

Системы искусственного интеллекта

«02.03.03 - Математическое обеспечение и администрирование информационных систем
направленность разработка и администрирование информационных систем»

<http://vikchas.ru>

Тема 2. Введение в искусственный интеллект и машинное обучение Лекция 5 «Внедрения искусственного интеллекта в бизнес»

Часовских Виктор Петрович

д-р техн. наук, профессор кафедры ШИиКМ

ФГБОУ ВО «Уральский государственный экономический
университет»

Екатеринбург 2022

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 организации нет, например их не сохраняли или же вообще ранее данная задача в не решалась. В таких случаях можно попытаться найти открытые данные в интернете или купить/заказать данные.

Другая возможная проблема — **безопасность**: данные могут в организации 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 организации часто закупают такие серверы (и нанимают специальную команду для их поддержки), небольшие организации могут арендовать облачные серверы, доступные через Интернет.

Специалисты

Для применения **машинного обучения** нужен опыт, поэтому настраивать алгоритм лучше доверить специалисту в этой области.

Крупные организации обычно создают собственный штат Data Science специалистов (от англ. Data Science — **наука о данных**, этому термину нет точного короткого перевода в русском языке, и он прижился в транслитерации).

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

Необходимо четко понимать, каковы цели решения задачи, какие **метрики** (*количественные показатели*) необходимо оптимизировать (точность предсказания, удовлетворенность клиентов, доход) и какие значения метрики приемлемы с точки зрения бизнеса.

Часто достижение хороших показателей на метриках машинного обучения не приводит к прорыву в бизнесе. Один из основных факторов, тормозящих внедрение машинного обучения в бизнес, - это несовместимость **метрик машинного обучения** и **показателей**, которыми оперирует топ-менеджмент.

Даже если математик обещает увеличение прибыли, необходимо еще понять, где **прибыль** увеличилась за счет **машинного обучения**, а где - за счет **других факторов**.

Кроме того, достижение желаемого уровня метрик может быть слишком **дорогим**.

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

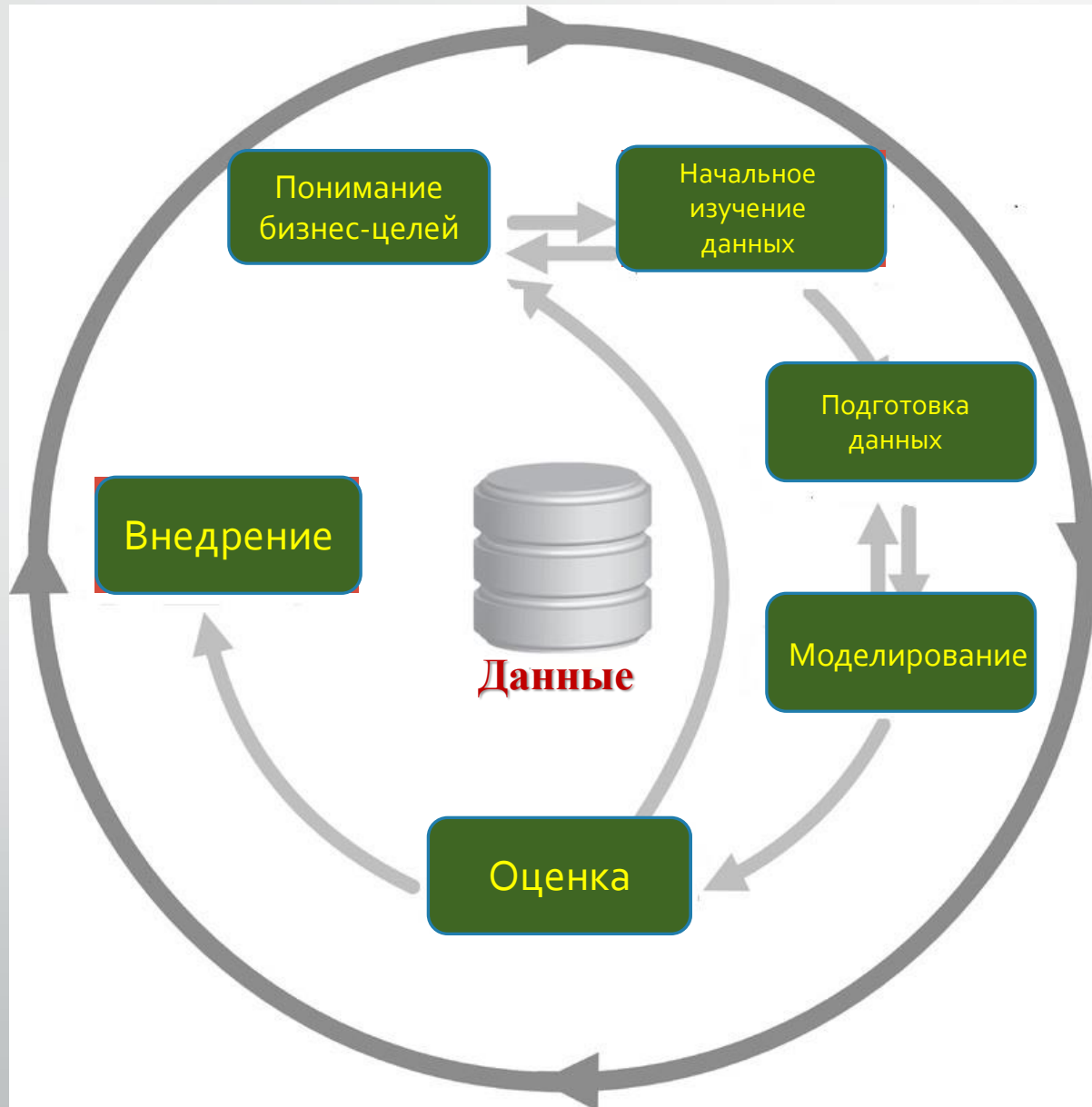
К сожалению, даже в случае выполнения указанных пунктов (имеется большое количество данных, ресурсов, и алгоритм настроен максимально качественно), проект может завершиться неуспехом, не достигнув желаемых значений метрик качества.

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.

Задачи машинного обучения - методология по исследованию данных CRISP-DM

англ. Cross-Industry Standard
Process for Data Mining-**CRISP-DM**



1.Понимание бизнес-целей
(Business Understanding)

2.Начальное изучение данных
(Data Understanding)

3.Подготовка данных (Data
Preparation)

4.Моделирование (Modeling)

5.Оценка (Evaluation)

6.Внедрение (Deployment)

Понимание бизнес-целей (Business Understanding) – на первом этапе работы с данными вам нужно понять, зачем вам собирать и анализировать данные, а также какие данные вам необходимы. Определение целей и предварительные гипотезы на данных затем лягут в основу вашего проекта.

Задачи фазы **Понимание бизнес-целей** :

Определить цели вашей организации

Оценить текущую ситуацию

Определить цели анализа данных

Составить план проекта

Начальное изучение данных (Data Understanding) – на втором этапе работы с данными вам нужно оценить качество ваших данных: насколько данные полные, есть ли в них ошибки, пробелы и пропуски.

Нужно понять, какими сведениями вы обладаете, сформулировать к ним вопросы и итоговые гипотезы о скрытых закономерностях

Задачи фазы **Начальное изучение данных** :

Собрать исходные данные

Описать данные

Исследовать данные

Проверить качество данных

Подготовка данных (Data Preparation) – на этом этапе нужно сформировать итоговый набор данных для анализа, “очистить” данные, привести их в единый формат из исходных разнородных и разно форматных данных.

Задачи фазы **Подготовка данных** могут выполняться много раз без какого-то заранее определенного порядка:

- Отобрать данные (таблицы, записи и атрибуты)
- Очистить данные, в т.ч. выполнить их конвертацию и подготовку к моделированию
- Сделать производные данные
- Объединить данные
- Привести данные в нужный формат

Моделирование (Modeling) – на этом этапе вам нужно выбрать методику, каким образом анализировать данные, построить модель анализа.

Модель должна отражать весь их процесс анализа (что вы хотите выяснить с помощью анализа данных, какие данные вы используете, как они организованы, как они обработаны, и так далее). У вас может возникнуть необходимость вернуться к фазе подготовки данных, так как разные методы анализа требуют различных форматов данных.

Задачи фазы **Моделирование** :

- Выбрать методику моделирования
- Сделать тесты для модели
- Построить модель
- Оценить модель

Оценка (Evaluation) – определение, удалось ли достигнуть целей с помощью разработанной модели и полученных результатов анализа.

Данный этап позволяет понять, действительно ли те шаги, которые вы запланировали, позволяют получить те результаты, которые вы хотели. На данном этапе могут быть выявлены более важные задачи организации, которые не были учтены.

Задачи фазы **Оценка**:

- Оценить результаты
- Сделать рецензию процесса
- Определить следующие шаги

Внедрение (Deployment) – этот этап может быть простым или сложным, в зависимости от целей организации.

Обычно это - разработка и внедрение решений на основе анализа данных. Это может быть как составление отчета, так и автоматизация процессов для решения ваших целей.

Задачи фазы **Внедрение**:

- Запланировать развертывание
- Запланировать поддержку и мониторинг развернутого решения
- Сделать финальный отчет
- Сделать рецензию проекта

Рассмотрим подробнее некоторые аспекты этапов подготовки и моделирования данных, инструменты подготовки данных и способы их моделирования

Сбор данных

Важный вопрос на этом этапе - **поиск данных**

- формулировка запроса - что ищем;
- запрос консультаций с целью помощи в поиске источников поиска;
- самостоятельный поиск;
- запрос и получение данных.

Как спрашиваем?	➤ конкретно ➤ к привязке к задаче ➤ с учетом типа желаемых данных ➤ локализуя границы запроса
Где консультируемся	➤ - у экспертов в предметной области ➤ - в консалтинговых компаниях ➤ - у органов власти ➤ - в социальных сетях
Где ищем сами?	➤ ГИС ➤ консалтинговые компании ➤ научные порталы и учреждения ➤ отраслевые источники информации ➤ профильные ассоциации и объединения ➤ органы власти ➤ СМИ
Как запрашиваем данные?	➤ запрос коммерческого предложения ➤ запрос данных у госорганов со ссылкой на ФЗ №8 от 9.02.2009 ➤ автоматизированный сбор данных (скрейпинг) ➤ получение открытых данных ➤ покупка данных по оферте

Хранение данных

Хранение данных - это процесс обеспечения доступности, целостности, защищенности данных.

Данные можно хранить разным способом:

- твердотельный съемный или несъемный носитель - нужен доступ к самому носителю или устройству, в которое он помещается, для получения данных;
- сервера баз данных;
- облачное хранилище данных - доступ к данным возможен из любой локации и др.

Выбор способа хранения данных зависит от объема данных, необходимой скорости доступа к ней, частоте обновлений данных, количества лиц, которым будет разрешен доступ к данным, стоимости хранения нужного объема данных.

Основной формой хранения данных является **база данных**. С помощью СУБД можно получить доступ к данным, записать их, переместить, изменить, удалить. Рекомендую SQL Server 2019, как включающую использование **технологии больших данных и машинного обучения**.

Обработка данных

Под обработкой данных понимается определенная последовательность операций с данными, выполненных для получения новой информации путем пересмотра и уточнения имеющейся результатов анализа данных, вычислений и пр.

На первом этапе осуществляется первичная обработка данных - приведение данных к единому формату, выделение общих признаков, структурирование данных. Затем выбирается наиболее актуальная для решения задачи модель работы:

- точечная обработка активных задач - операции только с выбранными категориями;
- потокковая обработка в реальном времени - операции с большим объемом данных, поступающих непрерывно, в процессе чего результаты анализа меняются каждый раз когда поступают новые данные;
- пакетная обработка исторических данных - обработка данных, накопленных за определенный срок.

В зависимости от выбранной модели, решаемой задачи подбираются технологии, тип базы данных, которые будут наиболее эффективны в конкретном случае.

К процедурам обработки данных относятся: создание данных, модификация данных, поиск информации, принятие решений, создание отчетов, создание документов, повышение безопасности данных.

При обработке данных обращают внимание на их качество.

Выделяют чистые и грязные данные. Грязные данные отличает наличие обработки, дополнительной, не связанных с первоначальными данными, информации, недостаток первичных данных. Все это мешает полному анализу данных, так как грязные данные уже содержат в себе некоторые критерии анализа, “обнулить” значение которых нельзя.

Визуализация данных

Визуализация данных - процесс представления данных в агрегированном, понятном для восприятия человеком виде.

Визуализация может быть презентационной - готовой для демонстрации аудитории, исследовательской - готовой для получения некоторых промежуточных результатов обработки данных.

Визуализация может быть использована на всех этапах работы с данными: визуализация результатов первичной обработки, визуализация промежуточных результатов, визуализация окончательных результатов.

В связи с объемом анализируемых данных визуализация — это необходимый способ оформления данных в понятный человеку вид. Поэтому инструменты визуализации важны в работе с данными.

Вид визуализации данных:

- 1.Графики: линейный, график рассеивания и др.
- 2.Диаграммы: столбиковая, круговая, гистограмма, кольцевая, лепестковая, облако тегов и др.
- 3.Инфографика.
- 4.Схемы.
- 5.Презентации.
- 6.Карты: фотографическая, географическая, дорожная, тематическая, картограмма.
- 7.Дашборды (*инструмент, который не только визуализирует, но и анализирует любые данные. Пример - студенты могут посмотреть результаты своей учёбы, сравнить их с результатами других студентов, увидеть статистику по учебным курсам.*).
- 8.Иллюстрации.

Выводы о методологии исследования данных **CRISP**

Методология включает **шесть этапов**: понимание бизнеса, начальное изучение данных, подготовка данных, моделирование, оценка решения и внедрение.

- **Поиск данных** включает четыре этапа: формулировка запроса, консультации, самостоятельный поиск, запрос и получение данных
- **Хранить данные** можно на твердых носителях, серверах или в облачных хранилищах.
- **Обработка данных** включает в себя: первичную обработку и очистку, выделение общих признаков, уплотнение данных, выбор модели для анализа.
- **Анализ данных** - совокупность действий исследователя, направленных на получение определенных представлений о характере явления, описываемых этими данными.
- **Визуализация данных** - процесс представления данных в агрегированном, понятном для восприятия человеком виде.

ИИ - задача классификации

В машинном обучении разработаны методы для решения задач установленного вида, с конкретными входными и выходными величинами, и необходимо задать, что будет этими входными и выходными величинами в данном процессе.

Наиболее часто решаемая задача — **классификации**.

Разместили объявление о вакансии, однако желающих поработать в проекте оказалось очень много. Чтобы не тратить время на изучение резюме, напишем программу, которая поможет.

Программа должна отбирать нужные резюме по определенным признакам. То есть, делить всех кандидатов на два класса — подходящие и неподходящие. В машинном обучении это называется задачей классификации.

Построим алгоритм, который позволит системе определить, есть ли в резюме кандидата необходимые параметры. Если есть – отправить в папку «Собеседование», если нет – в папку «Отказать».

Теперь ЭВМ должна определить, по каким критериям будет происходить отбор.

Для предложим алгоритму группу резюме, уже обработанных вручную - это обучающая выборка.

В этой выборке содержатся как подходящие кандидаты, так и получившие отказ.

Правильность работы алгоритма будем проверять тестовой выборкой – это размеченные резюме.

Для каждого резюме в тестовой выборке, если она подходит и алгоритм ответил правильно, то ответу дается характеристика **True**, если неправильно - **False**.

И сам алгоритм делит резюме на два класса, подходящие, **positive**, и неподходящие, **negative**.

Все возможные сочетания этих характеристик и классов дают 4 типа результатов:

TP — true positive, алгоритм верно пометил резюме как подходящее

TN — true negative, алгоритм верно отнес резюме к неподходящим

FP — false positive, алгоритм ошибочно считает подходящим резюме, в котором нет нужных качеств

FN — false negative, алгоритм ошибочно отбраковал подходящее резюме

Используя эти категории, мы сможем посчитать определенные метрики, которые позволяют оценить качество работы алгоритма.

Важно выбрать наиболее подходящую.

Самая простая метрика – **доля** правильных предсказаний: сколько раз прогноз ЭВМ и разметка теста совпали между собой.

Метрика хорошо работает, если количество подходящих и неподходящих резюме **совпадает**.

В данном случае эта метрика не подойдет: понятно, что подходящих кандидатов будет мало, и качество алгоритма больше зависит от точного отбора приглашенных на собеседование кандидатов.

Другая метрика – **точность**.

Она показывает **отношение количества** верно угаданных подходящих резюме к количеству тех, кого ЭВМ вообще отнесла к группе «собеседование». То есть ЭВМ могла пригласить на встречу кого-то неподходящего.

Точность будем рассчитывать по формуле на экране.

Кроме точности есть еще **метрика полноты**.

Полнота показывает отношение количества верно угаданных подходящих резюме, к другому значению: количеству кандидатов, которых следовало пригласить по мнению программиста. То есть ЭВМ могла пропустить кого-то подходящего.

Полноту будем рассчитывается по формуле на экране.

С одной стороны, чем выше одновременно точность и полнота, тем лучше.

Но в реальности это почти невозможно, и приходится искать баланс между ними. Для этого можно использовать так называемую **F-меру**.

Она считается как среднее гармоническое между точностью и полнотой, и, как следствие, мало зависит от общего количества подходящих и неподходящих резюме.

Описание и пример приводиться в методических указаниях **«Оценка классификатора - точность, полнота, F-мера»**.

На этом процесс обучения можно закончить или продолжить, поскольку чем больше данных анализирует ЭВМ, тем более точные результаты мы получаем.