Математические и инструментальные методы машинного обучения

1. Задачи и методологии анализа данных

Большие данные

- Большие данные (Big Data, англ.) представляют собой новое качество обычных данных в электронном виде, накопленных в большом объёме в разнообразных информационных системах, корпоративных или государственных, сайтах, блогах
- Согласно исследованиям IDC, мировой рынок бизнес-аналитики и больших данных планомерно растет: в 2015 г. он достиг $122 млрд., в 2016 г. – уже $130 млрд. К 2020 г. аналитики прогнозируют рост объема рынка до $203 млрд
- Количество накопленных данных для российского рынка в 2020 году составит 980 экзабайт

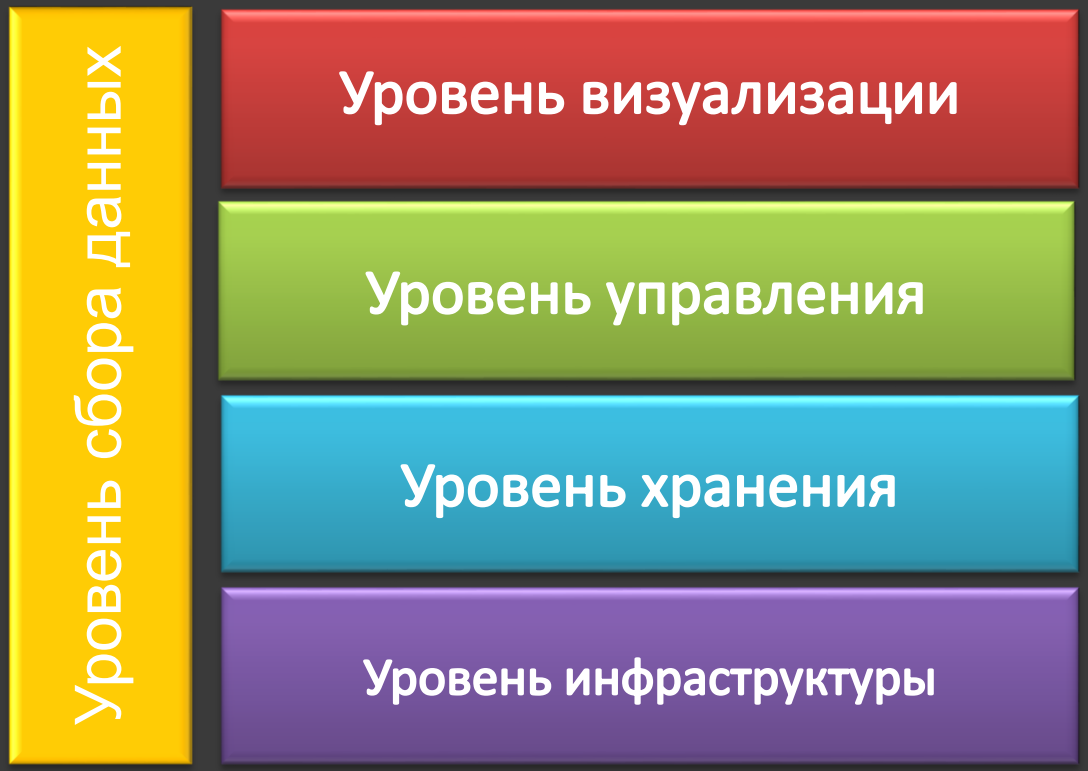
V-модель Больших данных

Доля компаний по отраслям, внедривших Большие данные

Направления развития профессиональной деятельности

- сбор и обработка больших данных
- аналитика
- инженерия больших данных
- архитектура больших данных и системная интеграция
- разработка продуктов и услуг на основе больших данных
- управление большими данными и системами на основе больших данных
- проведение исследований с целью получения новых математических и технических решений для работы с большими данными

Уровни архитектуры по обращению с Большими данными



Уровень сбора данных

Этот уровень отвечает за отделение шума от соответствующей информации, а также регулирование объема, скорости и разнообразия данных. Он должен иметь возможность проверять, очищать, преобразовывать, уменьшать и интегрировать данные в стек технологий больших данных для дальнейшей обработки. Это новое программное обеспечение, которое должно быть масштабируемым, устойчивым, отзывчивым и регулирующим в архитектуре больших данных.

Уровень инфраструктуры

На данном уровне располагается физическая инфраструктура, необходимая для функционирования и масштабируемости архитектуры больших данных. Фактически наличие надежной и недорогой физической инфраструктуры привело к появлению таких важных тенденций, как big data. Для поддержки непредвиденного или непредсказуемого объема, скорости или разнообразия данных физическая инфраструктура для больших данных должна отличаться от инфраструктуры для традиционных данных.

Уровень хранения

Использование массового распределенного хранилища и обработки является фундаментальным изменением в способе обработки больших данных предприятием. Распределенная система хранения данных обещает отказоустойчивость, а распараллеливание позволяет высокоскоростным алгоритмам распределенной обработки выполнять крупномасштабные данные. Для работы с «большими данными» используется несколько реализаций распределенных файловых систем. Основными из них можно считать реализацию от open-source проекта Hadoop (Hadoop Distributed File System - HDFS) и реализацию от Google (Google File System - GFS).

Уровень управления и обработки

На уровне управления и обработки находятся инструменты и языки запросов для доступа к базам данных NoSQL с помощью файловой системы хранения HDFS, находящейся поверх уровня физической инфраструктуры Hadoop. С развитием вычислительной техники, теперь можно управлять огромными объемами данных, которые ранее могли бы быть обработаны только суперкомпьютерами за большие деньги. Цены на системы (ЦП, ОЗУ и диск) упали. В результате, новые методы для распределенных вычислений стали основным направлением.

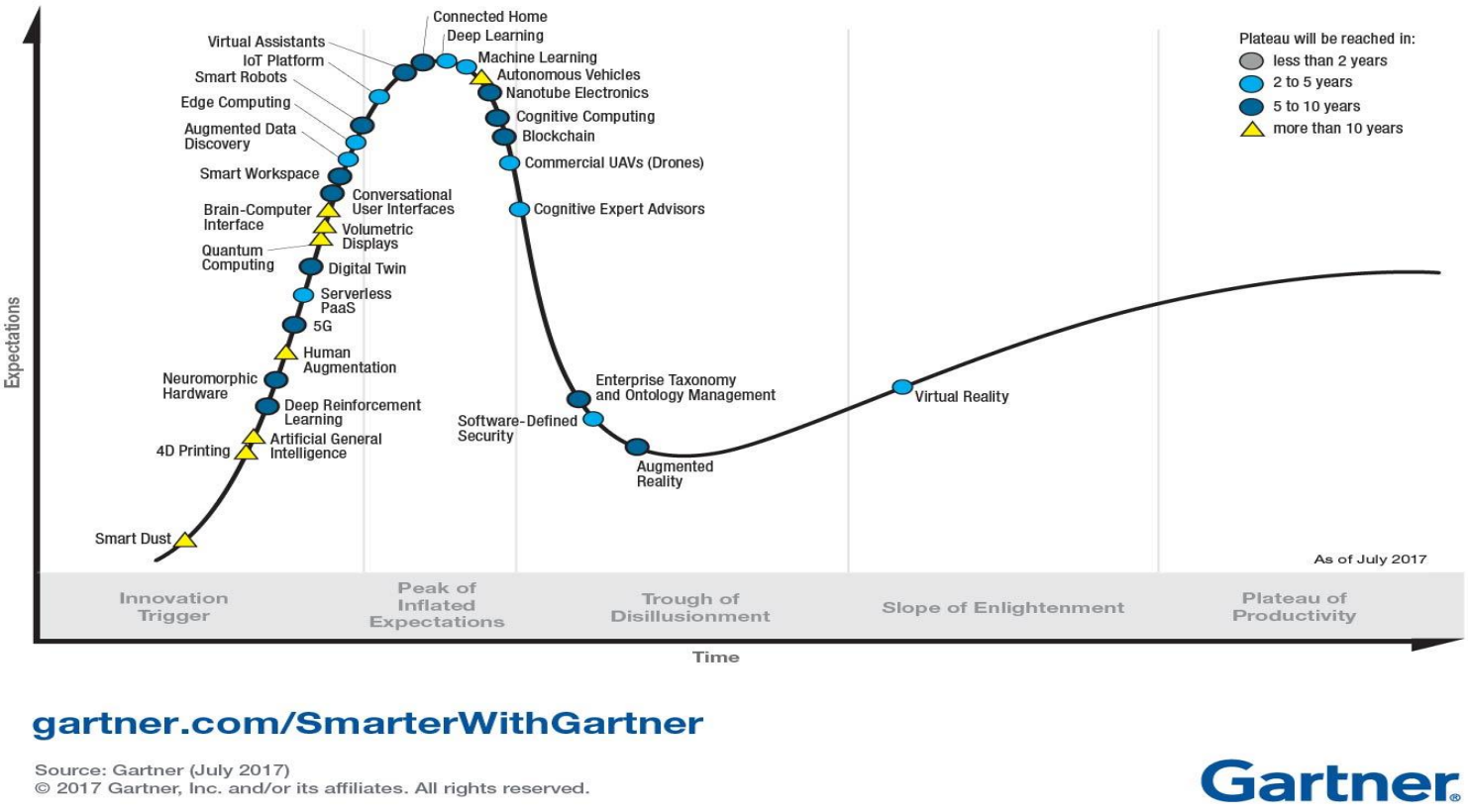
Угрозы и риски использования Больших данных

- риск конфиденциальности
- риск снижения эффективности больших данных
- риск неготовности к переменам
- риск потери данных
- риск формирования неэффективного набора данных
- риск внешнего консультанта
- риск переполнения хранилища
- риск мошенничества
- риск экономической нецелесообразности
- риск ошибок больших данных
- риск ошибок бизнес-модели

Снижение риска ошибок Больших данных

- проводить периодические ревизии данных
- контролировать ключевые параметры данных
- вести журнал выявленных ошибок и их устранения
- разрабатывать инструменты и алгоритмы устранения или нивелирования ошибок и некорректных состояний данных
- оценивать результативность инструментов
- проводить независимую оценку и экспертизу
- применять специальные средства тестирования данных и инструментов, которые разрабатываются самостоятельно
- использовать инструменты последовательно, подконтрольно и пошагово с постоянным контролем обрабатываемых данных в целом или по выборкам

Что происходит в мире?



Искусственный интеллект

Иску́сственный интелле́кт (англ. *Artificial intelligence (AI)*)— раздел информатики, изучающий возможность обеспечения разумных рассуждений и действий с помощью вычислительных систем и иных искусственных устройств. При этом в большинстве случаев заранее неизвестен алгоритм решения задачи. Обычно к реализации интеллектуальных систем подходят именно с точки зрения моделирования человеческой интеллектуальности.

Искусственный интеллект

Таким образом, в рамках искусственного интеллекта различают два основных направления:

- символьное (семиотическое, нисходящее) основано на моделировании высокоуровневых процессов мышления человека, на представлении и использовании знаний;
- нейрокибернетическое (нейросетевое, восходящее) основано на моделировании отдельных низкоуровневых структур мозга (нейронов).

Таким образом, сверхзадачей искусственного интеллекта является построение компьютерной интеллектуальной системы, которая обладала бы уровнем эффективности решений неформализованных задач, сравнимым с человеческим или превосходящим его. В качестве критерия и конструктивного определения интеллектуальности предложен мысленный эксперимент, известный как тест Тьюринга.

Тест Тьюринга

Тест Тьюринга — эмпирический тест, идея которого была предложена Аланом Тьюрингом в статье «Вычислительные машины и разум» (англ. Computing Machinery and Intelligence), опубликованной в 1950 году в философском журнале «Mind». Тьюринг задался целью определить, может ли машина мыслить.

Стандартная интерпретация этого теста звучит следующим образом: «Человек взаимодействует с одним компьютером и одним человеком. На основании ответов на вопросы он должен определить, с кем он разговаривает: с человеком или компьютерной программой. Задача компьютерной программы — ввести человека в заблуждение, заставив сделать неверный выбор».

Тест Тьюринга

Все участники теста не видят друг друга. Если судья не может сказать определенно, кто из собеседников является человеком, то считается, что машина прошла тест. Чтобы протестировать именно интеллект машины, а не её возможность распознавать устную речь, беседа ведется в режиме «только текст», например, с помощью клавиатуры и экрана (компьютера-посредника). Переписка должна производиться через контролируемые промежутки времени, чтобы судья не мог делать заключения исходя из скорости ответов. Во времена Тьюринга компьютеры реагировали медленнее человека. Сейчас это правило необходимо, потому что они реагируют гораздо быстрее, чем человек.

Конвенционный и Вычислительный ИИ

Можно выделить две научные школы с разными подходами к проблеме ИИ: Конвенционный ИИ и Вычислительный ИИ. В конвенционном ИИ главным образом используются методы машинного самообучения, основанные на формализме и статистическом анализе. Методы конвенционного ИИ:

- Экспертные системы: программы, которые действуя по определенным правилам, обрабатывают большое количество информации, и в результате выдают заключение на её основе.
- Рассуждение на основе аналогичных случаев (Case-based reasoning).
- Байесовские сети.
- Поведенческий подход: модульный метод построения систем ИИ, при котором система разбивается на несколько сравнительно автономных программ поведения, которые запускаются в зависимости от изменений внешней среды.

Конвенционный и Вычислительный ИИ

Вычислительный ИИ подразумевает итеративную разработку и обучение (например, подбор параметров в сети связности). Обучение основано на эмпирических данных и ассоциируется с не-символьным ИИ и мягкими вычислениями. Основные методы:

- Нейронные сети: системы с отличными способностями к распознаванию.
- Нечёткие системы: методики для рассуждений в условиях неопределенности (широко используются в современных промышленных и потребительских системах контроля)
- Эволюционные вычисления: здесь применяются понятия традиционно относящиеся к биологии такие как популяция, мутация и естественный отбор для создания лучших решений задачи. Эти методы делятся на эволюционные алгоритмы (например, генетические алгоритмы) и методы роевого интеллекта (например, муравьиный алгоритм).

Направления разработок в области искусственного интеллекта

Машинное обучение	Обработка текстов на естественном языке	Машинный перевод	Рекомендательные системы
Представление знаний	Глубокое обучение	Распознавание изображений	Визуализация
Имитационное моделирование	Робототехника	Интернет вещей	Нейросетевые технологии

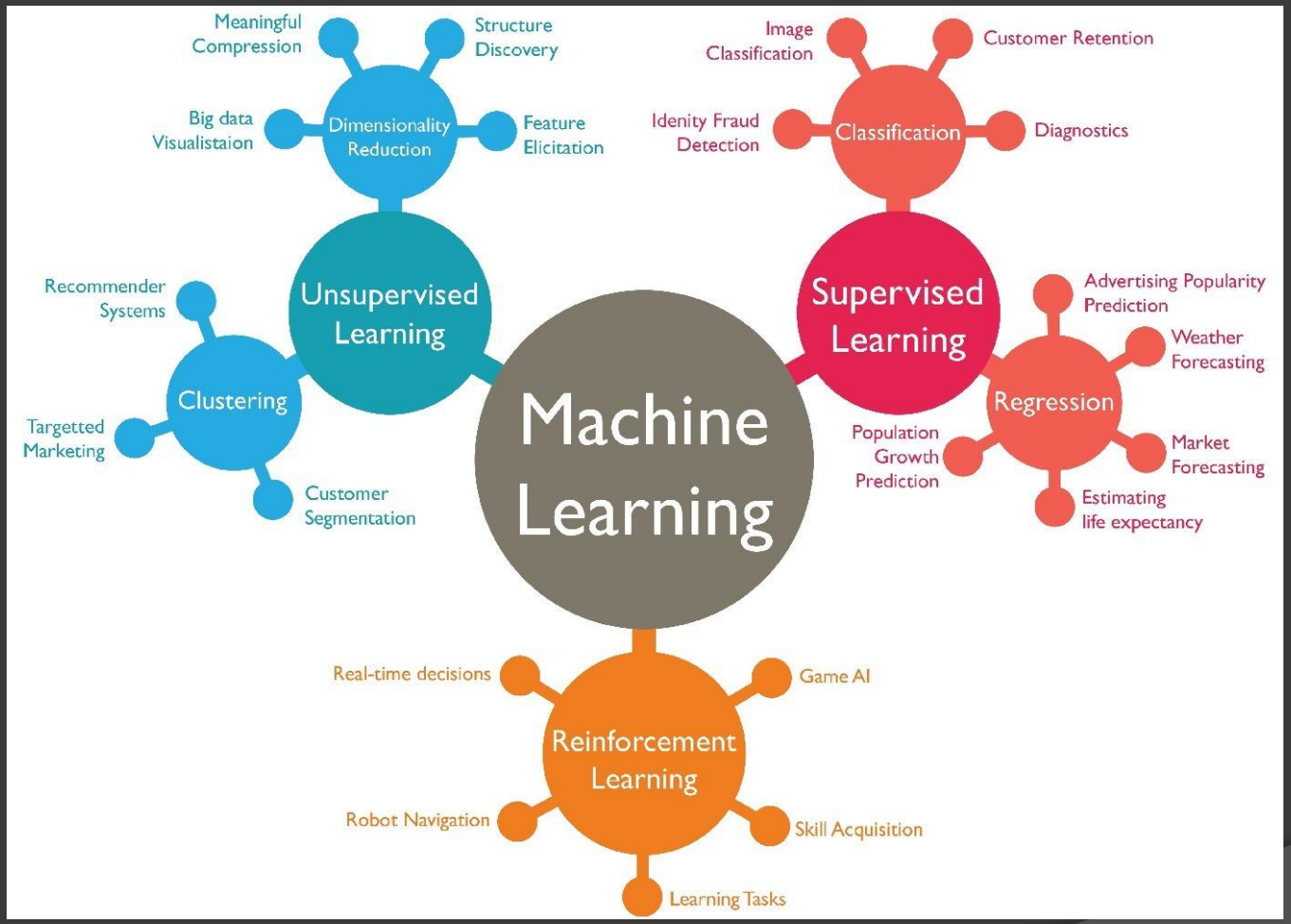
Машинное обучение

- Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Различают два типа обучения.
- Обучение по прецедентам, или индуктивное обучение, основано на выявлении общих закономерностей по частным эмпирическим данным. Дедуктивное обучение предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний. Дедуктивное обучение принято относить к области экспертных систем, поэтому термины машинное обучение и обучение по прецедентам можно считать синонимами.

Машинное обучение

- Машинное обучение находится на стыке математической статистики, методов оптимизации и классических математических дисциплин, но имеет также и собственную специфику, связанную с проблемами вычислительной эффективности и переобучения. Многие методы индуктивного обучения разрабатывались как альтернатива классическим статистическим подходам. Многие методы тесно связаны с извлечением информации и интеллектуальным анализом данных (Data Mining).

Зачем все это нужно?



Как это использовать?