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 6 «Внедрения искусственного интеллекта в бизнес»

Часовских Виктор Петрович
д-р техн. наук, профессор кафедры ШИиКМ

ФГБОУ ВО «Уральский государственный экономический университет»

Екатеринбург 2022

Табличные данные

В машинном обучении, как правило, работают с табличными данными, их еще называют структурированными.

Таблица Excel, база данных SQL 2019 (поддерживает машинное обучение). Интеграция машинного обучения с SQL Server обеспечивается благодаря поддержке языка программирования R, коммерческая версия R называется RevoScaleR.

Была создана платформа машинного обучения, используя данные, размещенные на SQL Server. Была выполнена интеграция с языком программирования Python. Благодаря этим новым возможностям R и Python совместно стали службами машинного обучения SQL Server (SQL Server Machine Learning Services).

Службы SQL Server ML Services предлагают новую возможность для создания и использования масштабируемых приложений машинного обучения в соответствии со следующими концепциями:

- приложения машинного обучения выполняются на том же компьютере, что и SQL Server, но в независимых от SQLSERVR.EXE процессах;

- SQL Server предоставляет интерфейс T-SQL посредством системной хранимой процедуры sp_execute_external_script для выполнения кода машинного обучения;
- SQL Server предоставляет архитектуру, делающую возможными интеллектуальный обмен данными и масштабируемость, используя специальный программный код для машинного обучения.

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.

Создание 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).

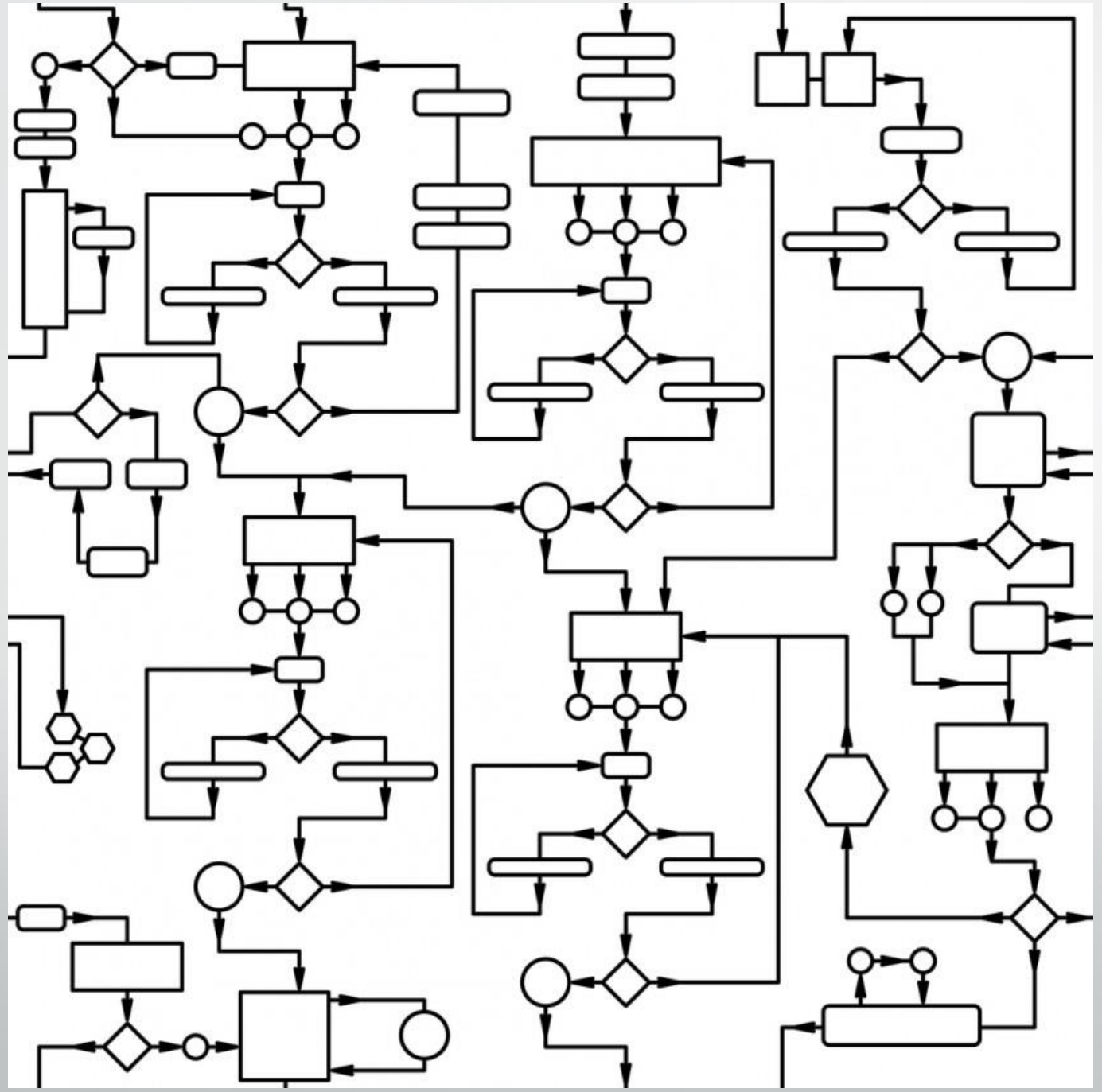
В задаче категоризации отзывов клиентов объект — это один отзыв, классы — категории отзывов (к примеру, «жалоба», «предложение», «позитивный отзыв»), и каждый отзыв относится к какой-то категории (классу).

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 2020 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.

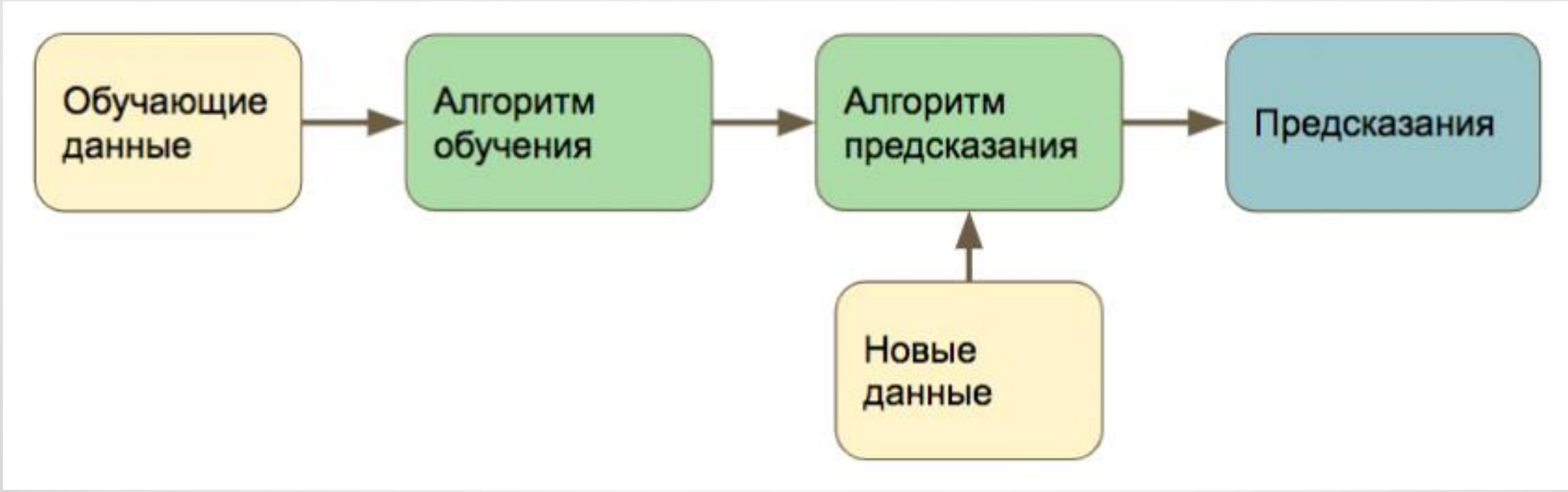
Машинное обучение предлагает альтернативное решение на основе обработки данных:

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.

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.

Задачи машинного обучения, например классификация, служат интерфейсом между бизнес-процессами и технологиями машинного обучения.

Чтобы воспользоваться существующими методами решения задачи классификации,

нужно выбрать объект,

выбрать для него признаковое описание и определиться, что будет предсказываемым классом,

а также собрать размеченные и полные данные.

Задача регрессии

Задача регрессии во многом похожа на задачу классификации, но подразумевает предсказание других величин.

Данные в задаче классификации представлены в виде таблицы, каждая строка — объект, столбцы — признаки объектов (все объекты описываются одним и тем же набором признаков, но значения признаков у каждого объекта свои). В задаче классификации также имеется отдельный столбец с классами объектов. Этот столбец еще называют столбцом значений целевой переменной, то есть величины, которую нужно предсказывать.

В задаче регрессии данные имеют такой же вид, но целевая переменная числовая.

Иными словами, задача регрессии состоит в том, чтобы на основании различных признаков предсказать вещественный ответ, т.е. для каждого объекта нужно предсказать число.

Например, в задаче предсказания спроса на товар нужно предсказать, какое количество единиц товара потребуется в торговой точке в определенный промежуток времени, например в конкретную неделю.

Другой пример — предсказание стоимости квартиры (стоимость в рублях — числовая величина).

Еще один пример — предсказание возраста человека по фотографии (возраст — число).

Отличие задачи классификации от задачи регрессии выглядит незначительным и в принципе действительно таковым является, однако, как мы увидим ниже, алгоритмы предсказания и обучения будут работать по-разному для этих двух задач.

Как и в задаче классификации, потребуются обучающие данные с признаками объектов и известными значениями числовой целевой переменной.

Для примера будем рассматривать задачу предсказания стоимости квартиры с данными следующего вида:

Признаки				Целевая переменная
Площадь ($м^2$)	**Этаж**	**Число комнат**	**Число лет с последнего ремонта**	**Стоимость (млн)**
115	3	4	2	46.0
55	5	2	5	10.3
72	6	3	12	17.7
55	20	1	0	32.0

В этой задаче объектом является квартира, признаки: площадь, число комнат, этаж и число лет с последнего ремонта, целевая переменная — стоимость квартиры.

На основе обучающих данных алгоритм обучения составляет алгоритм предсказания.

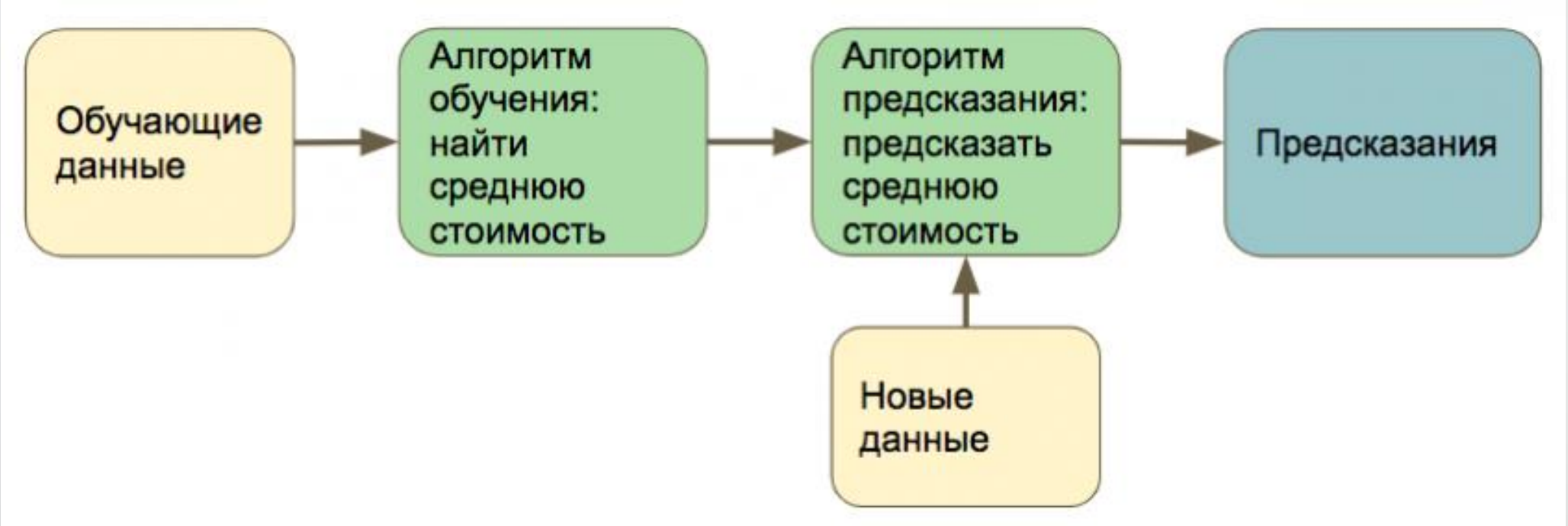
Алгоритм предсказания по признакам новой квартиры (площадь, число комнат, этаж и информация о ремонте) определяет ее стоимость.

Итак, предположим, мы формализовали задачу в виде задачи регрессии и собрали обучающие данные. Сосредоточимся на том, как на основе обучающих данных выполнить обучение и создать алгоритм предсказания.

Самый простой (и на самом деле бесполезный) алгоритм предсказания, который можно придумать, — это такой, который для всех объектов предсказывает одно и то же.

В нашей задаче это означает, что для всех квартир предсказывается одна и та же стоимость.

Определить эту стоимость можно по обучающим данным, например, можно использовать среднюю стоимость квартир в обучающих данных — в вычислении этой средней стоимости и будет заключаться обучение; все лучше, чем предсказывать, например, стоимость ноль.



Конечно, описанный алгоритм не принесет никакой пользы на практике, потому что никак не учитывает особенности объектов, но его можно использовать в качестве базового решения. Более сложный алгоритм не должен делать предсказания хуже, чем базовое решение. Понять, насколько хорошие предсказания делает алгоритм, можно, измерив среднюю ошибку.

Линейные модели

Самый известный метод регрессии — это линейные модели.

Основной механизм предсказания с помощью линейной модели формулируется следующим образом: необходимо умножить значения всех признаков на веса и сложить.

Предположим, что мы хотим предсказать стоимость квартиры со следующими значениями признаков:

Площадь ($м^2$)	Этаж	Число комнат	Число лет с последнего ремонта
70	2	3	5

Также предположим, что мы знаем веса признаков: 0,25 для площади, 1,8 для этажа, 0,5 для числа комнат и (-0,2) для числа лет со дня ремонта.

Получается, что вес признака задает вклад признака в предсказание: например, вес 0,25 означает, что каждый квадратный метр квартиры добавляет 0,25 единицы в ее стоимость, а вес 0,5 — что каждая комната в квартире добавляет 0,5 единиц в ее стоимость. Вес -0,2 означает, что каждый год, прошедший с последнего ремонта, вычитает 0,2 единицы из стоимости квартиры.

Тогда будет предсказана стоимость $0,25 \cdot 70+1,8 \cdot 2+0,5 \cdot 3-0,2 \cdot 5=21,6$ условных единиц, т.е. мы умножили признаки на веса и сложили.

Большое по модулю значение веса означает, что соответствующий признак вносит сильный вклад в предсказание, а вес, близкий к нулю, означает, что соответствующий признак практически не влияет на предсказание.

Поиск значений весов

Именно веса будем искать в процессе обучения по данным: алгоритм обучения автоматически подберет значения весов, такие, что ошибка предсказания с этими весами будет меньше, чем с другими весами.

При этом ошибка измеряется по обучающим данным: чем больше обучающих объектов (квартир в обучающих данных), тем в большем количестве случаев будет хорошо работать обученный алгоритм.

Если обучающих данных достаточно много, найденные в процессе обучения веса будут точно лучше, чем веса, которые предложил бы использовать специалист по недвижимости, потому что алгоритм обучения «видел» гораздо больше примеров, чем специалист.

РАНГОВЫЙ МЕТОД

Ранг - это номер объекта в упорядоченной по некоторому качеству совокупности однородных объектов. При использовании рангового метода эксперту предлагается определить ранги целей в пределах каждого частного множества по степени убывания важности достижения каждой из них сравнительно с остальными.

Таким образом, если в частном множестве находится ***n*** целей, каждой из них будет поставлен в соответствие ранг r_i , i = $\overline{1,n}$, $r_i = \overline{r_1, r_n}$.

Рекомендуется проводить ранжирование последовательным выделением наиболее важной из рассматриваемых целей, присвоением ей минимального ранга с последующим исключением этой цели и этого ранга из рассмотрения.

Пусть определены цели {ц1, ц2, ц3, ц4}

Заготовим таблицу ранжирования

Номер цели	1	2	3	4
Ранг				

Пусть **ц2 -** наиболее важная цель, тогда r_2=1

Номер цели	1	2	3	4
Ранг		1		

Рассмотрим оставшиеся цели {ц1, ц3, ц4}, пусть Ц4 наиболее важная цель, тогда $r_4 = 2$

Номер цели	1	2	3	4
Ранг		1		2

и так далее, до заполнения таблицы ранжирования

По оценкам важности определяются оценки значимости целей (веса целей). Оценки индивидуальной значимости

$$V_{i,k} = \frac{2}{n}\left(1 - \frac{r_{i,k}}{n+1}\right)$$

где n – количество целей, $r_{i,k}$ – ранг i – ой цели по мнению к – ого эксперта.

Групповые оценки значимости

$$V_i = \frac{2}{n}\left(1 - \frac{\sum_{r=1}^{m} r_{i,k}}{m(n+1)}\right)$$

где m – количество экспертов.

Ранговый метод характеризуется малой трудоемкостью, сравнительно высокой солидарностью и малой разрешающей способностью.

Метод парных сравнений

При использовании этого метода каждому эксперту предлагается оценить относительную важность достижения каждой из двух рассматриваемых целей заданной совокупности. Таким образом, эксперт выполняет $\frac{n}{2(n-1)}$ оценок, заполняя матрицу предпочтений вида:

	1	2	3	....	j	...	n
1	1	[1]/[2]	[1]/[3]		[1]/[j]		[1]/[n]
2		1	[2]/[3]		[2]/[j]		[2]/[n]
3			1				
...				1	[...]/[j]		[...]/[n]
i					1		[i]/[n]
...						1	
n							1

Как правило, при проведении экспертизы оговаривается, в какой форме представляется отношение важности i - той и j – ой целей [i]/[j], например, в форме отношения натуральных чисел от 1 до 10. Матрица предпочтений каждого эксперта преобразуется в матрицу рангов с использованием следующих соотношений:

$$r_{i,j} = \frac{2P_{i,j}}{P_{i.j} + 1}$$

$$r_{j,i} = \frac{2}{P_{i.j} + 1}$$

Оценки индивидуальной значимости целей для i - го эксперта

$$V_{i,k} = \frac{\sum_{j=1}^{n} r_{i,j,k}}{\sum_{i=1}^{n}\sum_{j=1}^{n} r_{i,j,k}}$$

Групповые оценки значимости

$$V_i = \frac{\sum_{k=1}^{m}\sum_{l=1}^{n} r_{i,j,k}}{\sum_{r=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} r_{i,j,k}}$$

Метод парных сравнений характеризуется существенно большей трудоемкостью и меньшей солидарностью, однако обеспечивает высокую разрешающую способность. При отсутствии априорной информации о степени изученности проблемы и квалификации экспертов рекомендуется проведение экспертизы с использованием рангового метода.

Солидарность экспертной группы оценивается коэффициентом конкордации

$$W = \frac{12 \sum_{i=1}^{n} \left(\sum_{r=1}^{m} r_{i,k} - m(n+1)/2 \right)^2}{m^2 n (n^2 - 1)}$$

W изменяется от 0 до 1

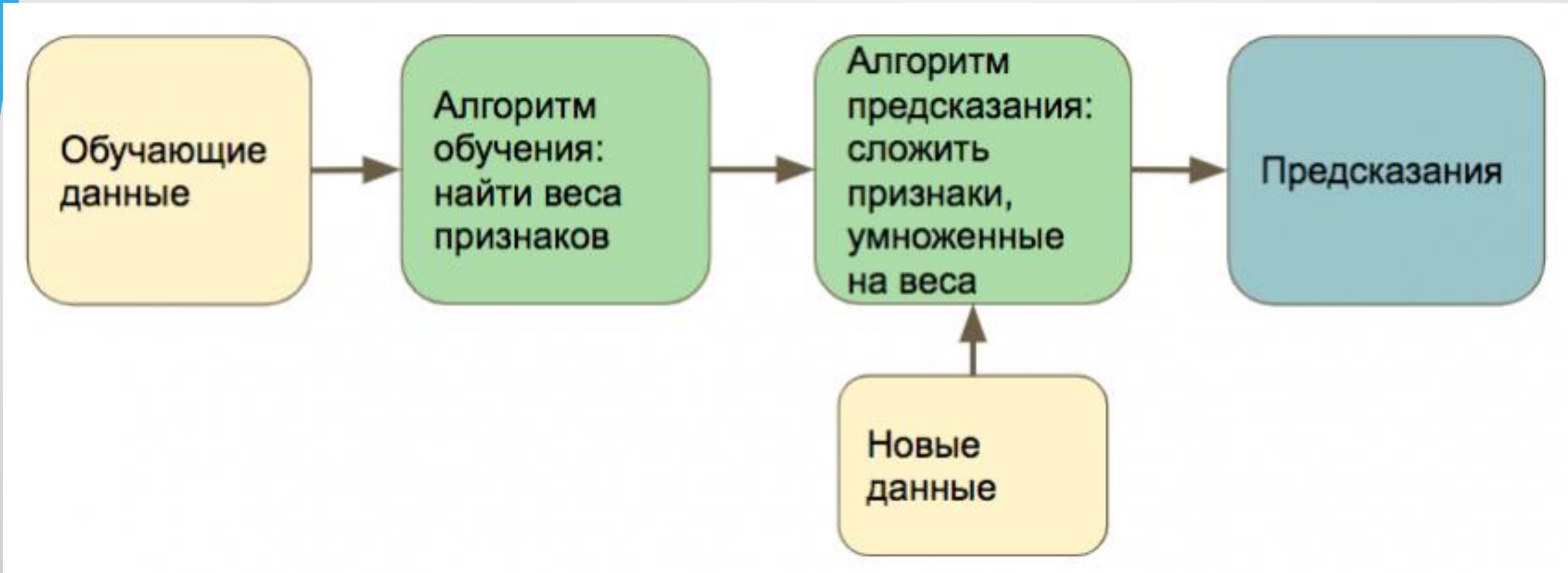
Если полученное значение W превышает 0,67, рекомендуется проведение повторной экспертизы с использованием метода парных сравнений. В противном случае проанализировать причины низкой солидарности. С этой целью могут быть использованы оценки солидарности двух экспертов (двух групп экспертов).

Коэффициент парной корреляции по Спирмену вычисляется по формуле

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} \left(R_{i,1} - R_{i,2}\right)^2}{n(n^2-1)}$$

где $R_{i,1}$ и $R_{i,2}$ - ранги i – той цели в оценках 1 и 2 экспертов (групп экспертов)

Диапазон от -1 до +1


Обучающие данные
Алгоритм обучения: найти веса признаков
Алгоритм предсказания: сложить признаки, умноженные на веса
Предсказания
Новые данные