MСистемы Litepaper

Проблема: Быстрое создания моделей ИИ

Увеличение спроса на модели ИИ, стимулированное развитием глубокого обучения, привело к необходимости быстрого создания, итеративного улучшения и масштабного внедрения моделей машинного обучения — как в больших, так и в малых масштабах. Современный подход в стиле «больших данных», применяемый крупными технологическими компаниями, заключается в объединении наборов данных для обучения систем глубокого обучения на огромных вычислительных кластерах, что привело к стремительному росту затрат на этот процесс. К примеру, обучение моделей, таких как ChatGPT, обходится в десятки миллионов долларов, а стоимость их поддержки превышает $100 тыс. в день. Эти огромные вычислительные требования, наряду с ограниченной интерпретируемостью и недостаточной надёжностью систем глубокого обучения, ограничили их применение в ряде научных и промышленных областей, где доминируют более простые, интерпретируемые модели, основанные на статистическом машинном обучении.

Такие модели поддерживают работу аналитика в системе принятия решений, не пытаясь заменить его, и создаются, внедряются и поддерживаются относительно небольшими командами. Эти системы намного более многочисленны и охватывают широкий спектр секторов, часто создаваясь с нуля под конкретную область применения. Однако, как и в случае с системами глубокого обучения, сложной задачей остаётся проектирование, разработка и развертывание таких систем. Для успешного проектирования, разработки, развертывания и поддержки таких систем компаниям требуются команды специалистов с разнообразными навыками — от экспертов в конкретных прикладных областях до математиков, специалистов по данным, инженеров DevOps и инженеров-электронщиков. Частично это происходит потому, что существует широкий выбор алгоритмов, применимых для решения конкретной задачи и отсутствуют информационные платформы типа САПР - системы автоматизированного проектирования для ИИ, позволяющие их быстро и качественно разрабатывать.

Цель: Платформа для интерпретируемых моделей ИИ, разработанная людьми для людей при поддержке машин

МСистемы стремится стать платформой для проектирования и развертывания моделей ИИ и ведущим поставщиком инструментов для создания, проектирования и развертывания моделей статистического машинного обучения. Это позволит отдельным специалистам, небольшим командам и крупным корпорациям ускорить разработку моделей ИИ в различных областях — от научных исследований и разработок до критической инфраструктуры с повышенными требованиями к безопасности. Мы достигнем этого, предоставляя интегрированную цепочку инструментов и технологий, которая позволит кардинально ускорить разработку интерпретируемый моделей искуственного интеллекта и на их основе синтезировать алгоритмы машинного обучения и низкоуровневый код, который впоследствии может быть развернут на различных вычислительных платформах — от маломощных встроенных устройств до высокопроизводительных вычислительных кластеров. Таким образом, мы намерены кардинально ускорить процесс проектирования и развертывания моделей, минимизируя затраты времени и ресурсов, необходимые для внедрения передовых аналитических систем.

Задача: Ускорение разработки и развертывания моделей

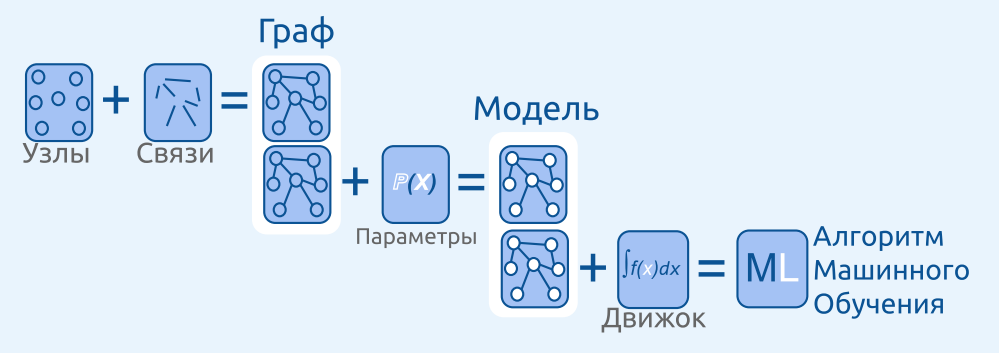
Хотя рост спроса на модели ИИ очевиден, практическая разработка таких моделей для конкретных приложений требует экспертных знаний в широком диапазоне дисциплин — от специалистов в области применения до статистиков, специалистов по данным, инженеров DevOps и, в некоторых случаях, инженеров по ПЛИС и электронике. Для успеха такие команды должны эффективно взаимодействовать между собой в междисциплинарном контексте для правильного определения требований и возможностей системы. После, следует длительный процесс проектирования и реализация компонентов, который завершается интеграцией отдельных частей в единую систему, формирующую конечный продукт с индивидуально подобранным алгоритмом и инфраструктурой для его поддержки. Этот процесс является трудоемким, подверженным ошибкам и связан с существенными рисками для заказчика.

Решение

Чтобы кардинально ускорить процесс разработки и развертывание моделей ИИ, необходимо отделить спецификацию модели от алгоритмов, которые выполняют вычисления и обработку данных. Это позволит разработчикам модели, доменных экспертам быстро вносить изменения в модель, не беспокоясь о низкоуровневых деталях реализации. Паралельно, работа инженеров над улучшением и тестированием универсальных движков для выполнения вычислений сможет принести пользу сразу всем разработанным прикладным моделям, не только одной конкретной модели.

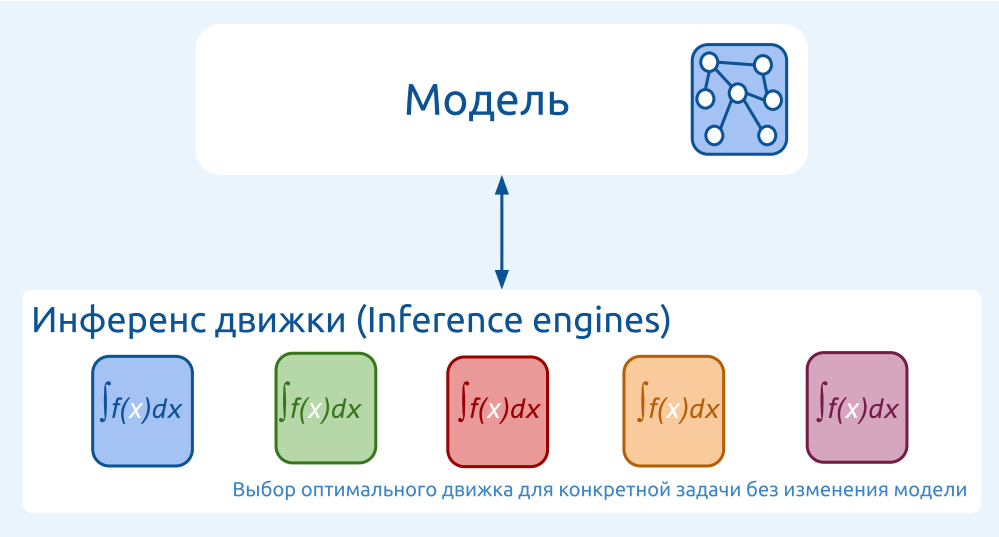
Вероятностное графовое моделирование

Мы можем достичь этого, используя язык вероятностных графовых моделей, которые позволяет легко формировать сложные многомерные вероятностные распределения с помощью графовой структуры, где узлы обозначают переменные, а ребра кодируют взаимозависимости между ними. Обрабатывая заданную вероятностную графическую модель с помощью вычислительного движка можно синтезировать алгоритм машинного обучения.



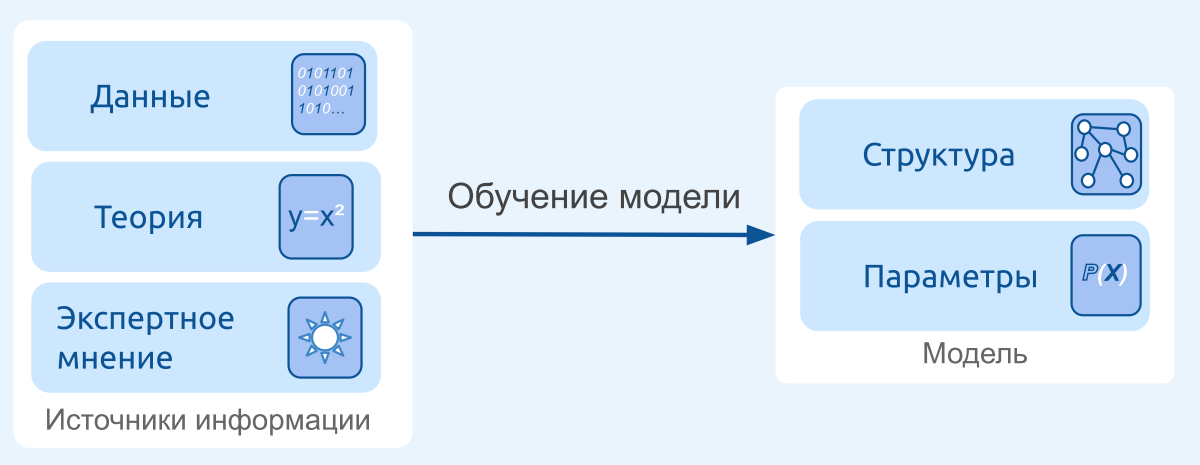
Синтез алгоритма машинного обучения из вероятностной графовой модели (модель).

Структура модели в форме графа позволяет пользователям визуализировать внутреннюю работу модели, а также увидеть предположения, на которых она основана. Если какое-либо предположение оказывается неподходящим, его можно изменить, обновив граф. Это автоматически приводит к генерации нового, обновленного алгоритма машинного обучения.



Разделение модели и движка позволяет выбрать движок под конкретную задачу

Кроме модификации самой модели, мы можем также оптимизировать вычислительный движок, чтобы адаптировать итоговый алгоритм машинного обучения к конкретному приложению и типу запросов. Это позволяет развертывать одну и ту же модель на различных вычислительных платформах и в различных условиях эксплуатации — от маломощных устройств в формате интернета-вещей до систем реального времени и систем с высокими требованиями к надежности.



Модель можно обучить как из данных, так и из других источников знаний

В отличие от традиционных систем глубокого и машинного обучения, вероятностные графические модели являются интерпретируемыми. Кроме того, вероятностная графическая модель может быть обучена с помощью комбинации методов, которые невозможно использовать в обычном машинном обучении и глубоких методах, включая использование субъективного экспертного мнения, теоретических результатов в виде инвариантов и ограничений, а также как «больших», так и «малых» данных. Последний момент представляет особый интерес для областей науки и техники, где наборы данных

трудно или дорого получить или они по определению недостаточно велики для адекватного обучения систем глубокого обучения.

Графический low-code редактор - mBayes

Первой составной частью решения является предоставление пользователям low-code редактора (метод создания приложений с помощью визуальных конструкторов), который позволит создавать и редактировать модели, а также проводить вычисления в ходе начальных этапов проекта. Разрабока прототипа в low-code среде позволит экспертам в области применения работать непосредственно с моделью. В то же время возможность запускать различные алгоритмы вывода (инференс) и генерировать код на основе модели ускорит последующие этапы разработки, позволяя разработчикам экспериментировать с различными конфигурациями, прежде чем вкладывать значительные ресурсы в реализацию решения в программном коде.

Редактор позволит быстро разработать и запустить прототип

Программные модули вероятностного моделирования - mCore

Для поддержки low-code редактора, а также предоставления продвинутым пользователям и разработчикам инструментов для проектирования моделей и выполнения вычислений, мы разработали программный пакет *mCore*. Он поддерживает пользователей на всех этапах разработки вероятностных моделей, включая обучение структуры и модули калибровки параметров, а также предоставляет набор вычислительных движков для синтеза моделей машинного обучение и модулей для генерации кода.

Система позволит разделить задачи по созданию и реализации модели между разными членами команды

В сочетании с low-code редактором *mBayes*, программный пакет *mCore* позволяет командам разделить этапы прототипирования и реализации моделей. Это даёт возможность различным командам сосредоточиться на задачах в рамках их компетенций, обеспечивая быструю итерацию и ускоренную разработку моделей.

Это достигается использованием вероятностной модели в качестве единого представления, которое могут использовать как эксперты в предметной области для проведения анализа, настройки и улучшения модели, так и разработчики, которые могут программно манипулировать графовой структурой данных, работая с экспортированным кодом или напрямую с вероятностными распределениями для переноса разработанной модели на компьютерную архитектуру, подходящую для конкретного приложения.

Запуск модели в облачном сервисе поможет быстро разработать прототип модели, а синтез кода ускорит реализацию в железе.