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)

Базовая работа с табличными данными.

Задача классификации

Машинное обучение позволяет усовершенствовать процессы в различных областях бизнеса, науки и общества. Однако машинное обучение — это не волшебный ящик, который берет данные и превращает их в умную программу. Чтобы применить технологии машинного обучения в конкретном процессе, нужно сформулировать проблему этого процесса в виде одной из задач машинного обучения, «перевести» суть проблемы с языка прикладной области на язык технологии. 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.

Табличные данные

Для начала разберемся с понятием структурированных данных. Услышав слово «данные», каждый представляет что-то свое: папку на компьютере с тысячами файлов, таблицу Excel, базу данных Oracle и т.д. В машинном обучении, как правило, работают с табличными данными, их еще называют структурированными.

В таблице данных по строкам расположены объекты — базовая единица данных; то, для чего необходимо выполнить прогноз.

В бизнес-задачах объектом часто является клиент, однако в качестве объекта может выступать что угодно: предприятие (предсказание банкротства предприятия), товар (определение категории товаров), отзыв клиента (определение тематики отзыва). Встречаются и более сложные объекты, например, в задаче предсказания спроса на товар объектом может быть пара "товар-день".

Столбцы таблицы задают признаки объектов — характеристики, позволяющие задавать особенности каждого конкретного объекта и отличать объекты друг от друга.

Для клиента банка в качестве признаков могут использоваться заработная плата, возраст, должность, уровень образования, город проживания, стаж работы. Для предприятия признаками могут выступать год основания, размер уставного капитала, тип юридического лица, годовой оборот, прибыль. Вообще говоря, значение признака должно быть числом, например, заработная плата клиента — 50 000. Однако используются и другие виды признаков, для них разрабатываются способы кодирования в виде чисел.

Заработная плата	**Возраст**	**Должность**	**Уровень образования**	**Город проживания**	**Стаж работы (годы)**	**Вернет ли клиент кредит**
100000	26	Риэлтор	Высшее	Санкт-Петербург	5	Да
50000	20	Продавец-консультант	Высшее	Москва	1	Нет
35000	39	Автомеханик	Среднее специальное	Воронеж	8	Нет
25000	23	Программист	Высшее	Самара	2	Да
75000	41	Юрист	Среднее	Москва	14	Да

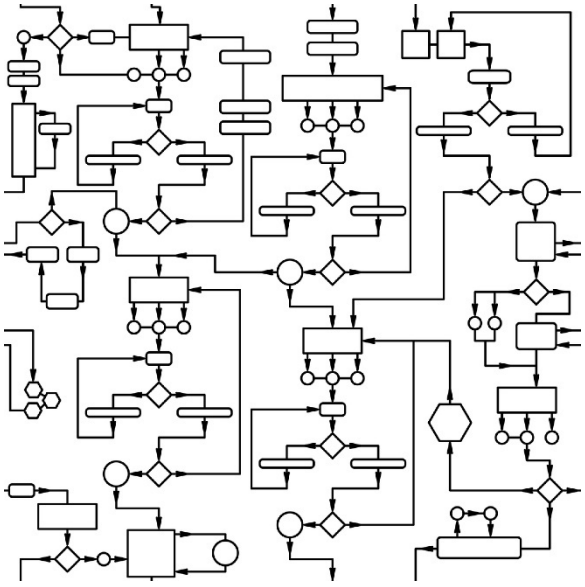
Для успешного применения алгоритмов машинного обучения важно, чтобы объектов было много. Например, в базах данных банка может храниться информация о миллионах клиентов, т.е. миллионы строк в таблице данных, — это достаточно большой объем данных. Наоборот, в медицинском учреждении может иметься информация лишь о небольшом количестве пациентов, например, нескольких сотнях — это уже относительно небольшой объем данных, обучить на нем качественный алгоритм может быть проблематично. Количество признаков, столбцов в таблице данных, может варьироваться, обычно их от нескольких десятков до нескольких сотен.

Концепция создания алгоритма классификации

Определившись с видом данных, мы можем перейти к постановке задачи классификации.

В этой задаче каждому объекту (строке в таблице данных) соответствует **класс** — значение из заданного набора классов. К примеру, в задаче предсказания оттока клиента объектом является клиент, классов два: «уйдет» и «не уйдет», и каждому клиенту соответствует один ответ (класс). В задаче категоризации отзывов клиентов объект — это один отзыв, классы — категории отзывов (к примеру, «жалоба», «предложение», «позитивный отзыв»), и каждый отзыв относится к какой-то категории (классу).

Задача классификации состоит в том, чтобы разработать алгоритм, который по признакам объекта будет предсказывать класс, например для клиента предсказывать его уход.

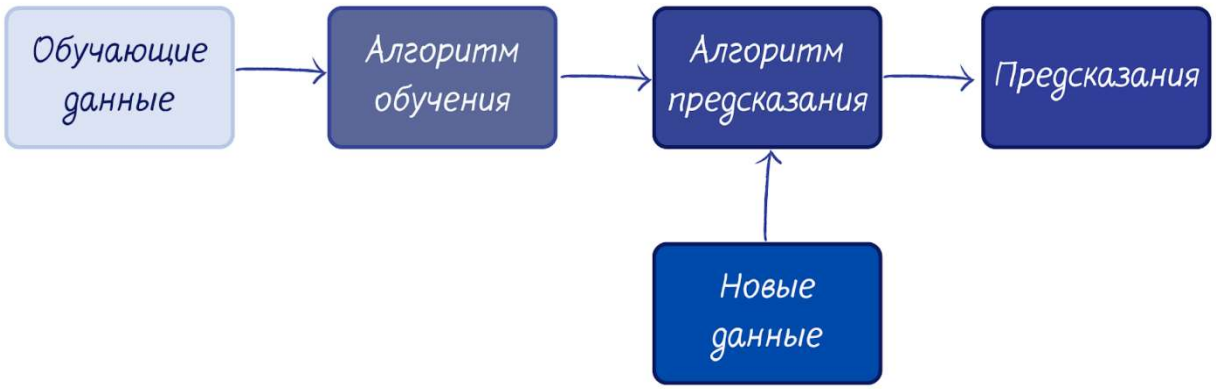


Простейший алгоритм предсказания ухода клиента может выглядеть так: если клиент пришел после 2018 года и совершил меньше трех покупок за последний год, то он уйдет, иначе считаем, что не уйдет. Иными словами, имея признаки клиента: «год прихода клиента» и «число покупок за последний год», мы задаем процедуру, как по этим признакам предсказать класс. Эта процедура, алгоритм, будет выполняться на компьютере автоматически, поэтому должна быть очень четкой. Разумеется, программа может быть сложнее: включать различные условия «если … иначе», вычисления с использованием признаков (например, суммирование значений

признаков) и т.д., он может обрабатывать сотни признаков, но обязательно должен быть четко сформулированным.

Возникает вопрос, как составить такой алгоритм предсказания класса. Можно было бы спросить специалистов, по каким правилам они определяют, уйдет ли клиент, и записать эти правила в виде компьютерной программы. Такой подход называется экспертной системой. Однако у него есть ряд недостатков: специалисты, которых мы будем спрашивать, могут быть не согласны друг с другом, они могут иметь опыт работы только с узким кругом клиентов и ничего не знать про других клиентов, в конце концов, их опыт может быть сложно формализовать в виде четкой программы, особенно если они во многом опираются на интуицию.

Машинное обучение предлагает альтернативное решение на основе обработки данных: по большому количеству примеров клиентов и информации, ушли они или нет, будет автоматически составлен алгоритм предсказания класса «уйдет» или «не уйдет» для нового клиента. То есть получается, что на основе данных программа создает другую программу: первая называется **алгоритмом обучения**, вторая — **алгоритмом предсказания**. Благодаря тому, что алгоритм обучения «видел» много примеров из реальной жизни, он будет способен делать качественные предсказания для новых клиентов.



Обучающие данные представлены в виде таблицы "объекты-признаки", также имеется отдельный столбец с классами объектов. Опираясь на данные,

алгоритм обучения автоматически составляет алгоритм предсказания, он умеет по признакам (столбцам) нового объекта определять его класс.

Еще немного о данных

Данные о клиентах, транзакциях и других объектах, как правило, имеются в крупных компаниях, хотя для некоторых задач может потребоваться собрать их отдельно (к примеру, скачать большие базы текстов новостей). Однако для задачи классификации важно, чтобы данные были размечены, то есть для каждого объекта должен быть известен его класс. В задачах кредитного скоринга или оттока клиентов разметка данных накапливается вместе с самими данными (клиент либо вернул кредит, либо нет), но для решения задач детекции мошеннических транзакций или категоризации отзывов клиентов имеющиеся данные могут не содержать разметки. В таких случаях приходится заказывать разметку, для этого существуют специальные сервисы, напримерЯндекс.Толока.

Также важно отметить, что табличные данные должны быть полными, то есть для каждого объекта должны быть известны значения всех (или почти всех) признаков. При наличии небольшой доли пропущенных значений признаков их можно заполнить средними значениями, но, если пропусков много, найти зависимости в данных может быть сложно. Аналогично, значения всех (или почти всех) признаков должны быть известны для каждого нового объекта. Например, в задаче предсказания типа следующей транзакции клиента хорошим признаком может оказаться сумма транзакции, но понятно, что пока клиент эту транзакцию не сделал, мы не знаем ни ее типа, ни суммы, то есть использовать признак «сумма транзакции» нельзя.