



Конспект > 1 урок > Введение в МО: каким оно бывает и каковы основные компоненты

> Оглавление

> Оглавление

> Основные понятия и определения

> Типы и задачи машинного обучения

Обучение с учителем (Supervised Learning)

Обучение без учителя (Unsupervised Learning)

Обучение с подкреплением (Reinforcement Learning)

> Компоненты классической ML задачи

> Выборка: объекты и признаки (objects and features)

> Ответ (Target)

> Основные понятия и определения

Машинное обучение (Machine learning) — комплекс методов и инструментов, позволяющий компьютеру строить модели с сильной экспертизой на основе входящих из реального мира данных и накопленных человеком знаний и опыта.

Два основных блока, рассматриваемых в машинном обучении — **классическое машинное обучение** и **глубокое обучение** (Deep learning).

Датасет (Dataset) — структурированный набор данных, как правило, представленный в виде одной или нескольких таблиц, где строкам соответствуют некие объекты (objects), а столбцам - их признаки (features).

Признаки или фичи (Features) — некоторые характеристики, на основе которых проводится обучение модели.

Алгоритм (Algorithm) — набор однозначных пошаговых инструкций, которые компьютер может выполнять для достижения определенной цели.

Модель (Model) — некая абстракция, с помощью данных обученная распознавать определенные типы закономерностей.

Ответ или целевая переменная (Target) — признак датасета, который предстоит предсказывать модели машинного обучения.

Множество — совокупность объектов (элементов множества). Запись вида $U = q, w, r, t$ означает, что множество U состоит из четырех элементов — q, w, r, t .

R — множество действительных чисел, числа, которые можно записать в виде конечной или бесконечной, периодической или непериодической десятичной дроби

$(R = 1; -1; -21; 0, 12; \pi; 2; \dots)$.

Z — множество целых чисел ($Z = \dots - 3, -2, -1, 0, 1, 2, 3, \dots$).

\subset — знак включения элемента как подмножества множества ($Z \subset R$).

\in — знак принадлежности отдельного элемента к множеству ($a \in Z$).

Подробнее о множествах и их условных обозначениях [тут](#)

> Типы и задачи машинного обучения

Примеры задач, решаемых с помощью машинного обучения:

- Прогноз стоимости акций
- Определение кредитоспособности (скоринг) банковских клиентов

- Поиск аномалий в процессах (например, на производстве)
- Отнесение объекта к той или иной группе по определенным признакам
- Ответы на сложные вопросы

Главные преимущества моделей машинного обучения по сравнению с человеческой экспертизой:

- Высокая скорость ответа
- Работа с большим объемом данных
- Масштабируемость и доступность

Можно выделить три основных типа машинного обучения:

Обучение с учителем (Supervised Learning)

Построение моделей, основанных на прецедентах (есть примеры правильных ответов, на которых учится модель).

- **Регрессия** — предсказание некоего числового значения (например, цена квартиры по ее площади и расположению)
- **Классификация** (бинарная/многоклассовая) — определение объекта в ту или иную заранее определенную группу

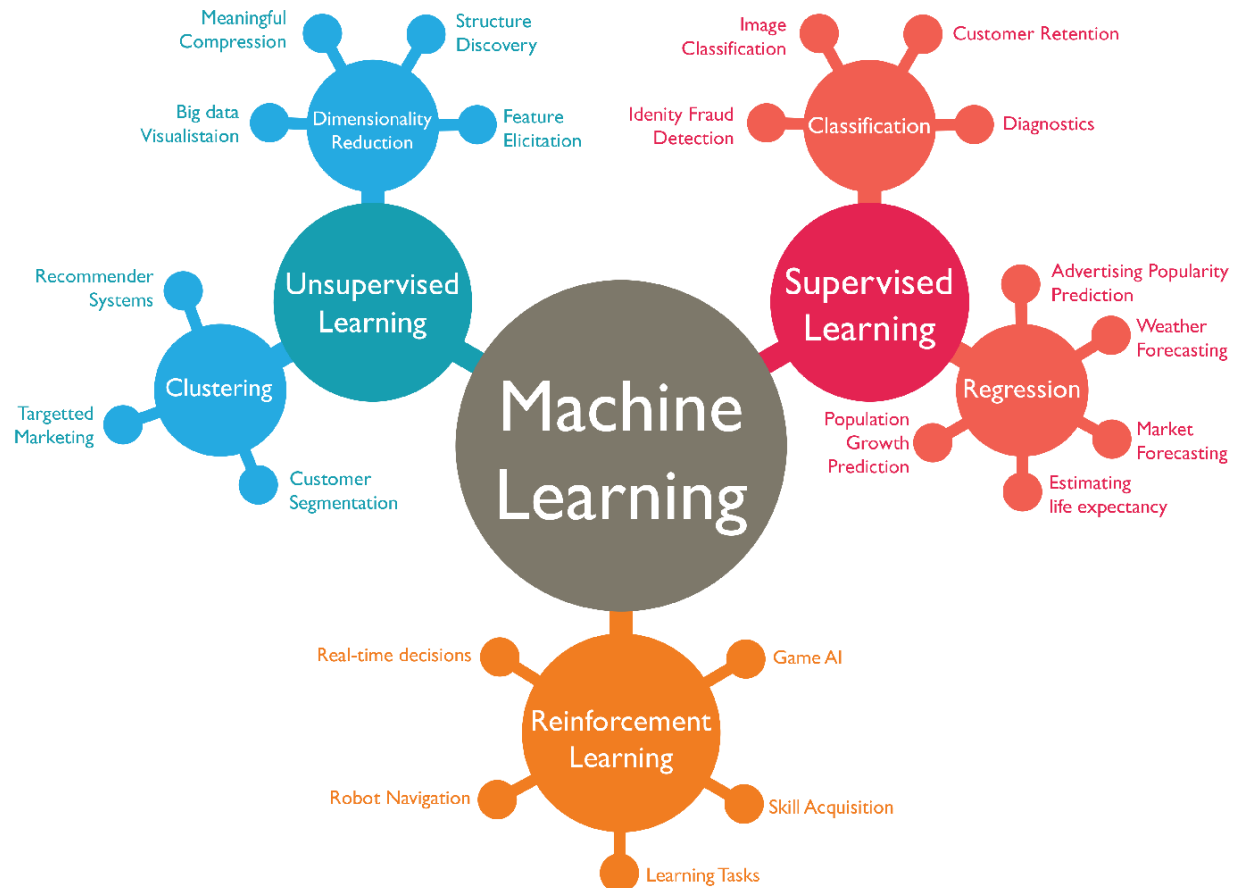
Обучение без учителя (Unsupervised Learning)

- **Кластеризация** — объединение объектов в группы, сходных по неким признакам (сегментация пользователей, группы генов с похожей экспрессией). При этом заранее не известно, какие это будут группы.
- **Рекомендательные системы** — подбор контента на основе предыдущего поведения пользователя
- **Поиск аномалий** — поиск объекта, сильно отличающегося от других

Обучение с подкреплением (Reinforcement Learning)

Ответ заранее неизвестен. Система обучается самостоятельно, взаимодействуя с некоторой средой, получая от нее негативные и позитивные сигналы, тем самым как бы выступая в роли учителя. Наглядный пример — нейросеть учится переходить дорогу.

Подробнее с типами машинного обучения можно познакомиться [ТУТ](#) и [ТУТ](#).



> Компоненты классической ML задачи

- Выборка
- Ответ (target)
- Функция потерь и функционал качества
- Метрики качества
- Алгоритм\семейство моделей
- Оценка модели

> Выборка: объекты и признаки (objects and features)

Выборка — набор объектов и признаков (objects and features), для которых мы хотим ответить на поставленный вопрос. В машинном обучении принято представлять каждый объект вектором, например объект x представляет собой совокупность признаков (d):

$$x_i = (d_1, d_2, \dots, d_k)$$

Выборка (X) будет обозначать множество всех объектов с ответами (y), если они есть:

$$X = \{(x_i, y_i)\}$$

Признаки или фичи (features) — некоторые характеристики, на основе которых проводится обучение модели.

Признаки бывают:

- **Вещественные** — значением признака является любое число на непрерывной числовой прямой (температура тела, рост, котировки акций)

$$d_j \in R$$

- **Бинарные** — когда признак может принимать одно из двух значений (М/Ж, болен/здоров)

$$d_j \in 0, 1$$

- **Категориальные** — признаки, которые невозможно упорядочить (профессиональная деятельность, город)

$$d_j \in D$$

- **Порядковые** — признак принимает целое значение в некой ограниченной области (группа инвалидности, количество детей в семье)

$$d_j \in D \subset Z$$

- **Множественные** — признак представлен набором чисел или слов (принимаемые лекарства)

> Ответ (Target)

- **Множественным** (сделать рекомендации по лечению)

Ответ (таргет) или целевая переменная – признак датасета, который предстоит предсказывать модели машинного обучения.

Множество всех ответов будем обозначать Y , а ответы к каждому конкретному объекту - y_i .

$$Y = y$$

Как и признаки, ответ, полученный с помощью машинного обучения, может быть:

- **Вещественным** (цена акции через год)

$$y_i \in R$$

- **Бинарным** (цена акций вырастет/не вырастет через год)




$$y_i \in 0, 1$$

- **Множественным** (какая стратегия будет лучшей: a,b или c).

$$y_i \in 0, 1, 2, 3, \dots, d$$

- **Задачи без явного таргета** (какие акции добавить в портфель: {Apple, Microsoft, Google}).

$$y_i \in \emptyset$$

	d^1 Сколько активов на счете компаний?	d^2 Дефолтила ли компания последние 5 лет?	d^3 Отрасль фирмы?	d^4 Какой у компании кредитный рейтинг?	d^5 Кто крупнейшие конкуренты?	y Сколько будет стоить одна акция через год?
	500M \$	Нет	Металлургия	AA	ИП Кузьминов Алексей	300\$
	100M \$	Да	Ремонт	BBB	∅	100\$
	250M\$	Нет	Телеком	AAA	Билайн, Мегафон	Не листингуется