Введение в Data Science и Machine Learning

Константин Ильченко • 24.03.2019

Обзор

Общие рассуждения:

-интеллект;
-свойство разумности;
-отличие машинного обучения от обычного программирования.

Намеки на базовые понятия:

-типы задач и методы обучения;
-методы обучения.

Основные “школы познания”.:

-символисты;
-коннекционисты;
-эволюционисты;
-б~~айесси~~анцы;
-аналогисты.

Общие рассуждения

Интеллект и разум

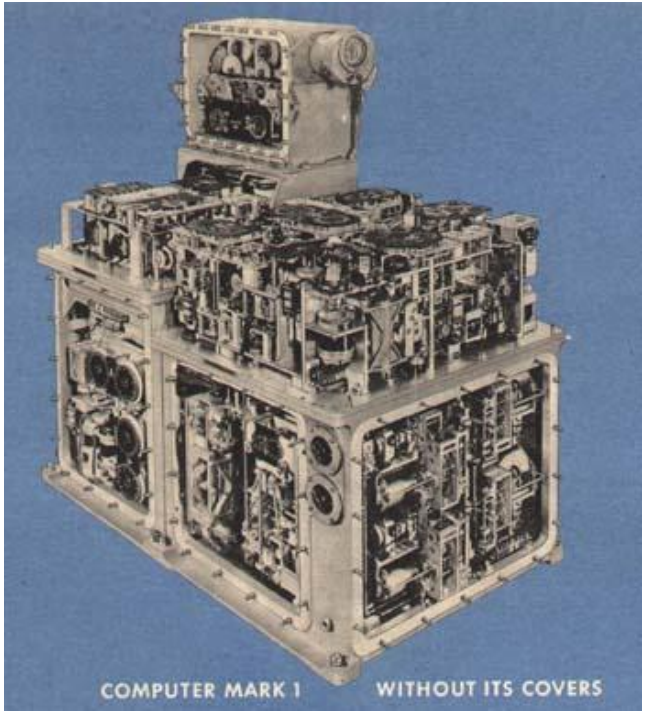
Интеллект

- Это способность воспринимать информацию и сохранять ее в качестве знания для построения адаптивного поведения в среде или контексте.
- Основная задача - “проложить путь” к “мишени” указанной механизмами целеполагания

и Разумность

- Восприятие
- Целеполагание
- Построение алгоритма действия для достижения цели

Ford Mark 1 - система управления огнем (ВМС США)



Интеллектуальные свойства

- Принимал данные о курсах и положении кораблей, а также метеоданные
- Проводил баллистические вычисления
- Выдавал параметры стрельбы на орудия

Отличие машинного обучения от обычного программирования

Обычное программирование

- Главная задача программиста - самому в ручную прописать все правила, которыми будет руководствоваться система в своем поведении и отладить их.

Машинное обучение

- на основании имеющихся данных(примеров решения данной задачи) подобрать метод обучения и подходящий алгоритм, обучить его и проверить на тестовых данных. То есть машина сама напишет себе программу по примерам из данных.

Намеки на базовые понятия

Типы задач и методы обучения

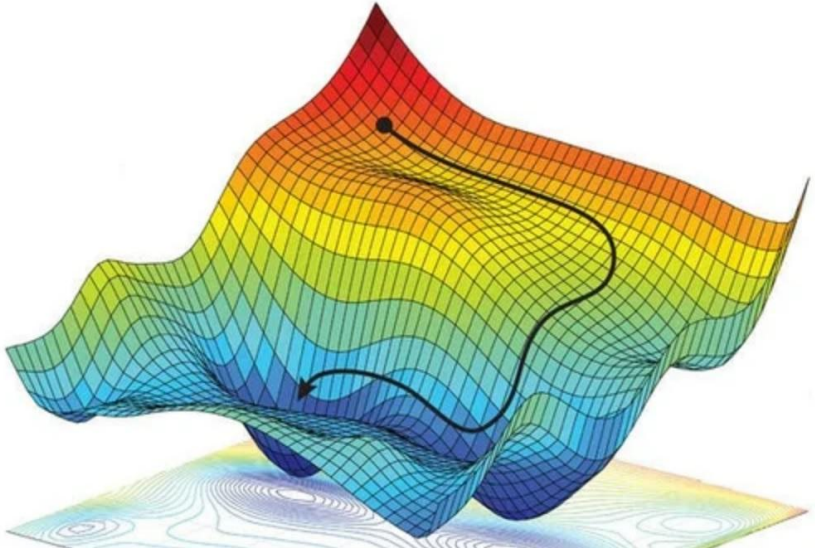
На примерах

Распознавание цифр относят к задаче **классификации обучение с учителем**. То есть имеем тренировочный набор рукописных циф, в котором каждая цифра соответствует своему класс и роль учителя заключается в том, что он соотнес каждому изображению цифры ее класс.

Программа генерации лиц представляет собой отработку метода анализа принципиальных компонент (**PCA**) **обучения без учителя** на чьем-то наборе студенческих фотографий. То есть автор “скормил” компьютеру набор фото студентов и попросил его разложить их самому по 80ти “полочкам”. Первое что бросилось в глаза методу - цвет футболки.

CRISP-DM

Стандартизованный жизненный цикл систем интеллектуальной обработки данных

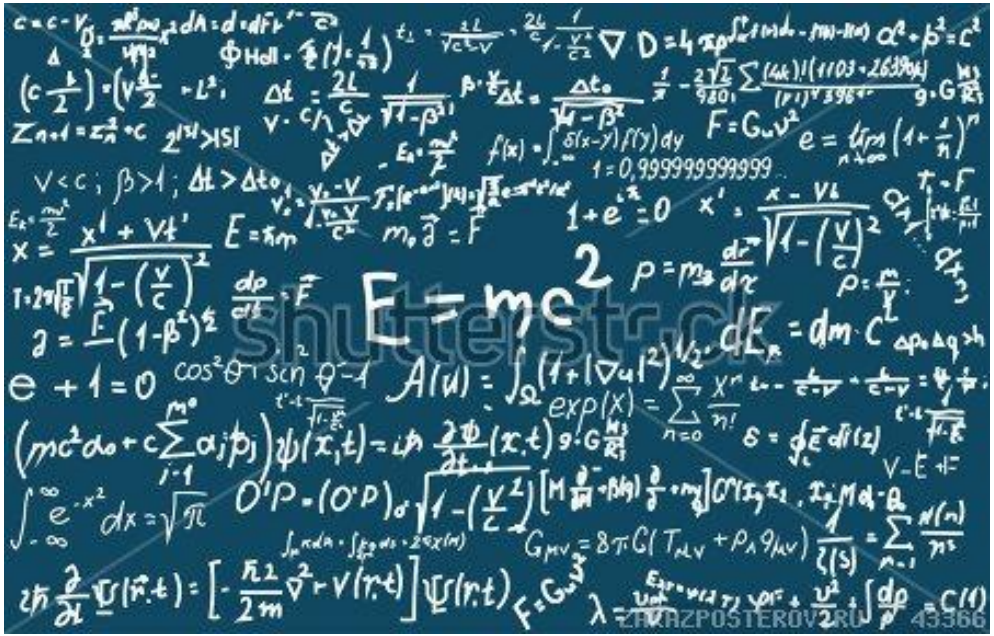


Какие бы задачи не решались методами машинного обучения, они проходят через 3 стадии:
-представление;
-оценка;
-оптимизация

Оптимизационный метод “градиентный спуск” оценивает данные представленные моделью по косвенной “функции ошибок” и выдает информацию о том, как изменить модель так, чтобы функция ошибок спустилась в тот минимум, который нас устроит.

Школы познания

Символисты

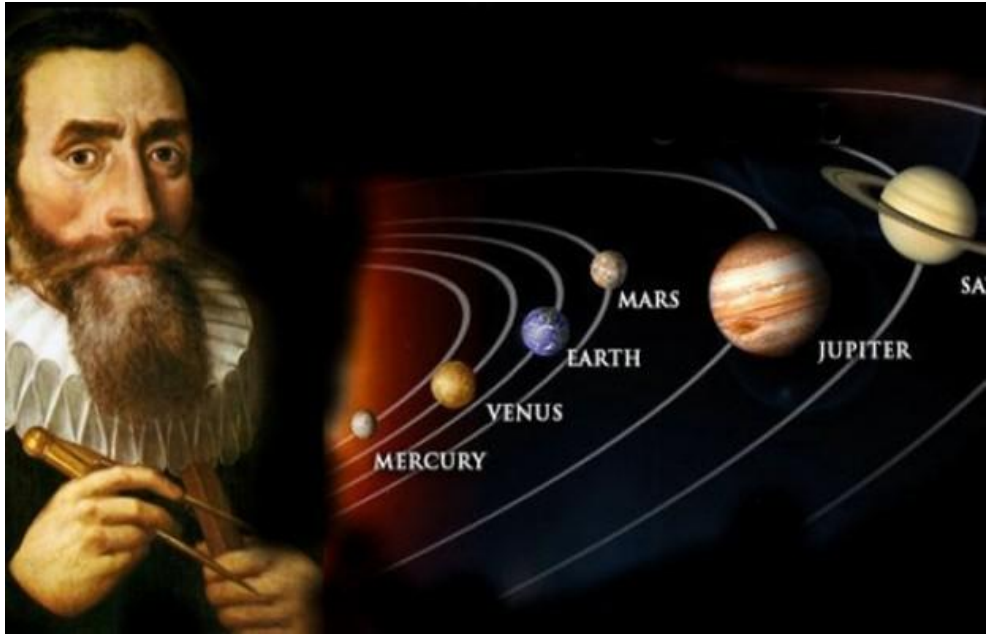


Получают знание применяя различные математические методы к известному аксиоматическому базису. Основное допущение - равнозначность явления и символа его обозначающего.

Стадии Браге, Кеплера и Ньютона

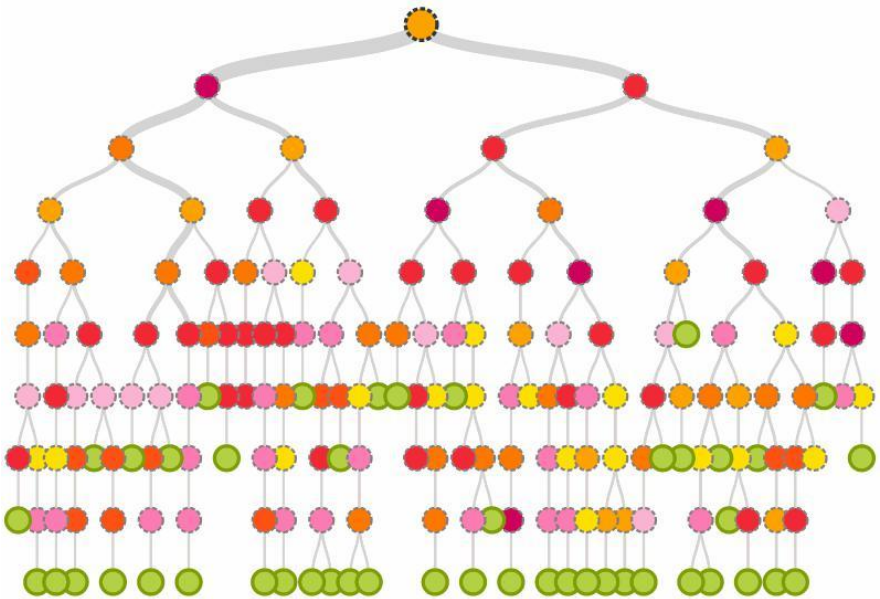
Тихо Браге

значительную часть жизни собирал астрономические данные достаточной точности о движении планет



Кеплер находит математические закономерности в данных Браге

Ньютон на основании найденных Кеплером закономерностей выводит известный аксиоматический базис

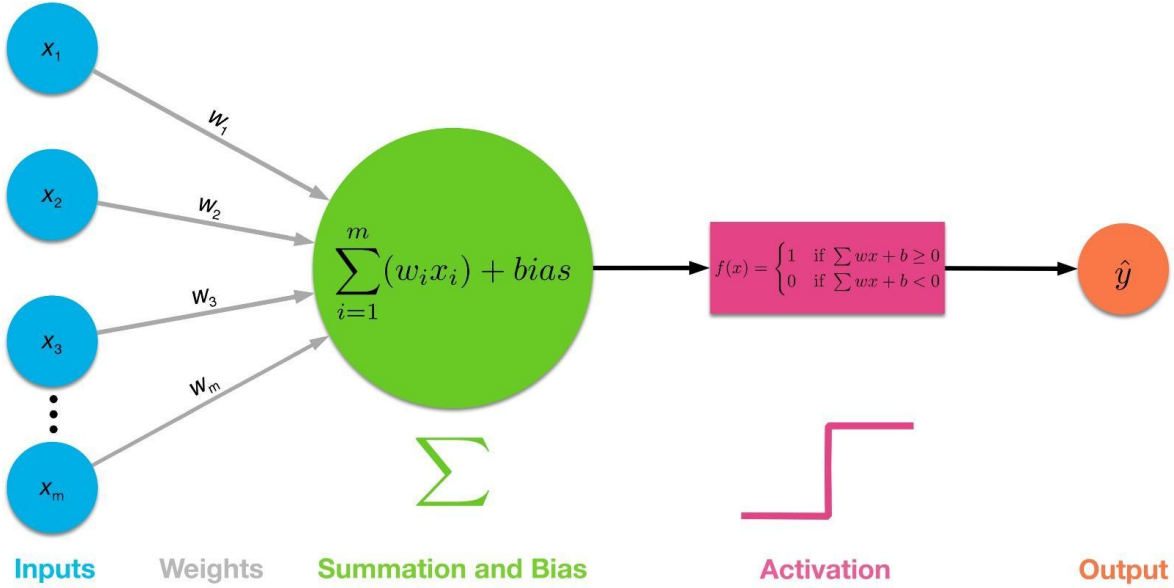


Основные методы:
обратная дедукция и решающие деревья

scikits learn
machine learning in Python

Коннекционисты

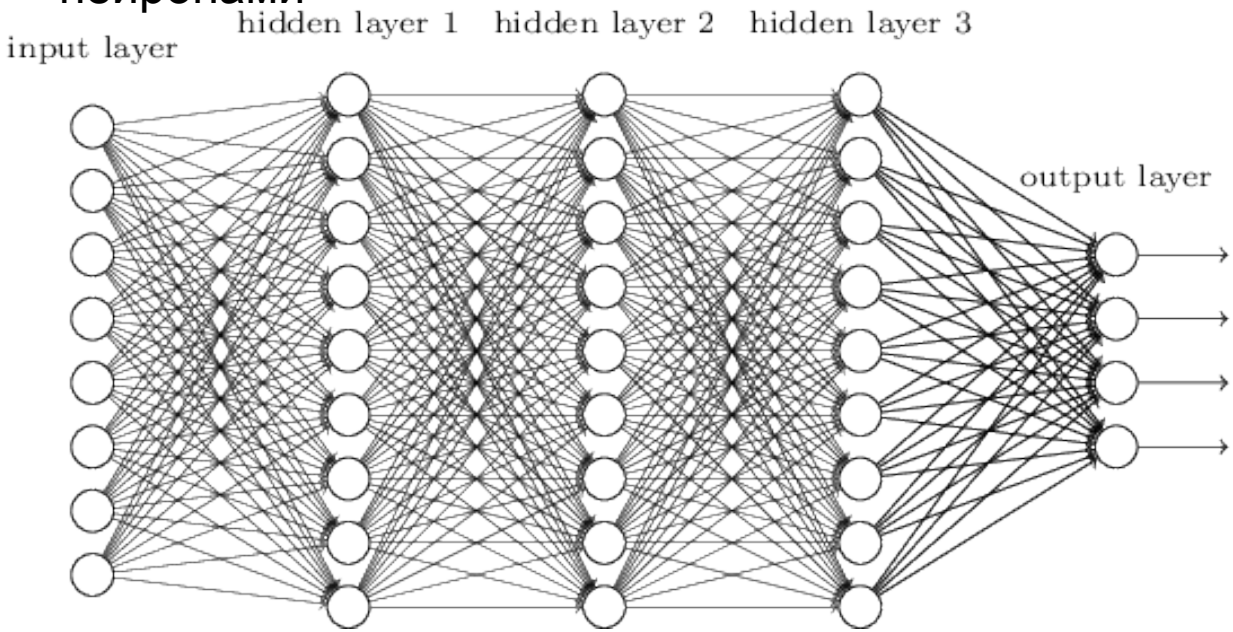
Перцептрон



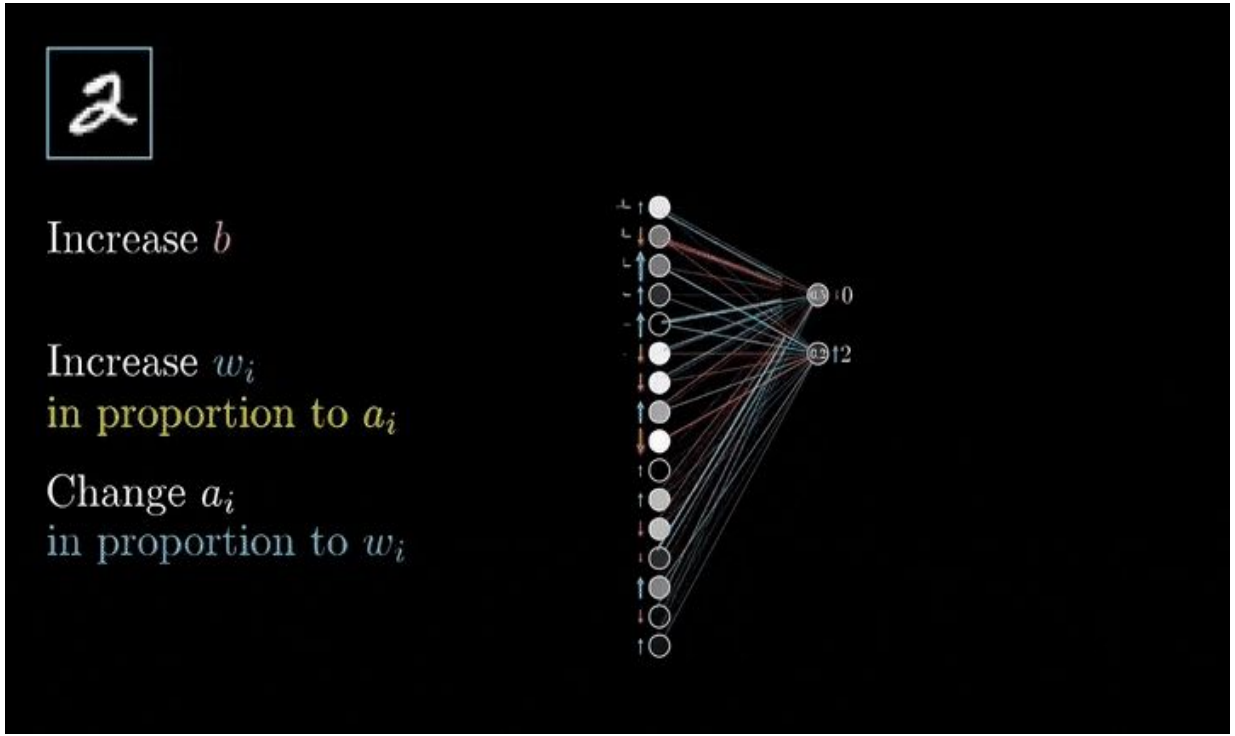
Эта школа вдохновляется данными о том как работает мозг. Как он строит знание в реальных условиях

Многослойный Перцептрон

Нейронная сеть хранит знание в связях между нейронами



Обратное распространение ошибки

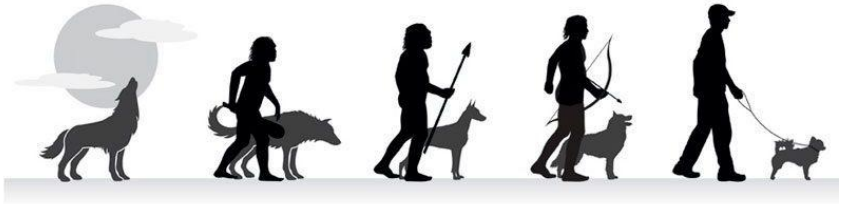


Основной метод построения знания в нейронных сетях, позволяющий на основе знания о том, как сильно ошиблась модель перестраивать все веса модели начиная с выходного

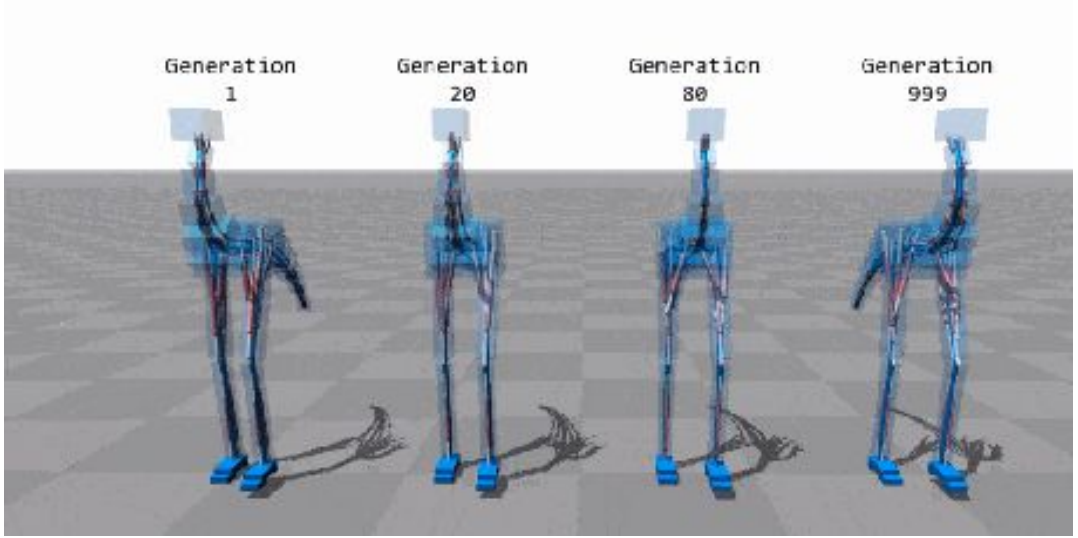
TensorFlow

Keras

Эволюционисты



Эволюция, как метод получения нового знания берет свои истоки из наблюдений за развитием живого.
В частности за тем, как человек сам стал влиять на развитие организмов его окружающих. Агрокультура издавна отбирала растения с самыми вкусными плодами, овец - с самой длинной шерстью. Одомашненные кошки, собаки, рыбки, кролики, попугаи и прочее также следствие эволюции с рукотворным отбором.



Эволюционисты рассматривают программы (алгоритм), как набор обращений к подпрограммам. Для решения конкретной задач они создают популяции программ и оценивают их способность к решению данной задачи с помощью **функции приспособленности.** Отобрав лучшие версии программ они “скрещивают” их в случайных местах вызовов подпрограмм и таким образом получают новую популяцию для следующей эпохи отбора.

DISTRIBUTED EVOLUTIONARY ALGORITHMS IN PYTHON

Сильнейшей стороной эволюционных алгоритмов является широчайший охват “пространства гипотез”, в котором каждая версия алгоритма прорабатывает свою версию ответа.
Также следует отметить взаимосвязь эволюции и коннекционизма.
Эволюционный рост ассоциативных зон коры головного мозга основан на нейронном обучении в сенсорных зонах - без этого он был бы бесполезен.
Эволюция усиливает в потомках те свойства, что в наибольшей мере помогли предкам выжить и размножиться..

Байесианцы

Naïve Bayes Classifier

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Thomas Bayes
1702 - 1761

Томас Байес(1702-1761) - британский математик, священник, член лондонского королевского общества. Сформулировал правило обновления уровня доверия к гипотезе при получении новых свидетельств.
Сама теорема была опубликована Лапласом спустя 10 лет после смерти Байеса, т. к. он посчитал ее недостойной публикации, но к счастью оставил ее в своих записях.

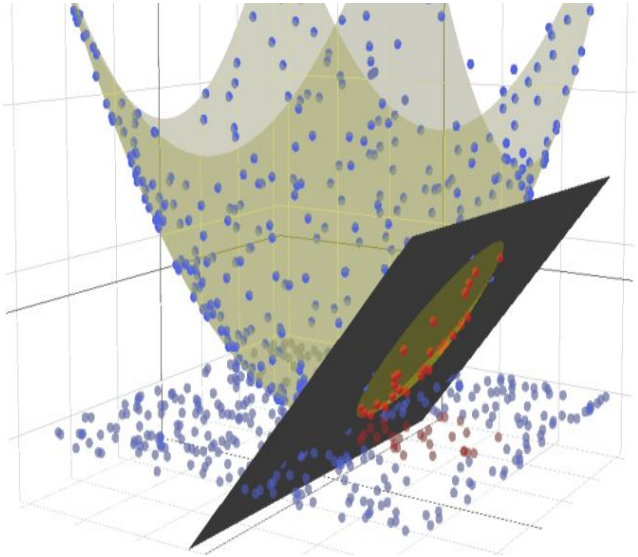
Проиллюстрируем работу теоремы на примере диагностики заболеваний.
Известно:
Тест на болезнь дает верный результат в 99% случаях заболеваний и дает ложноположительный результат в 1% случаев.
P(A)Распространенность заболевания - 0,1% среди всего населения (приорная вероятность)
P(B|A) - вероятность положительного теста при болезни
P(B) - вероятность положительного срабатывания теста = [P(A)*P(B|A)+P(-H)*P(E|-H)]
Тогда
P(A|B) - вероятность болезни в случае положительного теста составит примерно 9%.
Неочевидность результата объясняется игрой вероятностей - из 1000 человек будет болен 1 и тест это покажет(на 99%), но так же тест даст ложноположительное срабатывание для 10 человек, т.е. в результате тест из 1000 раз сработает 11 и только 1 из них будет болен, что и соответствует вероятности в 9%

Аналогисты

Рассуждения по аналогии - древнейший метод построения знания.
Первое упоминание относят к Аристотелю и его закону подобия “если две вещи схожи, мысль об одной из них будет склонна вызывать мысль о другой.”

Метод ближайших соседей

Джон Сноу — серьёзная вспышка эпидемии холеры, случившаяся в 1854 году в Лондоне. Событие вошло в историю благодаря методичным действиям доктора Джона Сноу выявившего источник эпидемии — загрязнённую воду из водозаброной колонки. Исследование Сноу послужило толчком к развитию эпидемиологии и совершенствованию систем водоснабжения и канализации.



Метод опорных векторов созданный Владимиром Вапником сотрудником Bell Labs в 1994м году решает задачу разделения классов “проводя аналогии” и секущие гиперплоскости из n+1 мерного пространства

scikits learn
machine learning in Python

VK.COM/AISNZ

school of ai

СНЕЖИНСК