Титульный лист материалов по дисциплине

(заполняется по каждому виду учебного материала)

ДИСЦИПЛИНА	**Технологии извлечения знаний из больших данных** (полное наименование дисциплины без сокращений)
ИНСТИТУТ	ИКБ
КАФЕДРА	**КБ-4 «Интеллектуальные системы информационной безопасности»** полное наименование кафедры)
ВИД УЧЕБНОГО МАТЕРИАЛА	**Лекция** (в соответствии с пп.1-11)
ПРЕПОДАВАТЕЛЬ	**Никонов В.В.** (фамилия, имя, отчество)
СЕМЕСТР	**3 семестр 2023/2024 уч. года** (указать семестр обучения, учебный год)

1. Интеграция в бизнес-процессы

Говоря о применении ИИ в бизнесе, сегодня чаще всего имеют в виду применение технологий машинного обучения (machinelearning) — именно они позволили достичь высоких результатов в анализе изображений, текстов, в играх и других нетривиальных задачах.

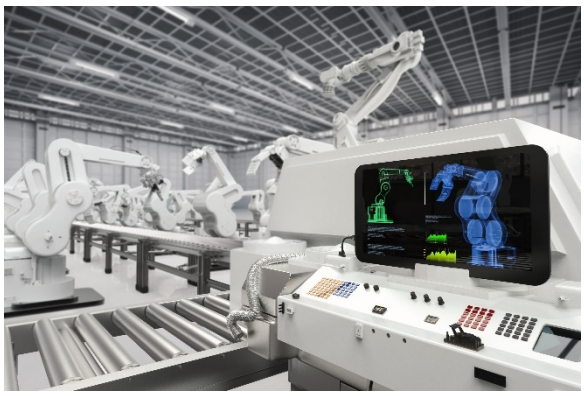
Эти технологии подразумевают извлечение знаний из огромных массивов информации (наборов данных, или по-английски dataset — датасетов).

Принцип работы алгоритма машинного обучения следующий: по большому количеству примеров вида вход — выход настраивают алгоритм, который сможет по входу предсказывать выход. Процесс настройки алгоритма называется обучением (learning).

Разработка способов обучения алгоритмов — это отдельная область прикладной математики на стыке с информационными технологиями. Однако заказчику досконально понимать, как устроены механизмы машинного обучения, не обязательно, главное — понимать, какие задачи эта технология может решать, и на высоком уровне ориентироваться в области. Этому и посвящен наш курс. В данном блоке мы охарактеризуем общие черты бизнес-задач, решаемых в машинном обучении, и обсудим, каковы ключевые компоненты успеха решения этих задач и с какими сложностями можно столкнуться.

Выделяют два основных направления применения: автоматизация и прогнозирование.

Автоматизация



ИИ успешно применяется в задачах, в которых вид входных и выходных данных всегда один и тот же, однако четкого алгоритма получения результата не существует. При этом поток входных данных настолько большой, что для решения задачи приходится нанимать большую команду — ИИ же может избавить от такой необходимости и значительно оптимизировать расходы.

Примеры автоматизации

Например, ИИ успешно сортирует отзывы о продукте, категоризирует их, собирает статистику и выделяет негативные отзывы, требующие срочного ответа. Другой пример — модерация контента на сайтах объявлений: вместо большого отдела модераторов объявления, не соответствующие правилам сервиса, отфильтровывает ИИ. Аналогичные примеры можно найти в такси или общепите: ИИ следит за скоростью водителя или за качеством мытья рук сотрудниками ресторана. Еще один пример — применение ИИ для организации хранения документов и для их обработки. В эту же группу можно отнести чат-боты, заменяющие сотрудников контакт-центра, отвечающих на обращения клиентов в контакт-центр.

Прогнозирование

Речь идет о решении задач, для которых ответ неизвестен, но его можно спрогнозировать на основе исторических данных. Для решения таких задач приходится нанимать опытных специалистов, но человеку нужно много времени, чтобы проанализировать большие объемы информации. ИИ справляется быстрее и качественнее, потому что быстрее выполняет

вычисления, «запоминает» больше зависимостей в данных, кроме того, ИИ оценивает ситуацию объективно и беспристрастно.

Примеры прогнозирования

К примерам относятся решение задачи кредитного скоринга, прогнозирование оттока клиентов, прогнозирование страховых рисков. ИИ успешно применяется для прогнозирования спроса на товары и услуги, например рестораны могут с высокой точностью предсказать количество заказов на следующий день/дни и оптимизировать закупки продуктов, а курьерские службы и такси — выводить на линию оптимальное количество сотрудников. К этому же направлению можно отнести рекомендательные системы — алгоритмы, прогнозирующие, какие товары/фильмы/продукты будут интересны клиенту, и маркетинговые инструменты, позволяющие предсказать, на какой баннер клиент более вероятно кликнет.

2. Кейсы применения

Актуальны следующие задачи:

- Кредитный скоринг

Уже сейчас при принятии решения для подавляющего количества кредитов используется ИИ, а к концу 2022 года ИИ будет выдавать все кредиты в банке. Для этого алгоритм будет анализировать кредитную историю клиента и информацию о его доходах и тратах.

AI Jorney: Рекомендательные системы и системы поддержки принятия решений в кредитном скоринге

- Agro AI

Перед агробизнесом стоит ряд задач:

Как повысить урожайность хозяйств;

Как купить или арендовать хорошее поле;

Как увеличить посевную площадь;

Как оценить спрос и предложение.

Алгоритмы могут анализировать большие данные со спутников или аэросъёмки и прогнозировать урожай, отмечать проблемные участки и влияние погоды. Решение для предприятий агропромышленного комплекса, разработанное Sber AI совместно со СберАналитикой, состоит из четырёх модулей, каждый из которых призван решать задачи определённого уровня:

AI-оценщик поля — проверка поля перед покупкой;

AI-агроном — прогноз урожайности и корректировка сельхозработ;

AI-аналитик — ежемесячный аналитический отчет по регионам РФ;

Мониторинг границ поля — определение фактических границ полей.

- Борьба с мошенничеством

Технологии позволяют более точно и быстро детектировать подозрительные операции и блокировать их. Для создания таких алгоритмов можно использовать как данные о реальных попытках мошенничества, так и строить модели поведения пользователя, основываясь только на «чистых» данных.

- Робот-юрист

Юристы занимаются вычиткой документов и тратят на рутинный процесс большое количество времени.

Алгоритмы машинного обучения находят в тексте имена, адреса, число голосов, наименование юр. лица и связывают эти фрагменты информации друг с другом. Например, робот способен «прочесть» протокол собрания акционеров и понять, какому акционеру какая доля акций принадлежит и набирается ли кворум для каждого вопроса, вынесенного на голосование. Благодаря такому скрупулёзному извлечению значимой информации достигается высокая точность работы робота-юриста.

- Чат-боты

Для упрощения взаимодействия клиента с банком используется автоматический сервис, задавать вопросы которому можно текстом или голосом. Такой чат-бот, конечно, не сможет решить сложные вопросы, но

упростит для клиента поиск ответов на вопросы вида «Как узнать лимиты по данной карте», «Где находится ближайший банкомат» или «Как получить кредит».

AI Jorney: DialoGPT для генерации диалогов.

- Снижение аварийности на транспорте с помощью компьютерного зрения

Пассажирский транспорт мегаполиса входит в группу повышенного риска в связи с внушительным пассажиропотоком, плотностью движения и высокой нагрузкой на водителя. 70% дорожно-транспортных инцидентов происходит из-за потери внимания человеком.

Весь подвижной состав ГУП «Мосгортранс» оборудован программно-аппаратным комплексом «Антисон». Система использует компьютерное зрение для анализа видеопотока в реальном времени непосредственно на устройстве. Обнаружив признаки отвлечения, платформа подает звуковой сигнал водителю и оповещает центр мониторинга.

- Дополнительные кейсы для изучения

ИИ активно применяется в следующих сферах:

- промышленность: настройка оборудования под производство конкретных объектов, автоматическая диагностика оборудования и прогнозирование сбоев, контроль производственных процессов;
- торговля: предсказание спроса, разработка персонализированных программ лояльности;
- медицина: автоматическая диагностика и расшифровка результатов исследований, автоматизация составления медицинских отчетов и рекомендаций пациенту;
- транспорт: прогнозирование спроса, диагностика транспортных средств, маркетинговые компании;
- общепит, телеком и т.д.

Несмотря на всю широту возможностей, на текущей стадии развития технологий ИИ может решать только конкретные задачи. При разработке комплексных «умных» систем каждую такую задачу решают отдельно и собирают все решения вместе в управляющей программе.

3. Условия применения

Машинное обучение построено на данных, без них не получится настроить алгоритм. Конечно, существует альтернатива — экспертные системы, в которых алгоритм предсказания реализуется на основе знаний специалиста (эксперта), однако на практике такие системы обычно показывают значительно более низкий уровень качества, чем алгоритмы машинного обучения. Если речь идет об автоматизации или прогнозировании процесса, который давно существует в компании (колл-центры, кредитный скоринг, прогнозирование спроса), то за это время, скорее всего, накопилось огромное количество данных, которые можно использовать для обучения. Бывает, что необходимых данных в компании нет, например их не сохраняли или же вообще ранее данная задача в компании не решалась. В таких случаях можно попытаться найти открытые данные в интернете или купить/заказать данные. Другая возможная проблема — безопасность: данные могут в компании быть, но выдать доступ разработчикам к ним нельзя. О решении подобных проблем стоит задуматься до начала работы над проектом. Также важен вопрос количества данных: чем их больше, тем точнее работает алгоритм, и чем сложнее задача, тем больше нужно данных.

Данные

Машинное обучение построено на данных, без них не получится настроить алгоритм. Конечно, существует альтернатива — экспертные системы, в которых алгоритм предсказания реализуется на основе знаний специалиста (эксперта), однако на практике такие системы обычно показывают значительно более низкий уровень качества, чем алгоритмы машинного обучения. Если речь идет об автоматизации или прогнозировании

процесса, который давно существует в компании (колл-центры, кредитный скоринг, прогнозирование спроса), то за это время, скорее всего, накопилось огромное количество данных, которые можно использовать для обучения. Бывает, что необходимых данных в компании нет, например их не сохраняли или же вообще ранее данная задача в компании не решалась. В таких случаях можно попытаться найти открытые данные в интернете или купить/заказать данные. Другая возможная проблема — безопасность: данные могут в компании быть, но выдать доступ разработчикам к ним нельзя. О решении подобных проблем стоит задуматься до начала работы над проектом. Также важен вопрос количества данных: чем их больше, тем точнее работает алгоритм, и чем сложнее задача, тем больше нужно данных.

Ресурсы

Обучение моделей машинного обучения часто требует больших вычислительных мощностей. Настроить модель на относительно небольших данных можно на персональном компьютере, но для обработки данных о миллионах клиентов потребуется специальный сервер, причем даже с наличием сервера обработка может занимать дни или недели. Крупные компании часто закупают такие серверы (и нанимают специальную команду для их поддержки), небольшие компании могут арендовать облачные серверы, доступные через интернет.

Специалисты

Для применения машинного обучения нужен опыт, поэтому настраивать алгоритм лучше доверить специалисту в этой области. Крупные компании обычно нанимают собственный штат дата-саентистов (от англ. DataScience — наука о данных, этому термину нет точного короткого перевода в русском языке, и он прижился в транслитерации), также существует множество вариантов для найма сторонней компании. В последнем случае заказчик снабжает исполнителя данными, описывает суть задачи и устанавливает метрики, а исполнитель занимается построением модели (и, возможно, внедрением). С другой стороны, построить несложную

модель на небольших данных может даже новичок, для этого достаточно пройти базовый онлайн-курс. В данной ситуации важно, чтобы данные были однородными, качественными и без ошибок. Также важно понимать, что специалисты в машинном обучении, как правило, не очень хорошо знают прикладную область, и для успеха проекта крайне желательно, чтобы заказчик понимал основные принципы и границы применимости машинного обучения — тогда двум сторонам будет проще найти общий язык и решить возникающие проблемы.

Метрики

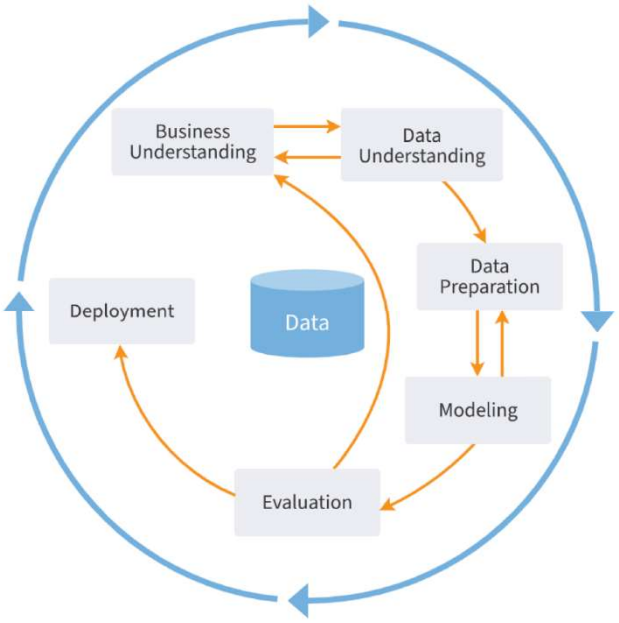
Заказчик должен четко понимать, каковы цели решения задачи, какие метрики необходимо оптимизировать (точность предсказания, удовлетворенность клиентов, доход) и какие значения метрики приемлемы с точки зрения бизнеса. Бывает, что достижение хороших показателей на метриках машинного обучения не приводит к прорыву в бизнесе. Кроме того, достижение желаемого уровня метрик может быть слишком дорогим (требовать работы команды в течение нескольких месяцев).

Согласно некоторым данным, около половины проектов, связанных с искусственным интеллектом, терпят неудачу, среди причин выделяют нереалистичные ожидания, нехватку данных и проблемы с данными.

К сожалению, даже в случае выполнения указанных пунктов (имеется большое количество данных, ресурсов, и алгоритм настроен максимально качественно), проект может завершиться неуспехом, не достигнув желаемых значений метрик качества. В этом случае говорят об отсутствии зависимости в данных: когда по входу невозможно предсказать выход. Такие случаи возникают, например, на бирже (значения индексов зависят от множества факторов, и их крайне сложно предсказать). Однако даже в биржевой торговле сегодня активно применяются алгоритмы ИИ.

4. Методология управления проектами по анализу данных

В завершение перечислим основные шаги решения задачи машинного обучения. Для этого используют популярную методологию CRISP-DM, Схема этапов работы над проектом на рисунке.



Бизнес анализ
(Business Understanding)

Анализ данных
(Data Understanding)

Подготовка данных
(Data Preparation)

Моделирование
(Modeling)

Оценка решения
(Evaluation)

Внедрение
(Deployment)

Итак, выделяют шесть этапов решения бизнес-задачи с использованием алгоритмов искусственного интеллекта:

1. понимание бизнес-целей;
2. сбор и изучение данных;
3. подготовка данных;
4. обучение и настройка алгоритма;
5. тестирование и оценка качества работы алгоритма;
6. внедрение разработанного алгоритма.

В данном курсе мы уделим особое внимание шагам 4 и 5. Более конкретно мы изучим:

- стандартные постановки задач искусственного интеллекта, чтобы слушателям было проще формализовать свои бизнес-задачи в виде задач, для которых существуют успешные технологии решения;
- основные идеи популярных алгоритмов машинного обучения, чтобы слушатели могли выбирать подходы для решения своих задач и анализировать результаты;

- наиболее распространенные метрики машинного обучения, чтобы слушатели могли оценить качество решения бизнес-задач и правильно поставить цели в проектах, связанных с искусственным интеллектом.