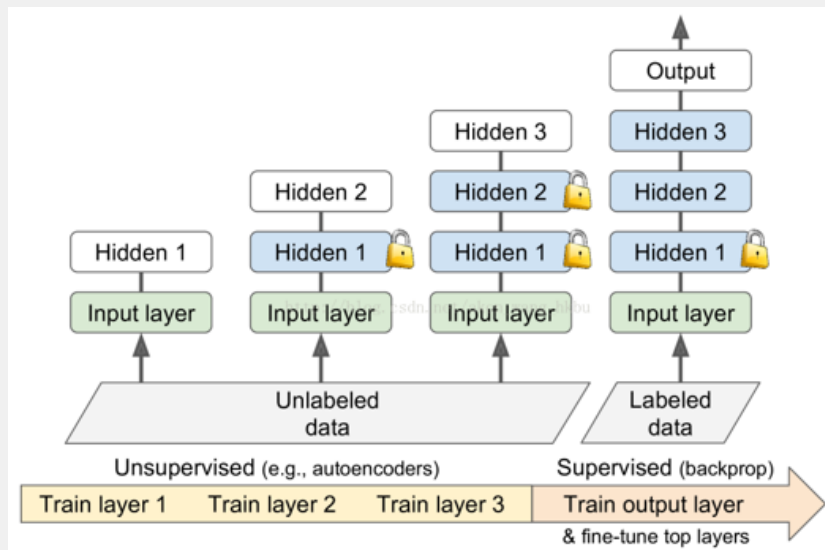


Рис. 2. Предобучение без учителя (unsupervised pretraining)

2.1. Предобучение.