Машинное обучение состоит из 3 компонент:

- Представление (Representation)
- Оценка (Evaluation)
- Оптимизация (Optimization)

Представление

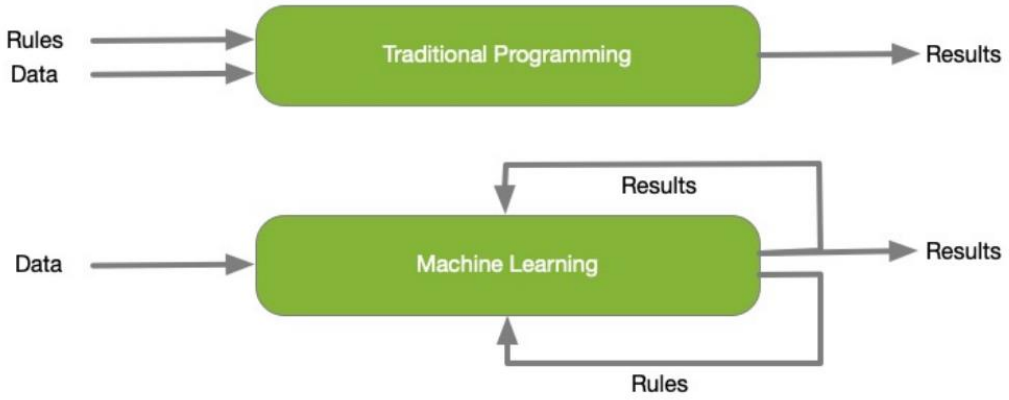
- Что собой представляет модель, какие классы задач она способна (и не способна) решать
- Пример: разделяющая гиперплоскость, деревья решений, нейросети.
- Как именно представляются данные? – Этапы выделения нужных признаков (feature extraction)

Оценка

- Как оценивать качество модели в контексте решения задачи, как выбирать лучшую модель из нескольких.
- Пример: RMSE (СКО), Accuracy/Precision/Recall, Logistic Loss

Оптимизация

- Как именно проводить обучение модели, каким именно образом осуществлять перебор пространства возможных моделей, чтобы найти лучшую.
- Пример: стохастический градиентный спуск, генетические алгоритмы, grid search



Искусственный интеллект (ИИ) — свойство интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека

ИИ связан со сходной задачей использования компьютеров для понимания человеческого интеллекта, но не обязательно ограничивается биологически правдоподобными методами.

Искусственный интеллект:

- ограниченный искусственный интеллект (Narrow AI);
- общий искусственный интеллект (AGI);
- сверхразумный искусственный интеллект

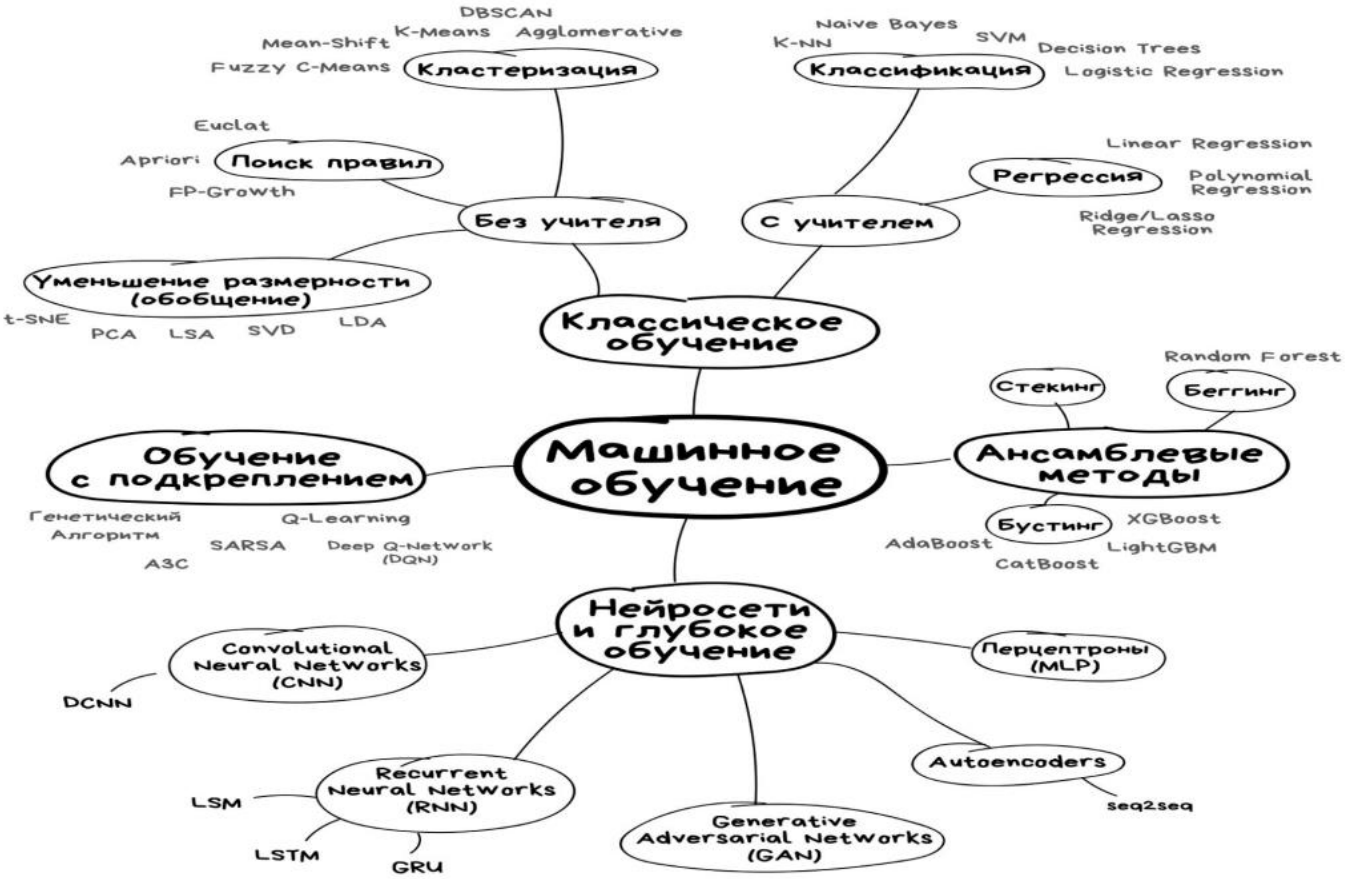
Машинное обучение — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач.

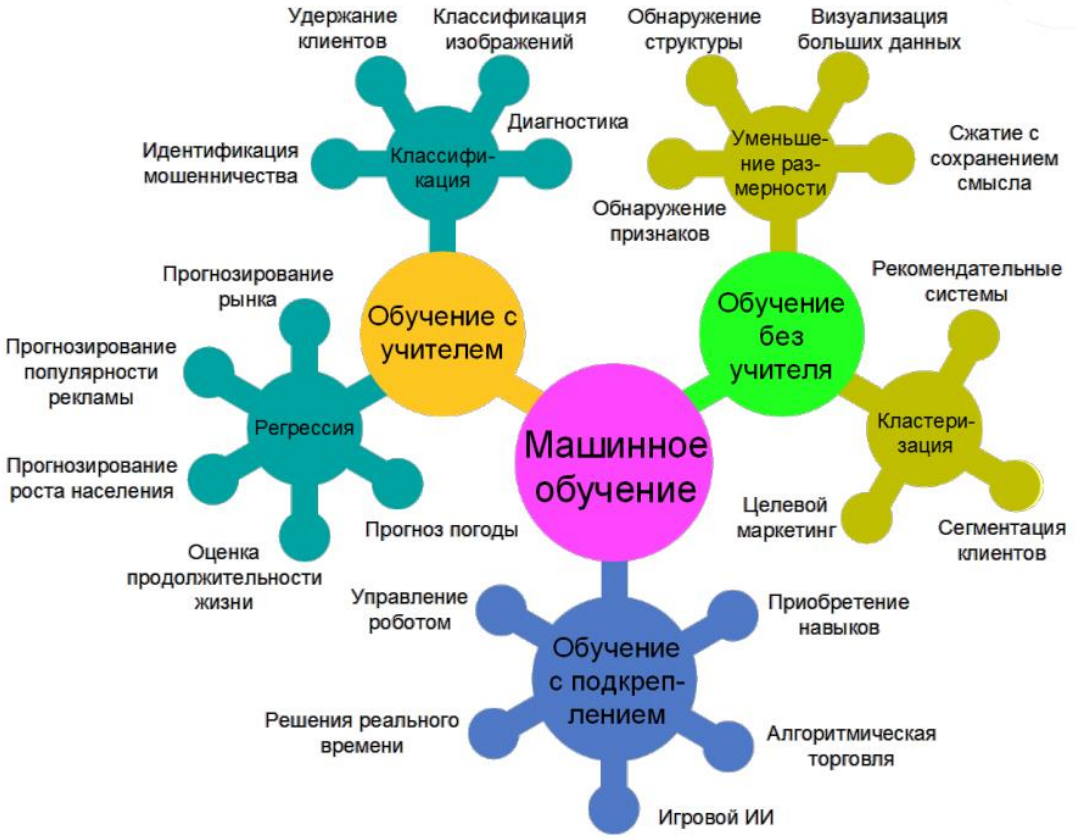
Машинное обучение - это подраздел искусственного интеллекта, который в широком смысле определяется как способность машины имитировать разумное поведение человека. Системы искусственного интеллекта используются для выполнения сложных задач аналогично тому, как люди решают проблемы.

Для построения таких методов используются средства математической статистики, численных методов, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

- Решение задач путем обработки прошлого опыта
- Альтернатива построению математических моделей

- Основное требование – наличие обучающих данных
- Часто в качестве обучающей информации выступает выборка прецедентов – ситуационных примеров из прошлого с известным исходом, где требуется построить алгоритм, который позволял бы обобщить опыт прошлых наблюдений/ситуаций для обработки новых, не встречавшихся ранее случаев, исход которых неизвестен.





Классификация

- В классической задаче классификации обучающая выборка представляет собой набор отдельных объектов $X = \{x_i\}_{i=1}^{n}$, характеризующихся вектором значимых признаков $x_i = (x_{i,1}, \dots, x_{i,d})$
- В качестве исхода объекта x фигурирует переменная t, принимающая конечное число значений, обычно из множества $\mathcal{T} = \{1, \dots, l\}$
- Требуется построить алгоритм (классификатор), который по вектору признаков x вернул бы метку класса $\hat{t}$ или вектор оценок принадлежности (апостериорных вероятностей) к каждому из классов $\{p(s|x)\}_{s=1}^{l}$

  - Медицинская диагностика: по набору медицинских характеристик требуется поставить диагноз
  - Геологоразведка: по данным зондирования почв определить наличие полезных ископаемых
  - Оптическое распознавание текстов: по отсканированному изображению текста опре делить цепочку символов, его формирующих
  - Кредитный скоринг: по анкете заемщика принять решение о выдаче/отказе кредита
  - Синтез химических соединений: по параметрам химических элементов спрогнозировать свойства получаемого соединения

Задача медицинской диагностики

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

    - Бинарные: пол, головная боль, слабость, тошнота и т. д.
    - Порядковые: желтушность, тяжесть состояния и т. д.
    - Количественные: давление, пульс, возраст и т. д.
- Особенности задачи
    - Обычно много «пропусков» в данных
    - Нужен интерпретируемый алгоритм классификации
    - Нужно выделять синдромы – сочетания симптомов
    - Нужна оценка вероятности отрицательного исхода

Задача категоризации текстовых документов

- Объект – текстовый документ
- Классы – рубрики иерархического тематического каталога
- Пример признаков
    - Номинальные: автор, издание, год и т. д.
    - Количественные: для каждого термина – частота в тексте, в аннотации, в заголовках и т. д.
- Особенности задачи
    - Лишь небольшая часть документов имеет метки
    - Документ может относиться к нескольким рубрикам
    - В каждом ребре дерева свой классификатор на 2 класса

Регрессия

- Основывается на влиянии одной группы непрерывных случайных величин на другую группу непрерывных случайных величин
- В классической задаче восстановления регрессии обучающая выборка представляет собой набор отдельных объектов $X = \{x_i\}_{i=1}^{n}$, характеризующихся вектором значимых признаков $x_i = (x_{i,1}, \dots, x_{i,d})$
- В качестве исхода объекта x фигурирует непрерывная вещественнозначная переменная t
- Требуется построить алгоритм (регрессор), который по вектору признаков x вернул бы точечную оценку значения регрессии $\hat{t}$, доверительный интервал (t_-, t_+) или апостериорное распределение на множества значений регрессионной переменной $p(t|x)$

- Оценка стоимости недвижимости: по характеристике района, экологической обстановке, транспортной связности оценить стоимость жилья
- Прогноз свойств соединений: по параметрам химических элементов спрогнозировать температуру плавления, электропроводность, теплоемкость получаемого соединения
- Медицина: по постоперационным показателям оценить время заживления органа
- Кредитный скоринг: по анкете заемщика оценить величину кредитного лимита
- Инженерное дело: по техническим характеристикам автомобиля и режиму езды спрогнозировать расход топлива

Задача прогнозирования стоимости недвижимости

- Объект – квартира в Москве
- Примеры признаков:
    - Бинарные: наличие балкона, лифта, охраны и т. д.

    - Номинальные: район, тип дома (кирпичный, монолит, блочный) и т. д.
    - Количественные: число комнат, жилая площадь, расстояние до центра, возраст дома и т. д.
- Особенности задачи:
    - Выборка неоднородна, стоимость меняется со временем
    - Разнотипные признаки
    - Для линейной модели нужны преобразования признаков

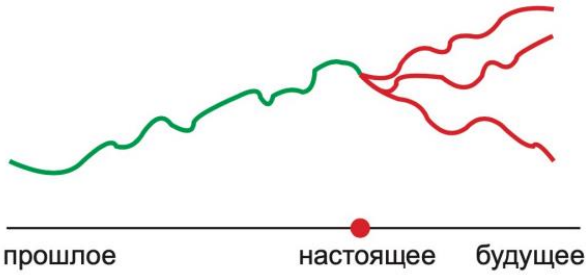
Идентификация

- Используется в классификации, когда есть необходимость отделить объекты, обладающие определенными свойствами, от основных данных
- В задаче идентификации обучающая выборка представляет собой набор отдельных объектов $X = \{x_i\}_{i=1}^{n}$, характеризующихся вектором значимых признаков $x_i = (x_{i,1}, \dots, x_{i,d})$, обладающим некоторым свойством $\chi_A(x) = 1$
- Особенностью задачи является то, что все объекты принадлежат одному классу, причем не существует возможности сделать репрезентативную выборку из класса все остальные данные
- Требуется построить алгоритм (идетификатор), который по вектору признаков x определил бы наличие свойства A у объекта x, либо вернул оценку степени его выраженности $p(\chi_A(x) = 1|x$

- Медицинская диагностика: по набору медицинских характеристик требуется установить наличие/отсутствие конкретного заболевания
- Системы безопасности: по камерам наблюдения в подъезде идентифицировать жильца дома
- Банковское дело: определить подлинность подписи на чеке
- Обработка изображений: выделить участки с изображениями лиц на фотографии
- Искусствоведение: по характеристикам произведения (картины, музыки, текста) определить, является ли его автором тот или иной автор

Прогнозирование

- Используется для предсказания временных рядов через какой-то промежуток времени
- В задаче прогнозирования обучающая выборка представляет собой набор измерений $X = \{x[i]\}_{i=1}^{n}$, представляющий собой вектор вещественных величин $\mathrm{x[i]} = (x_{1[i]}, \dots, x_{d[i]})$, сделанных в определенный момент времени
- Требуется построить алгоритм (предиктор), который вернул бы точную оценку $\{\hat{x}[i]_{i=n+1}^{n+q}$ доверительный интервал $\{(x_-[i], x_+[i])\}_{i=n+1}^{n+q}$ или апостериорное распределение $p(x[n+1], \dots, x[n+q]|x[1], \dots, x[n])$ прогноза на заданную глубину q
- В отличие от задачи восстановления регрессии, здесь осуществляется прогноз по времени, а не по признакам.



- Биржевое дело: прогнозирование биржевых индексов и котировок
- Системы управления: прогноз показателей работы реактора по данным телеметрии
- Экономика: прогноз цен на недвижимость

- Демография: прогноз изменения численности различных социальных групп в конкретном ареале
- Гидрометеорология: прогноз геомагнитной активности

Извлечение знаний

- Используется при исследовании взаимозависимостей между косвенными показателями одного и того же явления
- В задаче извлечения знаний обучающая выборка представляет собой набор отдельных объектов $X = \{x_i\}_{i=1}^{n}$, характеризующийся вектором вещественных признаков $x_i = (x_{i,1}, \dots, x_{i,d})$
- Требуется построить алгоритм, генерирующий набор объективных закономерностей между признаками, имеющих место в генеральной совокупности
- Закономерности обычно имеют форму предикатов «ЕСЛИ ... ТО ...» и могут выражаться как в цифровых терминах $((0.36 \le x4 \le 31.2)\&(-6.73 \le x7 \le -6.29) \Rightarrow (3.21 \le x2 \le 3.345))$, так и в текстовых («ЕСЛИ Давление – низкое И (Реакция – слабая ИЛИ Реакция – отсутствует) ТО Пульс – нитевидныий»)

- Медицина: поиск взаимосвязей (синдромов) между различными показателями при фиксированной болезни
- Социология: определение факторов, влияющих на победу на выборах
- Генная инженерия: выявление связанных участков генома
- Научные исследования: получение новых знаний об исследуемом процессе
- Биржевое дело: определение закономерностей между различными биржевыми показателями

Кластеризация

- Возникла из задачи группировки схожих объектов в единую структуру (кластер) с последующим выявлением общих черт
- В задаче кластеризации обучающая выборка представляет собой набор отдельных объектов $X = \{x_i\}_{i=1}^{n}$, характеризующийся вектором вещественных признаков $x_i = (x_{i,1}, \dots, x_{i,d})$
- Требуется построить алгоритм (кластеризатор), который разбил бы выборку на непересекающиеся группы (кластеры) $X = \cup_{j=1}^{k} C_k, C_j \subset \{x_1, \dots, x_m\},\ C_i \cap C_j = 0$
- В каждый класс должны попасть объекты в некотором смысле похожие друг на друга

- Экономическая география: по физико-географическим и экономическим показателям разбить страны мира на группы схожих по экономическому положению государств
- Финансовая сфера: по сводкам банковских операций выявить группы «подозрительных», нетипичных банков, сгруппировать остальные по степени близости проводимой стратегии
- Маркетинг: по результатам маркетинговых исследований среди множества потребителей выделить характерные группы по степени интереса к продвигаемому продукту
- Социология: по результатам социологических опросов выявить группы общественных проблем, вызывающих схожую реакцию у общества, а также характерные фокусгруппы населения

Алгоритм логической регрессии

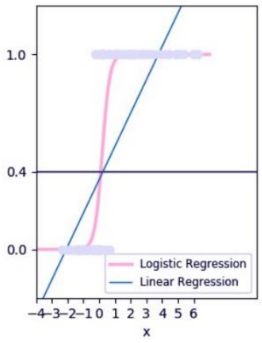


График логистической функции:

$$f(x) = \frac{1}{(1 + e^{-x})}$$

Пример уравнение логистической регрессии:

$$y = \frac{e^{(b_0 + b_1 * x)}}{1 + e^{(b_0 + b_1 * x)}}$$

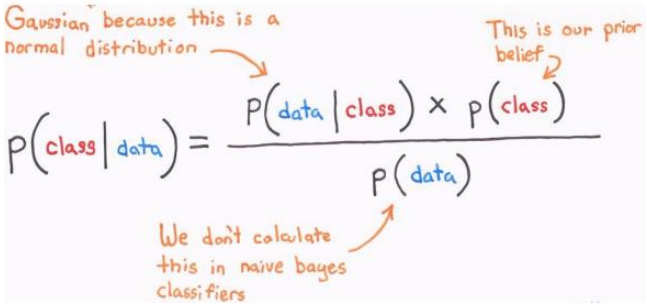
где b_0 и b_1 коэффициента по входному x.

Наивный байесовский алгоритм (Naïve Bayes Algorithm)

- Naive Bayes - это простой и быстрый способ предсказать класс набора данных. С его помощью можно выполнять многоклассовое прогнозирование.
- Когда предположение о независимости справедливо, Naive Bayes гораздо более эффективен, чем другие алгоритмы, такие как логистическая регрессия.
- Требуется меньше обучающих данных.

Недостатки:

- Если категориальная переменная принадлежит к категории, которая не была изучена в обучающем наборе, то модель присвоит ей вероятность 0, что не позволит ей сделать какое-либо предсказание.
- Naive Bayes предполагает независимость между признаками. В реальной жизни трудно собрать данные, которые включали бы полностью независимые признаки.



Алгоритм дерева решений (Decision Tree Algorithm)

- Дерево для классификации, когда предсказываемый результат является классом, к которому принадлежат данные;
- Дерево для регрессии, когда предсказываемый результат можно рассматривать как вещественное число (например, цена на дом, или продолжительность пребывания пациента в больнице).

Алгоритмы построения деревьев:

- Алгоритм ID3
- Алгоритм C4.5 (улучшенная версия ID3)
- Алгоритм CART и его модификации — IndCART, DB-CART
- Автоматический детектор взаимодействия Хи-квадрат (CHAID). (Выполняет многоуровневое разделение при расчёте классификации деревьев;
- MARS: расширяет деревья решений для улучшения обработки цифровых данных

Метод k-ближайших соседей (англ. k-nearest neighbors algorithm, k-NN) — метрический алгоритм для автоматической классификации объектов или регрессии. В случае использования метода для классификации объект присваивается тому классу, который является наиболее распространённым среди соседей данного элемента, классы которых уже известны. В случае использования метода для регрессии, объекту присваивается среднее значение по ближайшим к нему объектам, значения которых уже известны

Метод Опорных Векторов или SVM (от англ. Support Vector Machines) — это линейный алгоритм используемый в задачах классификации и регрессии. Данный алгоритм имеет широкое применение на практике и может решать как линейные так и нелинейные задачи. Алгоритм создает линию или гиперплоскость, которая разделяет данные на классы.

Алгоритм стохастического градиентного спуска

Градиентный спуск — метод нахождения локального минимума или максимума функции с помощью движения вдоль градиента.

Стохастический градиентный спуск (англ. Stochastic gradient descent, SGD) — это итерационный метод для оптимизации целевой функции с подходящими свойствами гладкости (например, дифференцируемость или субдифференцируемость). Его можно расценивать как стохастическую аппроксимацию оптимизации методом градиентного спуска, поскольку он заменяет реальный градиент, вычисленный из полного набора данных его оценкой, вычисленной из случайно выбранного подмножества данных. Это сокращает задействованные вычислительные ресурсы и помогает достичь более высокой скорости итераций в обмен на более низкую скорость сходимости. Особенно большой эффект достигается в приложениях связанных с обработкой больших данных.

Глубокое обучение — совокупность широкого семейства методов машинного обучения, основанных на имитации работы человеческого мозга в процессе обработки данных и создания паттернов, используемых для принятия решений. Как правило, глубокое обучение предназначено для работы с большими объемами данных и использует сложные алгоритмы для обучения модели. На больших датасетах глубокое обучение показывает более высокую точность результатов в сравнении с традиционным машинным обучением.

Переобучение

Причины возникновения:

- Недоученные (underfitting)
    - Высокая ошибка при обучении
    - Ошибка обучения близка к тестовой
    - Высокое смещение\предвзятости (bias)
- Нормальное состояние
    - Ошибка обучения немного ниже ошибки тестирования
- Переобучение (overfitting)
    - Очень низкая погрешность обучения
    - Ошибка обучения значительно ниже ошибки тестирования
    - Высокая дисперсия

Возможные способы устранения (недоученное):

- Усложнить модель
- Добавить больше функций
- Обучаться дольше

Возможные способы устранения (переобучение):

- Выполнить регуляризацию
- Получить больше данных