



Введение

в ML

Цель курса

Научить основным подходам в области анализа больших данных и машинного обучения.

Дать практический опыт на реальных данных.

Пригласить самых сильных студентов в акселератор стажеров в январе\феврале 2021.

План спецкурса

1. Вводная лекция
2. Знакомство с Python. Базовые библиотеки для DS.
Визуализация
3. Линейный модели. Метрики
4. Деревья решений. Ансамбли
5. Обучение без учителя
6. Проверка гипотез, А/В тестирование
7. Работа с текстовыми данными
8. (*) Анализ графов
9. Работа с гео-данными
10. Базы данных, фреймворки для работы с большими данными
11. Инфраструктура и основы Hadoop
12. Data Science в телекоме

8

домашних заданий

2

прикладных проекта

Срок выполнения – 1 неделя

Захита – на последнем
занятии 10 декабря



Наш курс - это

- **Ориентация на практику** — максимум навыков, которые пригодятся при решении реальных задач
- **Много самостоятельной работы** — некоторые задания для самостоятельной работы подразумевают до 10 ч. работы
- **Четкие сроки** — по домашним заданиям, за списывания/невыполнения заданий мы отчисляем с курса
- **Доступность материалов** — все материалы курса — видео занятий, код и данные будут доступны на GitHub'е и YouTube

Используемые платформы



— для вопросов по курсу, материалам занятий



— для материалов занятий и сдачи дз



— для онлайн занятий



— для записей занятий

Наши спикеры



Горбань Иван
Lead Data Scientist



Шелепанов Сергей
Data Scientist



Тувалева Юлия
Data Scientist, Geo



Васильев Роман
Data Scientist



Володкина Екатерина
Data Scientist



Морозов Александр
Data Engineer



Карнушин Валерий
Data Scientist



Тюкавин Андрей
Data Scientist, Geo

Чему научитесь:

- Понимание классических алгоритмов машинного обучения (линейные модели, деревья, ансамбли и т.д.).
- Знание базовых метрик.
- Умение строить свои модели и проектировать эксперименты для их использования.
- Умение работать с данными разной природы.
- Понимание того, как DS помогает бизнесу.



- **1 этап** — первые 2 занятия + дз. Коммуникации в tg
- **2 этап**— 3 занятие и далее, попадут лишь те, кто сдал первое дз. Получат полный доступ к материалам.

План

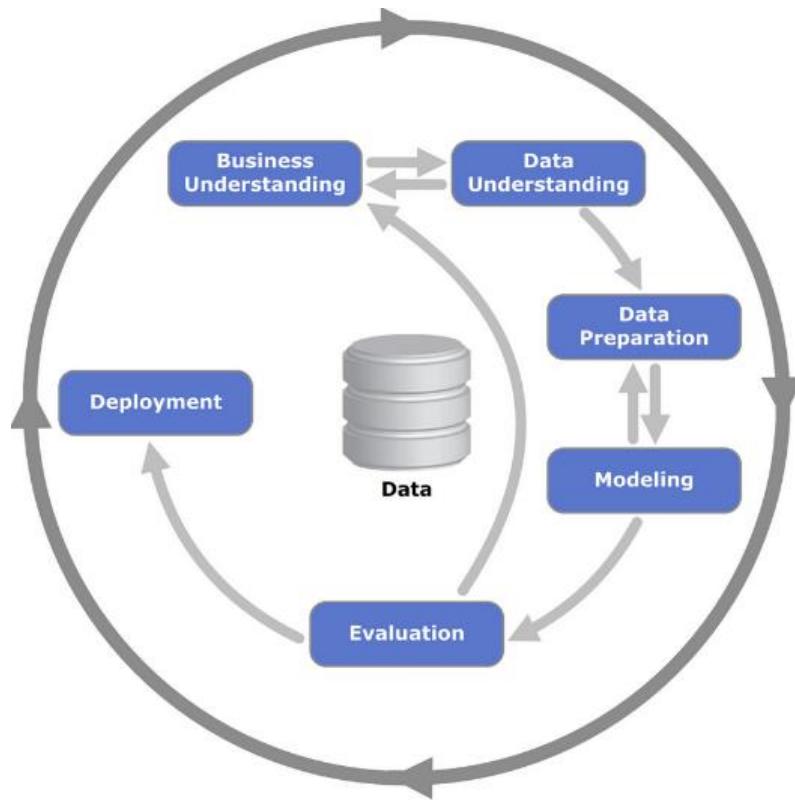
60 мин	Типы и примеры задач
10 мин	Перерыв
50 мин	Измерение качества моделей, работа с признаками
10 мин	Перерыв
40 мин	Наш первый алгоритм машинного обучения - kNN

Машинное обучение

Что нужно чтобы обучать машины?

- Цель
- Данные

Cross Industry Standard Process for Data Mining – CRISP-DM



типы задач

и примеры



Машинное обучение

1) Классическое обучение

а) С учителем

Есть объекты с признаками и верные ответы

б) Без учителя

Есть объекты с признаками

2) Обучение с подкреплением

Есть среда, с которой можно взаимодействовать

3) Нейросети и глубокое обучение

а) Работа с данными сложной структуры

б) Когда из классического обучения уже всё выжато

Машинное обучение

Классическое обучение

a) С учителем

Есть объекты с признаками и верные ответы

b) Без учителя

Есть объекты с признаками

2) Обучение с подкреплением

Есть среда, с которой можно взаимодействовать

3) Нейросети и глубокое обучение

a) Работа с данными сложной структуры

b) Когда из классического обучения уже всё выжато

Классическое обучение с учителем

Признаки (Features)

Target

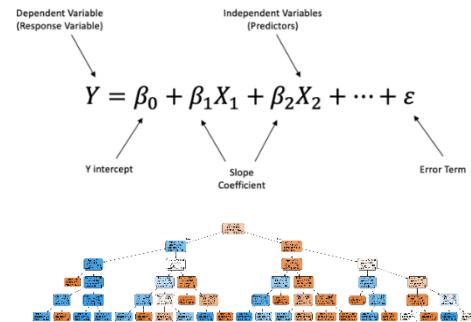
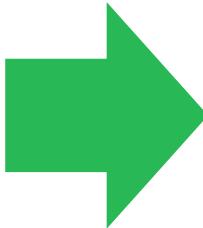
Признаки (Features)									Target
36	1	36	1	0	14	5	2	3 321 000	
56	2	4	0	0	3	3	4	13 000 000	
41	1	28	0	1	13	23	2.3	8 020 000	
148	4	13	1	1	7	3	5	21 412 000	
...	

Тренировка модели на Train

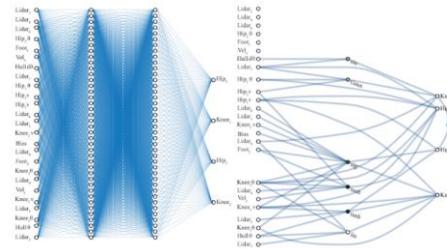
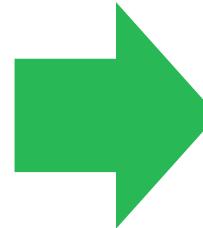
Алгоритм машинного обучения + его параметры

Признаки (Features)

Train



Target
Train



Использование модели



И ЭТО твоя система Машинного обучения?

Ага! Высыпаешь данные в эту большую кучу линейной алгебры, а потом с другой стороны собираешь ответы.

А если ответы неверные?

просто перемещай кучу, пока они не станут выглядеть правильно.



Каким может быть Target?

- 1). вещественным числом (\mathbb{R}) – задача регрессии
- 2). {0, 1} – задача бинарной классификации
- 3). {0, 1, ..., M-1} – задача многоклассовой классификации (на M классов)
- 4). {1, ..., k} для каждого id, k – число item'ов для id. Задача ранжирования

Как бизнес может использовать классическое обучение с учителем?



Примеры из задач МегаФона:

- Предсказание оттока пользователей
- Прогнозирование объемов продаж
- Прогнозирование интереса абонента к услуге
- Прогноз удовлетворённости сотовой связью (Customer Satisfaction Index)
- Next Best Action



- Предсказание оттока пользователей - **бин. классификация**
- Прогнозирование объемов продаж - **регрессия**
- Прогнозирование интереса абонента к услуге - **бин. классификация**
- Прогноз удовлетворённости сотовой связью (Customer Satisfaction Index) - **бин. классификация**
- Next Best Action - **ранжирование**

Примеры из проектов МегаФона:

Психотипирование личности

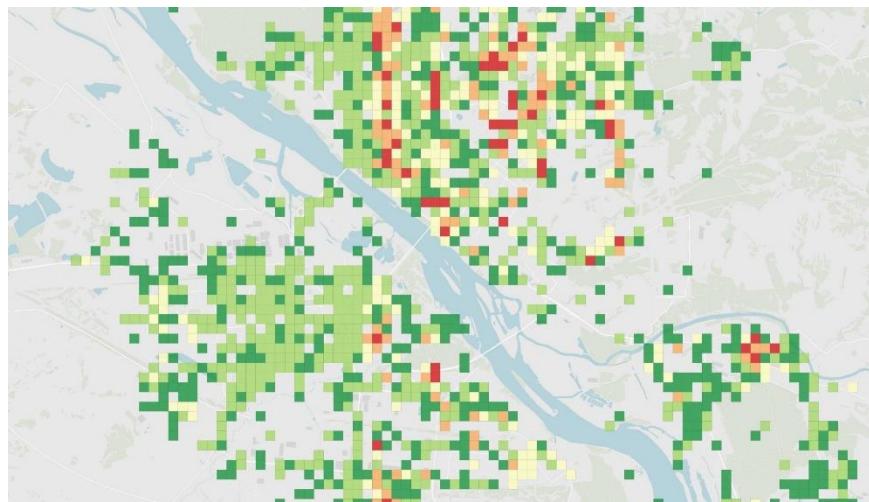
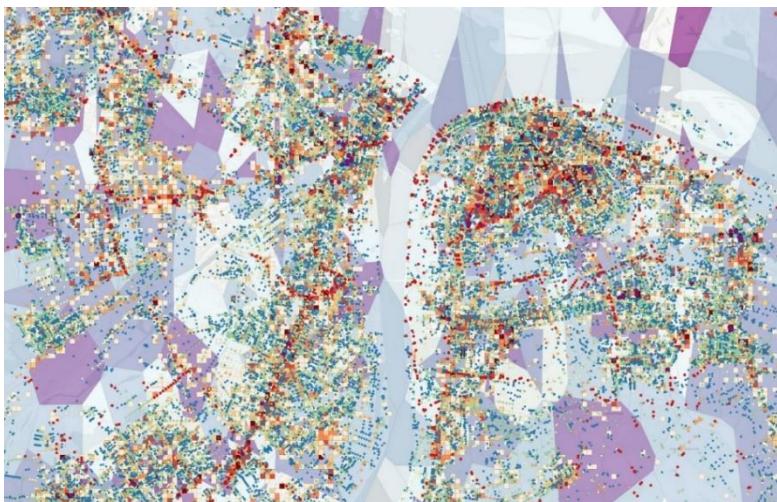


Психотипирование личности

Классификация на пересекающиеся классы

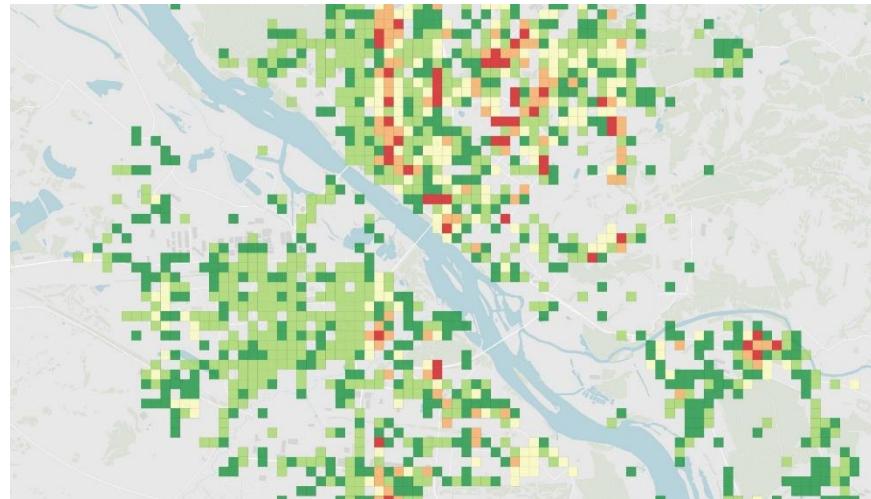
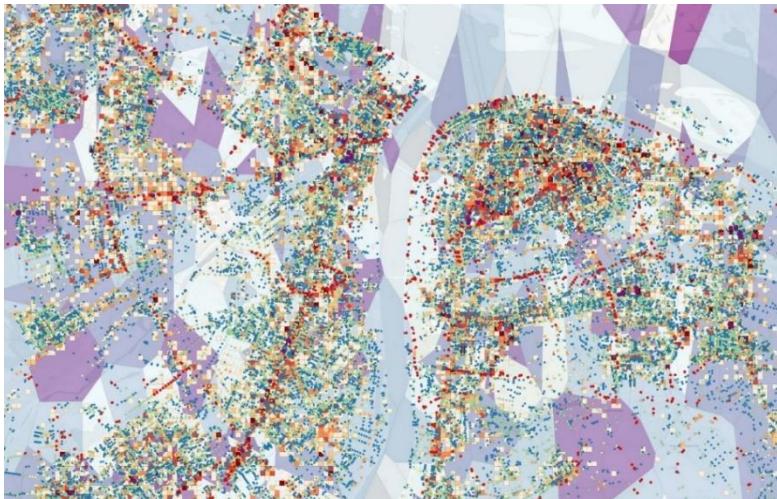


Прогнозирование клиентопотока по полигонам

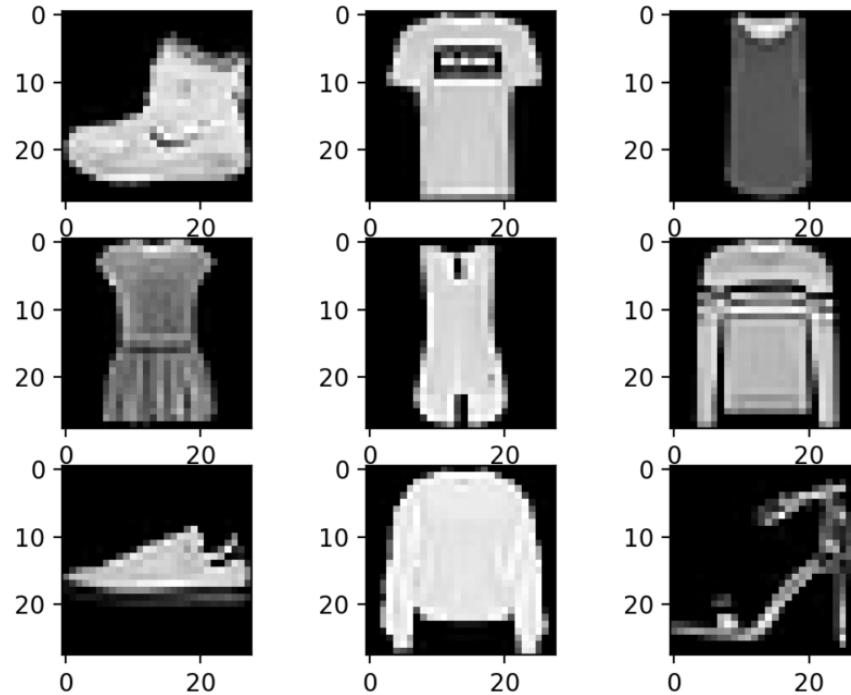


Прогнозирование клиентопотока по полигонам

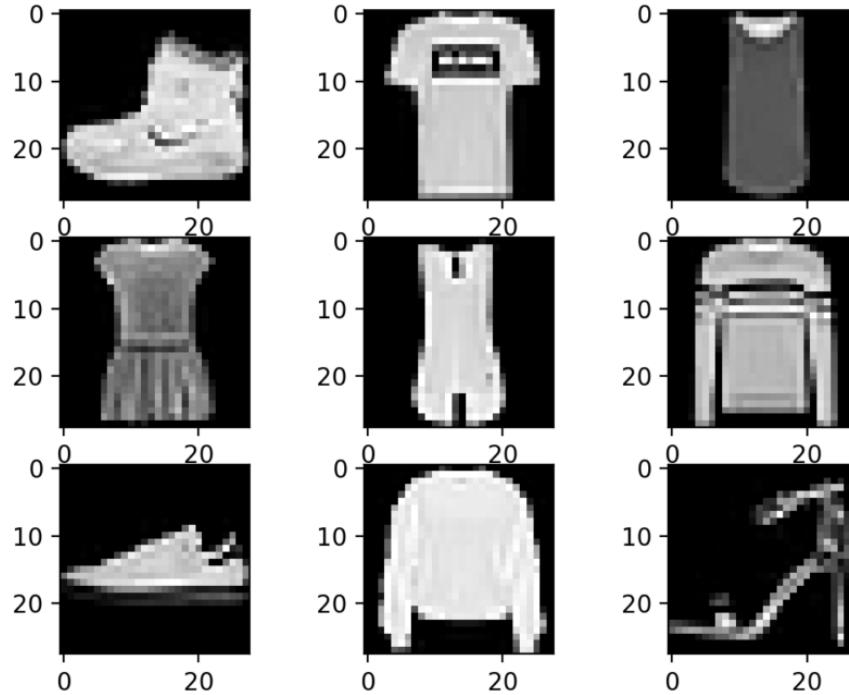
Регрессия



А если смотреть не только в МегаФоне?



А если смотреть не только в МегаФоне?



Многоклассовая
классификация

А если смотреть не только в МегаФоне?



А если смотреть не только в МегаФоне?



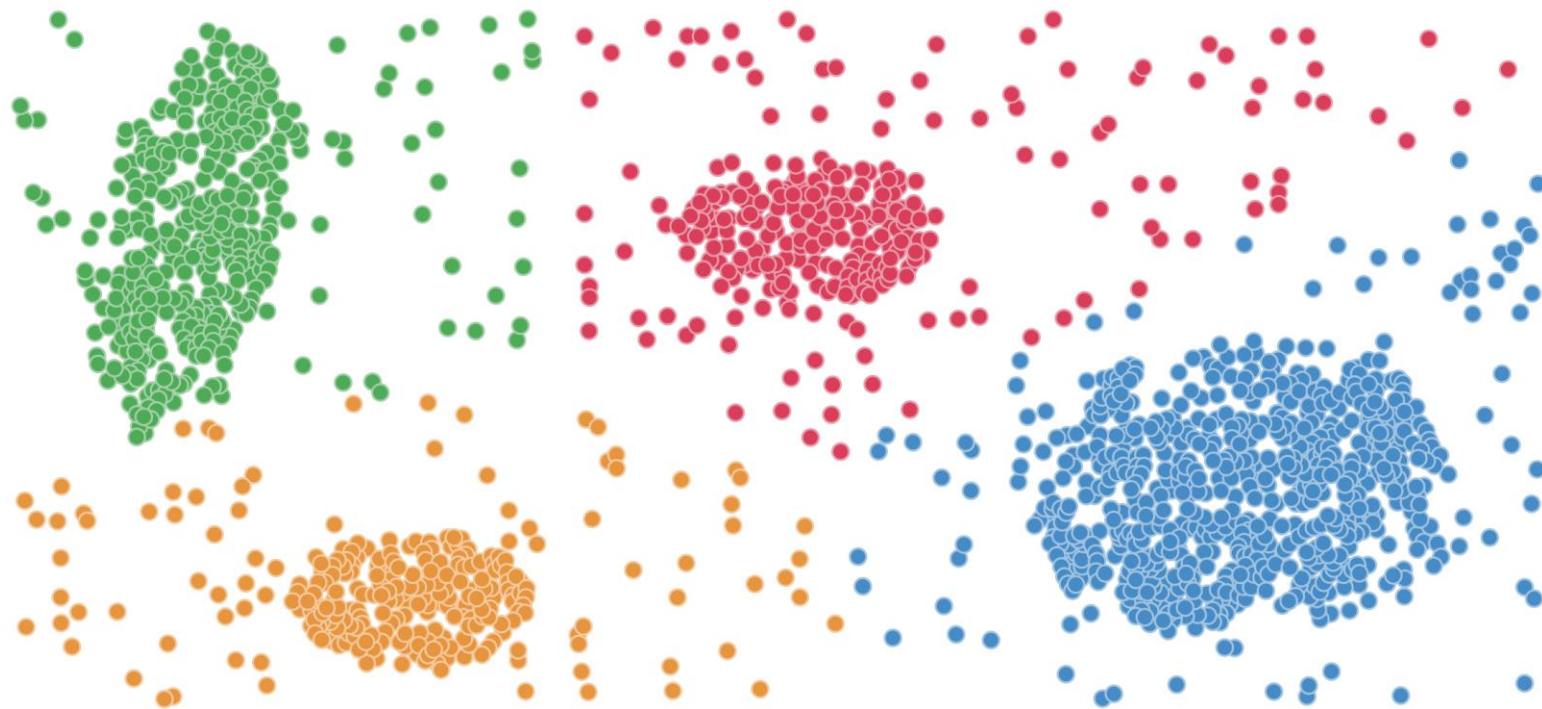
Бинарная классификация

Классическое обучение без учителя

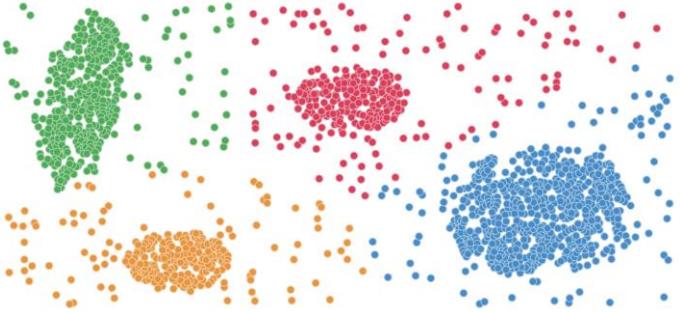
Признаки (Features)

Площадь, м ²	Число комнат	Расстояние до центра, км	Новостройка	Наличие балкона	Время до метро, мин	Этаж	Высота потолков, м
36	1	36	1	0	14	5	2
56	2	4	0	0	3	3	4
41	1	28	0	1	13	23	2.3
148	4	13	1	1	7	3	5
...

~~Target~~



Как связаны эти две картинки?



и

Москва · Недвижимость · Квартиры · Купить · 2-комнатные · Вторичка
2-к квартира, 50 м², 12/17 эт.
Добавить в избранное Добавить заметку Сегодня в 01:29

15 000 000 ₽
Возьми в Ипотеку
Подробнее
Показать телефон
в 938 XXX-XX-XX
Написать сообщение
Отвечает в течение дня

Анастасия
Агентство
На Авито с апреля 2013
№ 2010088953, 48 (+48)

Назад Следующее →

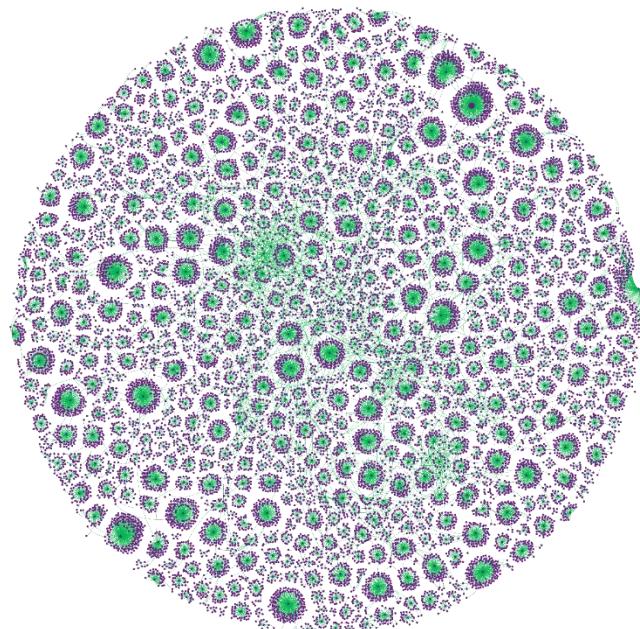
Похожие объявления

2-к квартира, 43 м ² , 1/5 эт.	2-к квартира, 50,2 м ² , 2/17 эт.	2-к квартира, 52 м ² , 9/12 эт.
11 700 000 ₽ Москва, Багратионовская 20 сентября 20:11	13 950 000 ₽ Москва, Багратионовская 9 сентября 10:55	13 900 000 ₽ Москва, Багратионовская 3 сентября 19:48
2-к квартира, 59 м ² , 5/8 эт.	3-к квартира, 75,6 м ² , 22/22 эт.	1-к квартира, 37 м ² , 7/16 эт.
18 300 000 ₽ Москва, Багратионовская 29 сентября 07:53	23 299 000 ₽ Москва, Багратионовская 23 сентября 15:20	11 200 000 ₽ Москва, Багратионовская 15 сентября 12:51



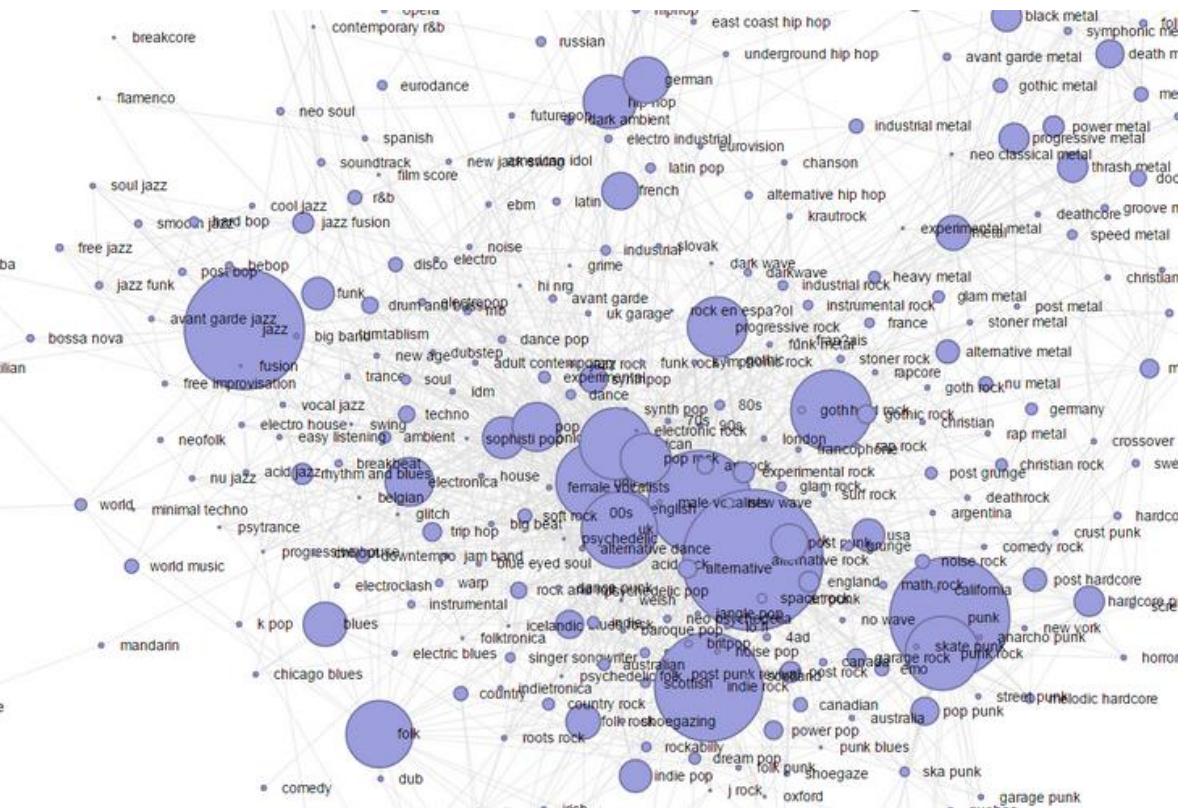
Примеры из задач МегаФона:

Выявление трендов, групп по интересам



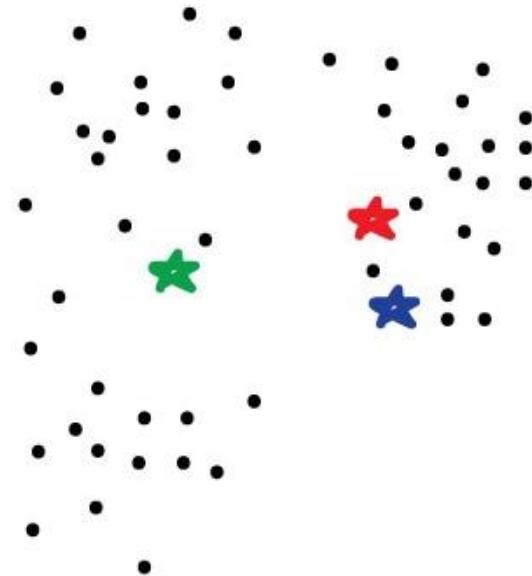
А если смотреть не только в МегаФоне?

- Поиск музыки близких жанров
- Поиск товаров близких по характеристикам



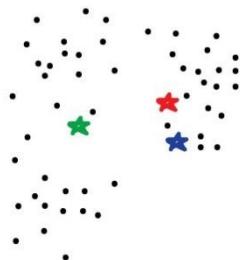
А если смотреть не только в МегаФоне?

Хотим поставить 3
ларька с шаурмой в
маленьком городе

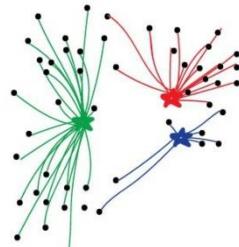


Ставим три ларька с шаурмой оптимальным образом

(илюстрируя метод К-средних)



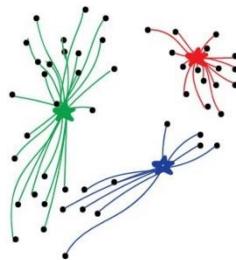
1. Ставим ларьки с шаурмой
в случайных местах



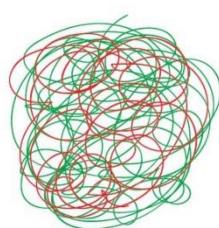
2. Смотрим в какой
кому ближе идти



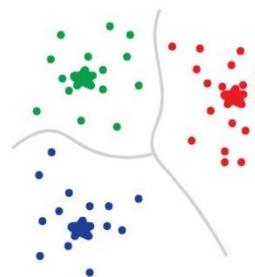
3. Двигаем ларьки ближе
к центрам их популярности



4. Снова смотрим и двигаем



5. Повторяем много раз



6. Готово, вы великолепны!



Обучение без учителя

Где ещё используется?

01

Генерация фич через переход в
другое пространство

02

Объединение близких точек на
карте

03

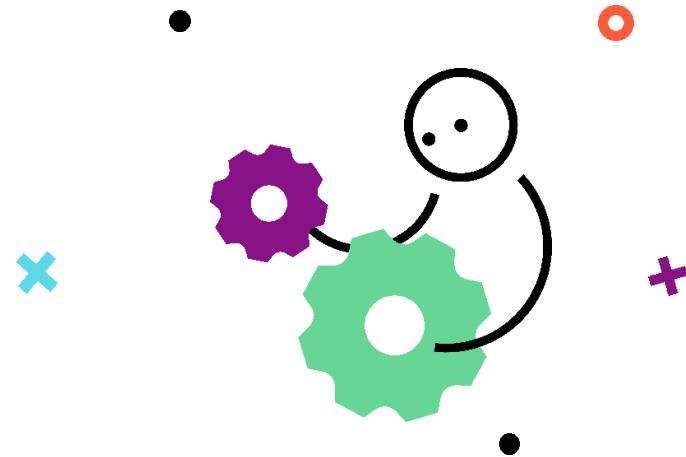
Сжатие изображений

04

Анализ и разметка новых
данных

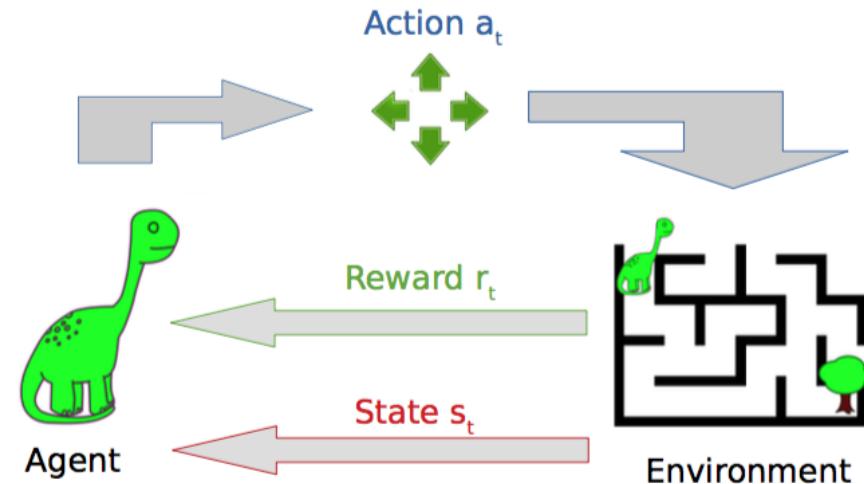
05

Детекторы аномального
поведения



Обучение с подкреплением

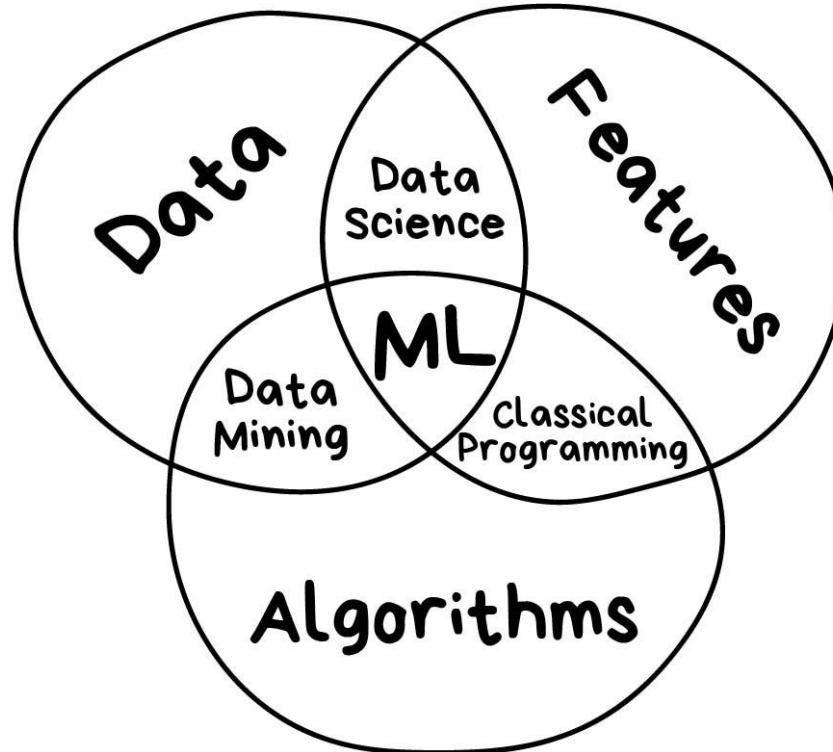
Обучение с подкреплением (reinforcement learning) – один из способов машинного обучения, в ходе которого испытуемая система (агент) обучается, взаимодействуя с некоторой средой.



Основные понятия



Машинное обучение



Модель

Предиктивная модель – это параметрическое семейство функций (семейство гипотез):

$$\mathcal{H} = \{h(x, \theta) \mid \theta \in \Theta\}$$

где

- $h : X \times \Theta \rightarrow Y$
- Θ — множество параметров

Из большого семейства гипотез мы должны выбрать одну, которая с точки зрения меры L является лучшей.

Процесс такого выбора назовем **алгоритмом обучения**:

$$\mathcal{M} : (X \times Y)^n \rightarrow \mathcal{H}$$

Алгоритм

Алгоритм обучения – это отображение из набора данных в пространство гипотез.

Процесс обучения с учителем состоит из двух шагов:

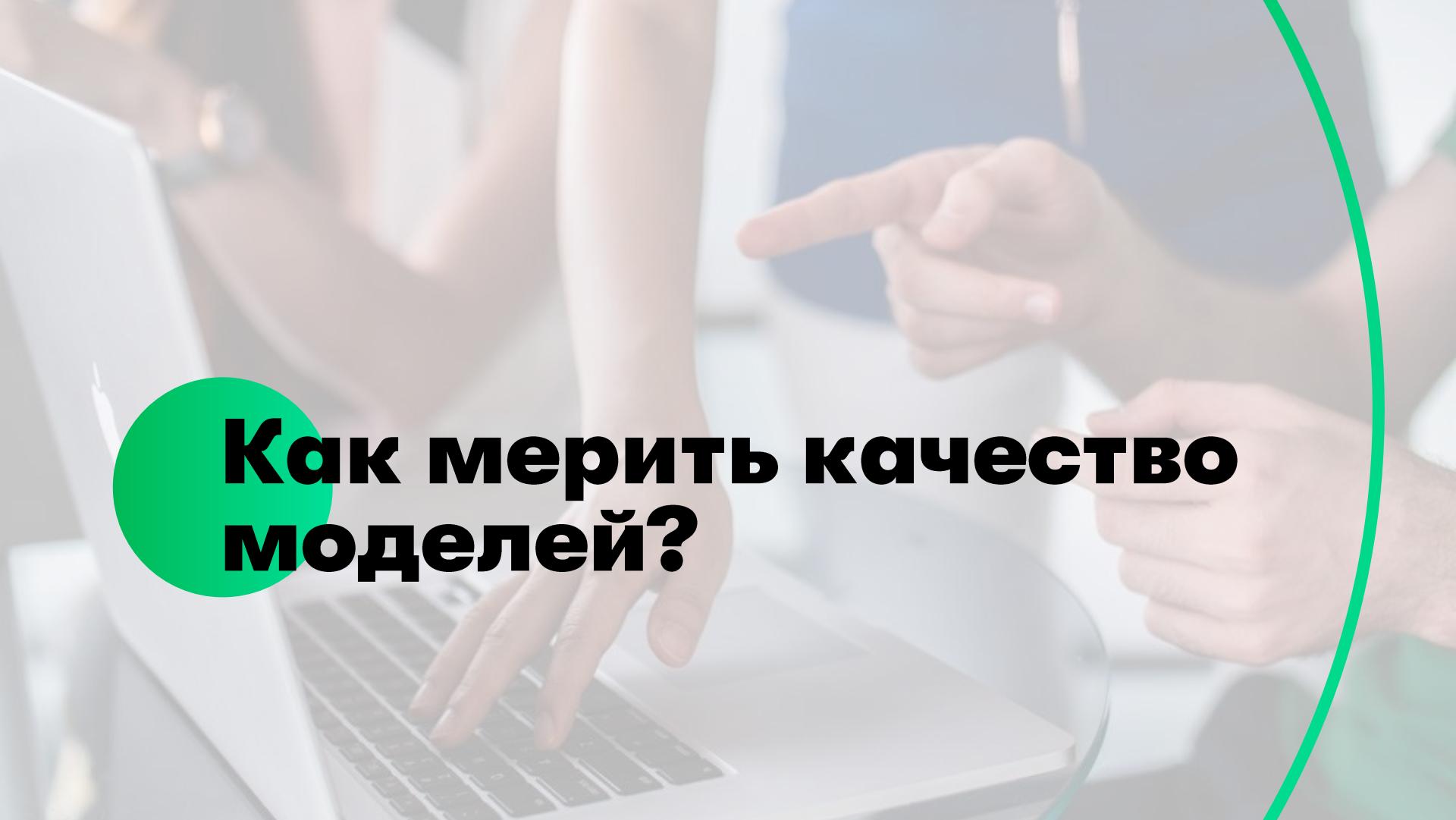
1. Обучение $h = M(D)$

2. Применение $\hat{y} = h(x)$

Часто для обучения модели пользуются **принципом минимизации эмпирического риска**.

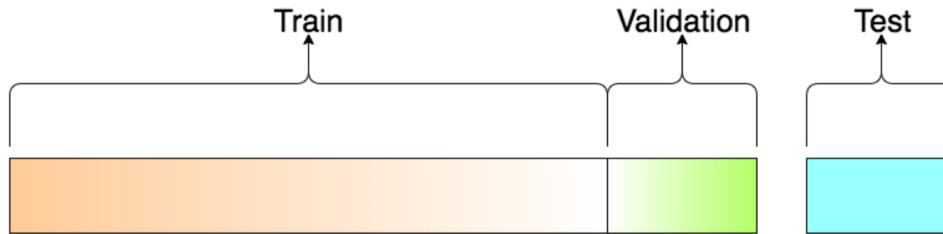
Риском гипотезы h называют ожидаемое значение функции стоимости L .

Модель обладает **обобщающей способностью**, тогда, когда ошибка на новом (тестовом) наборе данных (взятом из того же распределения $P(x,y)$) мала, или же предсказуема.



**Как мерить качество
моделей?**

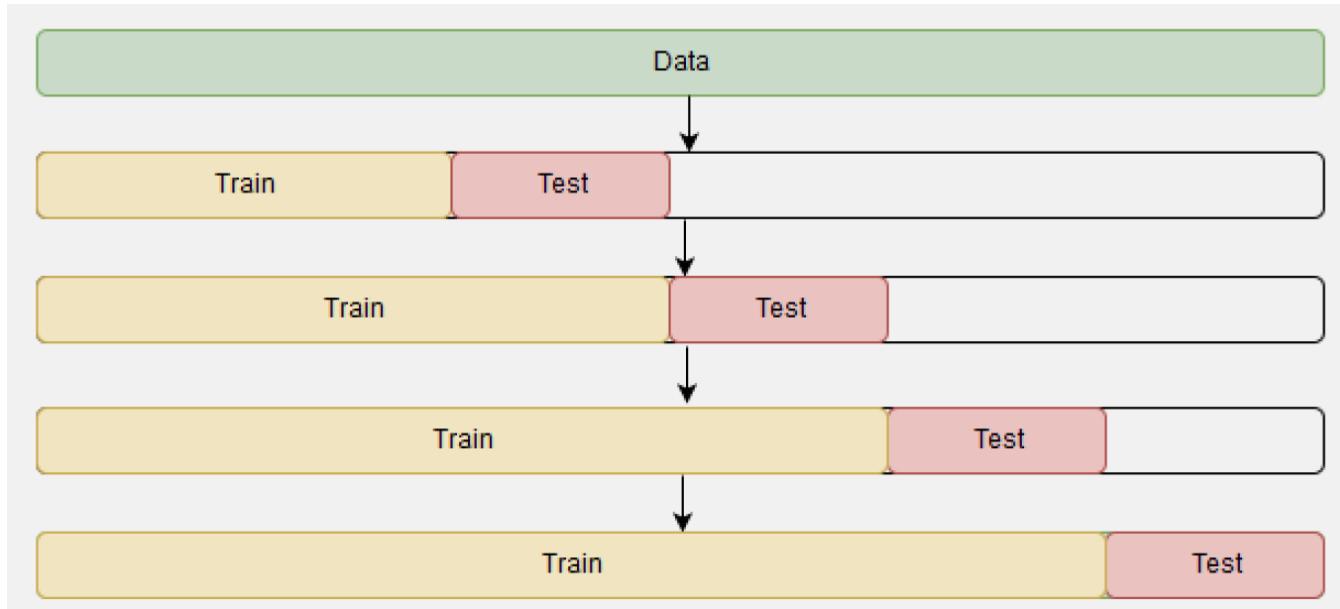
Обучающая и тестовая выборка



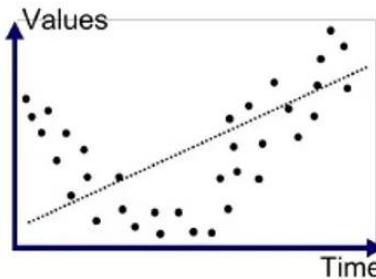
- **Обучающая выборка (training sample)** — выборка, по которой производится настройка (оптимизация параметров) модели.
- **Тестовая выборка (test sample)** — выборка, по которой оценивается качество построенной модели.
- **Проверочная выборка (validation sample)** — выборка, по которой осуществляется выбор лучшей модели из множества моделей, построенных по обучающей выборке.

Обучающая и тестовая выборка

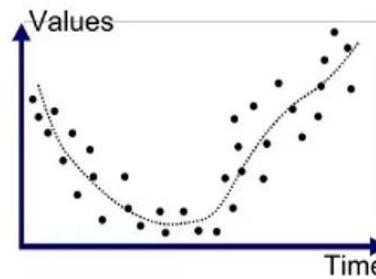
Разбиение временных рядов следует проводить по времени события



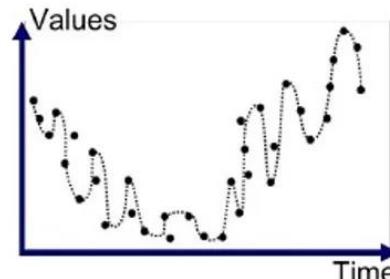
Обобщающая способность



Underfitted



Good Fit/Robust



Overfitted

Переобучение, переподгонка (*overtraining, overfitting*) – нежелательное явление, возникающее при решении задач обучения по прецедентам, когда ошибка обученного алгоритма на объектах тестовой выборки существенно выше, чем средняя ошибка на обучающей выборке.

Переобучение возникает при использовании избыточно сложных моделей.

Недообучение – нежелательное явление, возникающее, когда алгоритм обучения не обеспечивает достаточно малой величины средней ошибки на обучающей выборке.

Типы признаков

и их обработка

Типы признаков

1) Бинарные (флаг подключения услуги)

Binary $\{true, false\}$

2) Номинальные (тарифный план)

Categorical множество значений конечно

3) Количественные (количество мегабайт в месяц) Numerical

\mathbb{R}

4) Порядковые (месяцы, этажи)

Ordinal множество значений конечно и упорядочено

Извлечение признаков

Задача: Необходимо спрогнозировать расходы абонента при переходе на новый ТП

Признаки, характеризующие расходы клиента на связь:

Бинарные	Номинальные	Количественные	Порядковые
Наличие интернета	Тарифный план Город Тип устройства ОС	Количество звонков Количество сообщений Объем трафика	LTV (время жизни клиента)

Работа с изначальными данными

Подготовка данных (Data Preparation)

- 1) Удаление шума
- 2) Заполнение отсутствующих значений
- 3) Трансформация значений
- 4) Использование знаний о предметной области

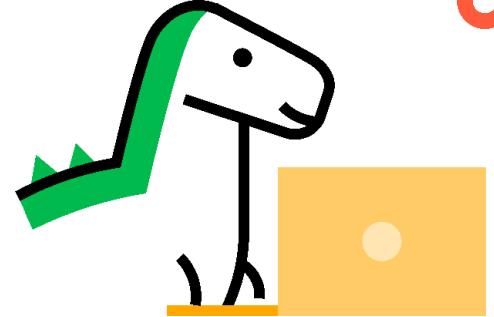
Извлечение признаков – дата/время



Даты и время

- 1) Абсолютное время события (2019:05:2127:31:00) ✖
- 2) Периодичность (месяц, неделя, день и т.д.)
- 3) Временной интервал до или после особого события
(праздник, распродажа, день выдачи зарплаты и т.д.)

Преобразование признаков



Категориальные признаки

Label Encoding

Пример: имеется текстовое описание признаков

Не подходит для линейных моделей



The diagram illustrates the process of transforming categorical features into numerical values. On the left, a table shows four rows of categorical data under the column 'Feature': 'School', 'Basic', 'University', and 'School'. A large blue arrow points from this table to another table on the right, which shows the same four rows but with numerical values: 1, 0, 2, and 1 respectively, under the column 'Feature'.

	Feature
1	School
2	Basic
3	University
4	School

→

	Feature
1	1
2	0
3	2
4	1



Преобразование признаков

Категориальные признаки

One-Hot Encoding

Пример: имеется текстовое описание признаков

The diagram illustrates the process of One-Hot Encoding. On the left, there is a table with four rows and two columns. The first column is labeled 'Feature' and the second column contains the values 'School', 'Basic', 'University', and 'School'. A large blue arrow points from this table to another table on the right. The right table has four rows corresponding to the rows in the first table. It has four columns: the first column is empty, the second column is labeled 'F=School', the third column is labeled 'F=Basic', and the fourth column is labeled 'F=University'. The values in the 'F=School' column are 1, 0, 0, and 1 respectively. The values in the 'F=Basic' column are 0, 1, 0, and 0 respectively. The values in the 'F=University' column are 0, 0, 1, and 0 respectively.

	Feature
1	School
2	Basic
3	University
4	School

	F=School	F=Basic	F=University
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0

Преобразование признаков

Категориальные признаки

Hashing trick

The diagram illustrates the 'Hashing trick' for transforming categorical features into binary vectors. On the left, a table lists four categories: School, Basic, University, and School. An arrow points from this table to a second table on the right. The second table has three columns: F=S (Feature is School), F=B,U (Feature is Basic or University), and an unlabeled column. The rows correspond to the four categories from the first table. The values in the F=S column indicate whether the feature is 'School' (1) or not (0). The values in the F=B,U column indicate whether the feature is 'Basic' or 'University' (1) or not (0).

	Feature		
1	School		
2	Basic		
3	University		
4	School		

	F=S	F=B,U
1	1	0
2	0	1
3	0	1
4	1	0

Нормализация

Различные модели по-разному реагируют на возможные значения входных признаков

1) Standart Scaling

$$x' = \frac{x - \bar{x}}{\sigma}$$

2) MinMax Scaling

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Очистка данных



1) Удаление или преобразование пропущенных (неопределенных) данных – многие модели не допускают во входных данных пропуски



2) Удаление «нуль-вариантных» переменных (числовых и номинальных) – для многих моделей это может привести к краху или к нестабильной работе.

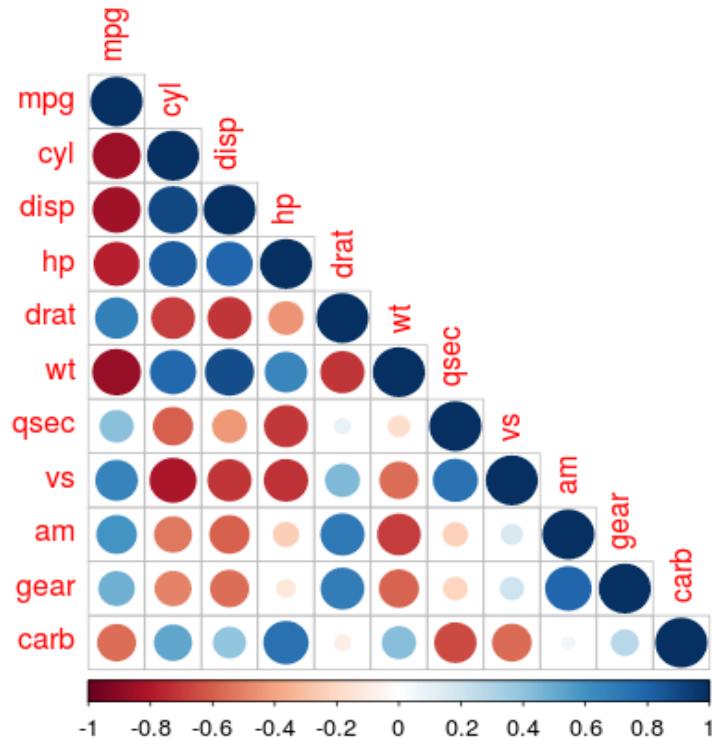


3) Выявление и удаление коррелированных предикторов (числовых) – некоторые модели отлично справляются с коррелированными предикторами (например PLS, LARS и подобные, использующие L1 регуляризацию), другие модели могут получить преимущества от снижения уровня корреляции между предикторами.

Корреляция

Корреляция Пирсона:

$$r_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2(Y - \bar{Y})^2}}$$



Пропуски



Большинство реальных данных имеют пропущенные значения:

- Ошибки при записи
- Ошибки при измерении
- Невозможность сбора

Далеко не все алгоритмы умеют работать с неполными данными

Заполнение пропусков

- Заменять наиболее вероятным – в случае непрерывных данных замена на среднее значение из наиболее вероятного интервала; в дискретном случае – выбирается значение с наибольшей вероятностью.
- Заменять случайными значениями – замена пропусков на случайное значение из распределения.
- Заменять средним/медианой
- Заменять значением Не задано – доступно только для дискретного поля, выполняется замена пропусков на значение «Не задано».
- Удалять записи – строки с выявленными пропусками исключаются из набора данных. Метод недоступен для упорядоченных рядов. **Ничего не испортим, но что если данных и так мало?**
- Заполняем прогнозным значением

Методы отбора признаков

1. Одномерный отбор признаков

Отбор признаков по взаимосвязи с целевой переменной, могут быть отобраны с помощью статистических критериев (например, хи-квадрат).

2. Рекурсивное исключение признаков

Метод рекурсивного исключения признаков (recursive feature elimination, RFE)

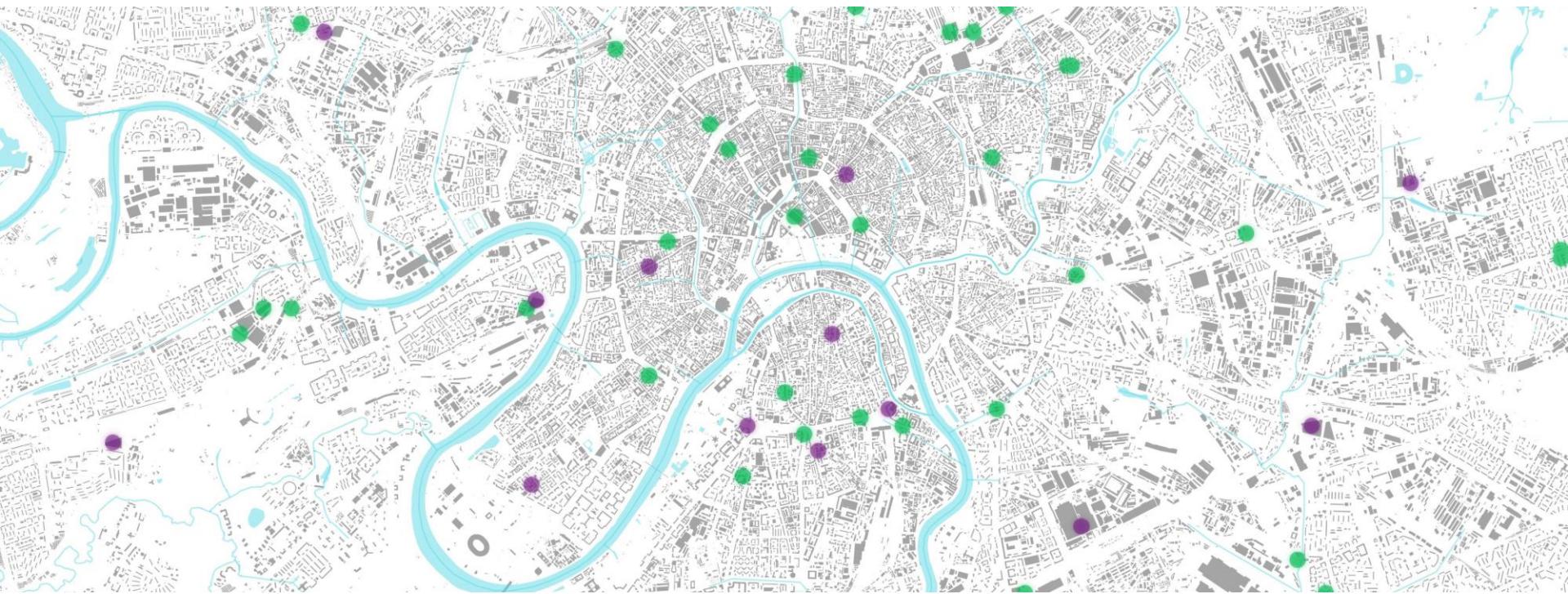
3. Метод главных компонент

Метод главных компонент (principal component analysis, PCA) позволяет уменьшить размерность данных с помощью преобразования на основе линейной алгебры

4. Отбор на основе важности признаков

Ансамблевые алгоритмы на основе деревьев решений, такие как случайный лес (random forest), позволяют оценить важность признаков

Извлечение признаков – гео-данные



Категории гео-данных



Данные по расположению
(широта, долгота,
номер базовой станции)



Данные дорожной
инфраструктуры



Данные коммерческой
инфраструктуры



Данные по населению



Данные по абонентам

Задачи анализа гео-данных



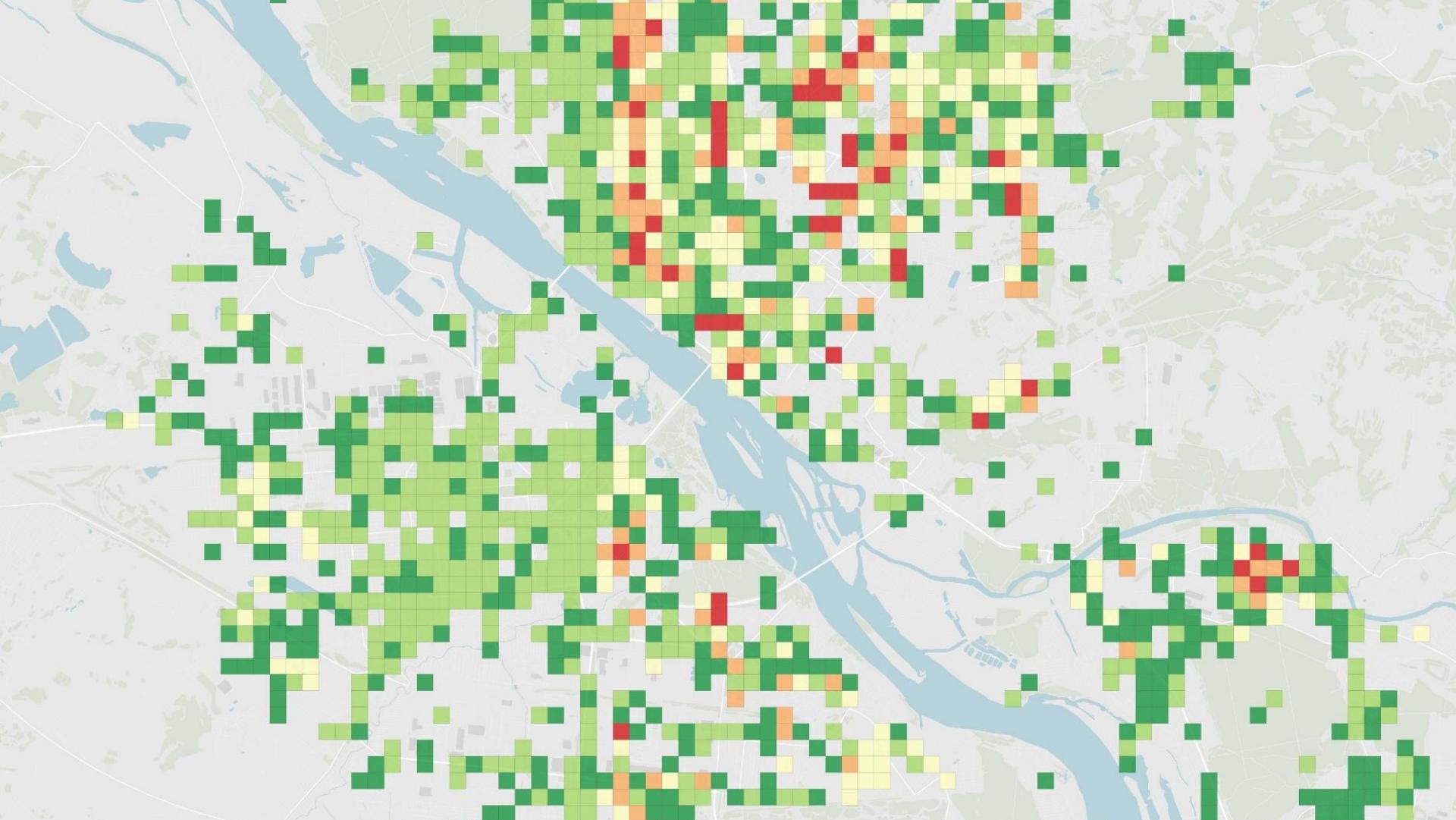
Клиентопоток в определенной локации зависит от географии места



Любимые локации



Скорость и направления перемещения абонентов – автовладельцы, пассажиры метро,...

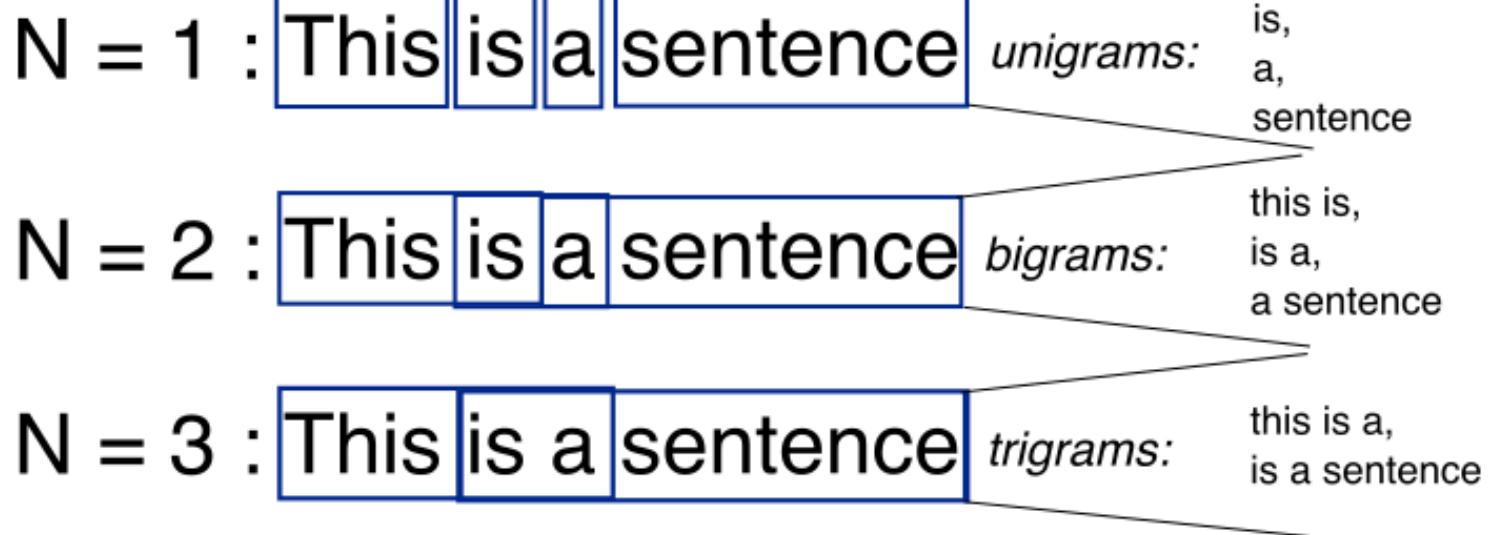


Извлечение признаков – тексты

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



Извлечение признаков – n-grams



Извлечение признаков – TF-IDF

TF (term frequency – частота слова)

Отношение числа вхождений некоторого слова к общему числу слов документа.

$$tf(t, d) = \frac{n_t}{\sum_k n_k} ,$$

n_t – число вхождений слова t в документ

$Z_d n_d$ – общее количество слов в документе

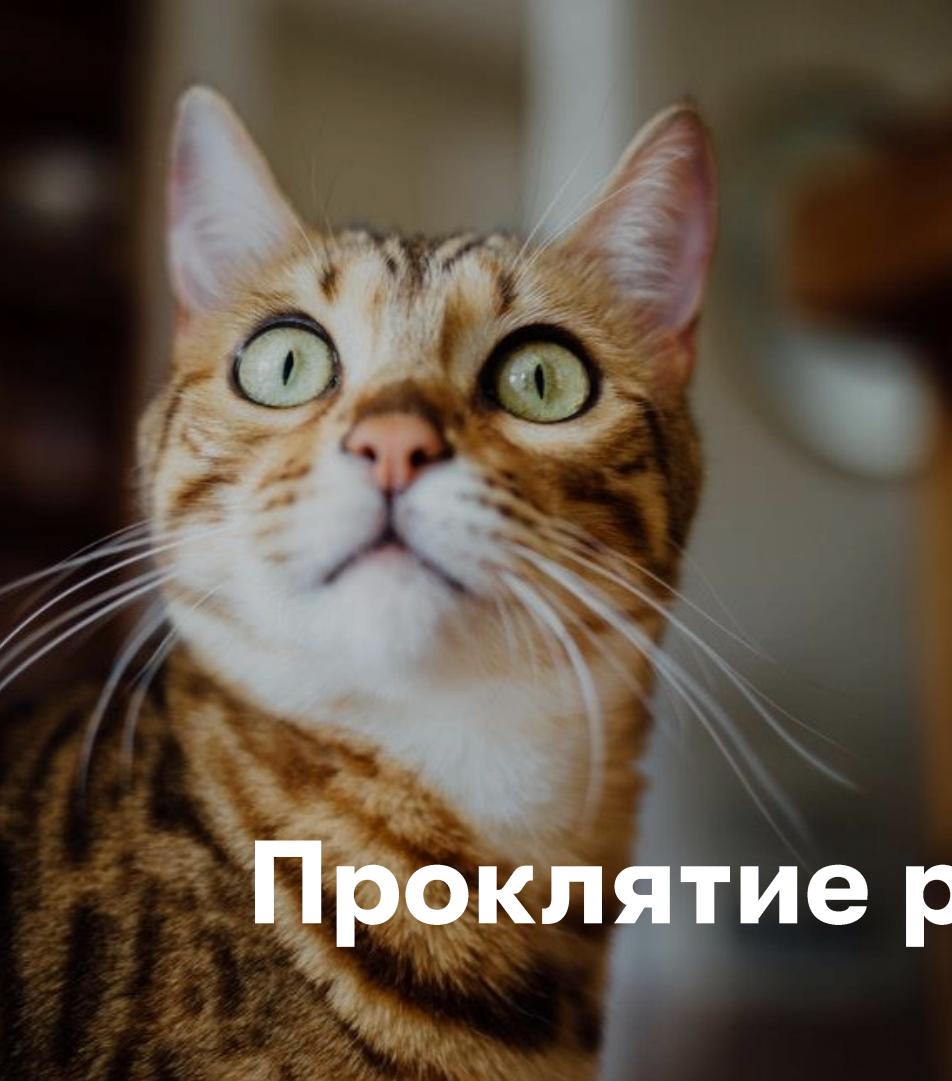
IDF (inverse document frequency – обратная частота документа)

Инверсия частоты, с которой некоторое слово встречается в документах коллекции.

$$idf(t, D) = \log \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|}$$

$|D|$ – количество документов в корпусе
 $|\{d_i \in D \mid t_i \in d_i\}|$ – число документов в коллекции, в которых встречается слово t_i

$$tf - idf_{t,d,D} = tf_{t,d} * idf_{t,D}$$



Проклятие размерности?

Проклятие размерности

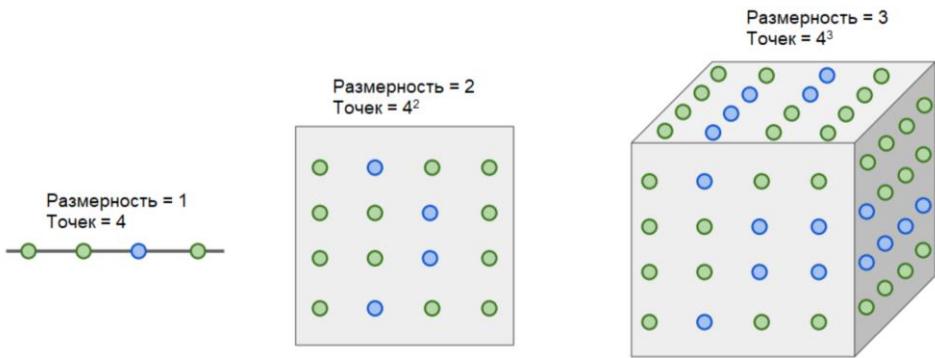
Проклятие размерности – проблема, связанная с экспоненциальным ростом количества данных из-за увеличения размерности пространства, в сложных системах с большим количеством параметров.

$$\lim_{d \rightarrow \infty} \frac{\text{dist}_{\max} - \text{dist}_{\min}}{\text{dist}_{\min}} \rightarrow 0$$

Влечет следующие трудности:

1. Трудоемкость вычислений
2. Хранение огромного количества данных
3. Увеличение доли шумов
4. В линейных классификаторах ведет к проблемам мультиколлинеарности и переобучения.
5. В метрических классификаторах к снижению информативности.

Основная идея — понизить размерность пространства, спроектировать данные на подпространство меньшей размерности (например, с помощью метода главных компонент).



Отбор признаков (feature selection)

Отбор признаков – это выбор признаков, имеющих наиболее тесные взаимосвязи с целевой переменной.

Обеспечивает три основных преимущества:

- 1. Уменьшение переобучения.** Чем меньше избыточных данных, тем меньше возможностей для модели принимать решения на основе «шума».
- 2. Повышение точности.** Чем меньше противоречивых данных, тем выше точность.
- 3. Сокращение времени обучения.** Чем меньше данных, тем быстрее обучается модель.

Стоимость признаков

Время на вычисление признаков

Использование оперативной памяти

Ресурсы на получение дополнительных данных

Дополнительная нестабильность

Добавление признаков в модели классификации позволяет повысить их качество, но требует дополнительного контроля их стабильности (Data Quality)

Метрические методы в обучении с учителем k-NN

A photograph of a man skydiving. He is wearing a black helmet with a mounted camera, dark goggles, and an orange and white patterned skydiving suit. He has his mouth wide open, shouting or screaming. He is holding onto a metal frame of a skydiving harness. In the background, another person is visible in the distance under a clear blue sky.

Скажи мне
друг ...

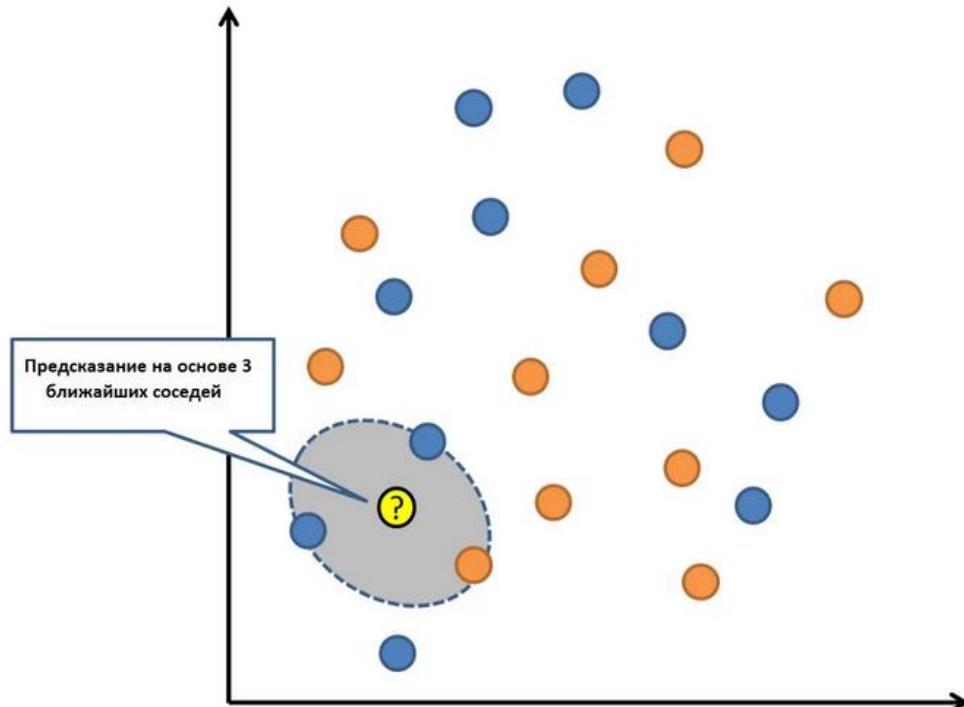
КТО ТВОЙ

и я скажу кто ты!

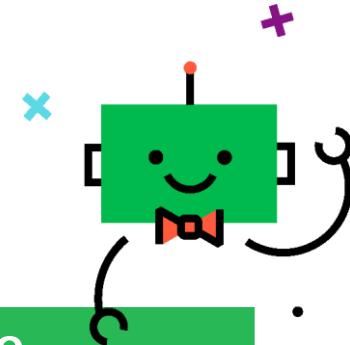
kNN

К-ближайших соседей — очень простой и очень эффективный алгоритм.

Предсказание для новой точки проводится путём поиска K ближайших соседей в наборе данных и усреднения выходной переменной для этих K экземпляров.



kNN – характеристики



Продукт	Сладость	Хруст	Класс
яблоко	9	8	фрукт
огурец	2	7	овощ
банан	10	1	фрукт
бекон	1	4	протеин
...

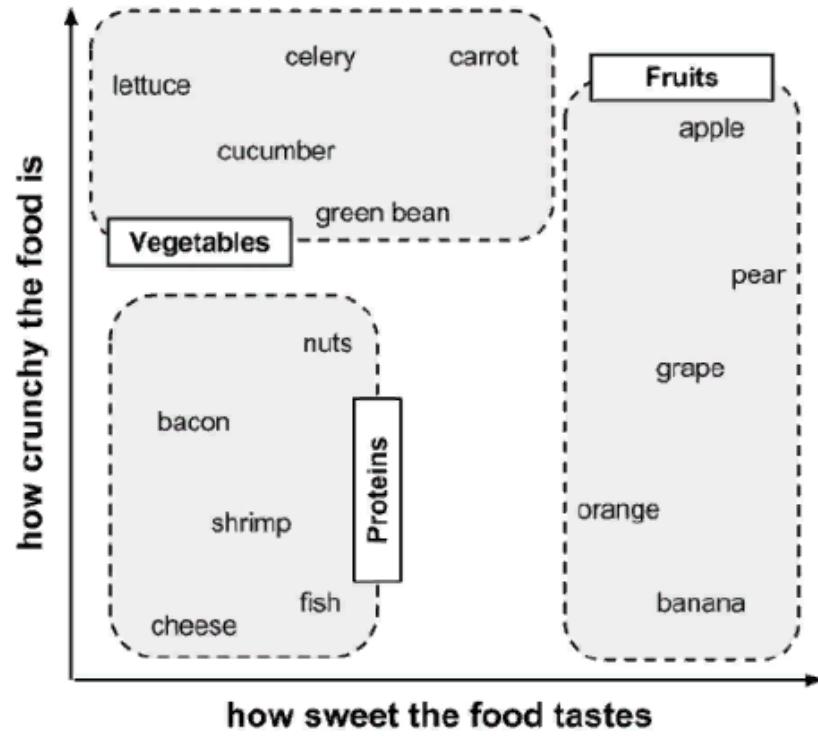
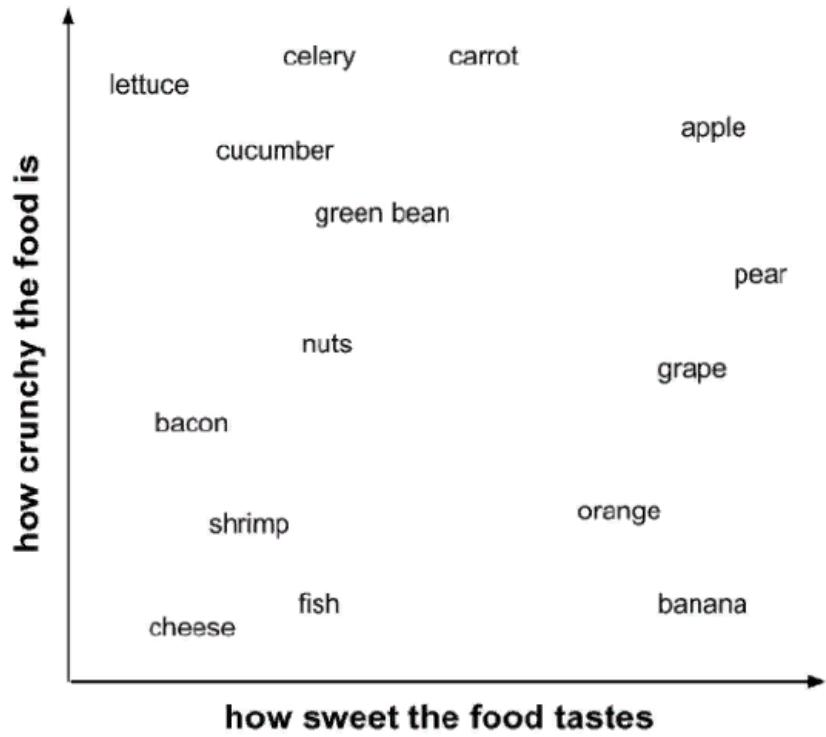
Гипотеза компактности

kNN – гипотеза компактности

Если мера сходства объектов введена достаточно удачно, то схожие объекты гораздо чаще лежат в одном классе, чем в разных.

В этом случае, граница между классами имеет достаточно простую форму, а классы образуют компактно локализованные области в пространстве объектов

kNN



kNN

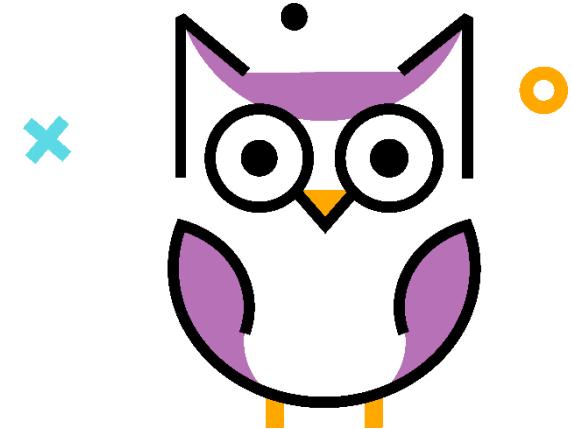
Для классификации каждого из объектов тестовой выборки необходимо последовательно выполнить следующие операции:

1. Вычислить расстояние до каждого из объектов обучающей выборки
2. Отобрать k объектов обучающей выборки, расстояние до которых минимально
3. Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей

kNN – вариации алгоритма

- 01** Метод ближайшего соседа — классифицируемый объект относится к тому классу, которому принадлежит ближайший к нему объект обучающей выборки.
- 02** Метод k-ближайших соседей (k-Nearest Neighbors) — для повышения надёжности классификации объект относится к тому классу, которому принадлежит большинство из его соседей — k ближайших к нему объектов обучающей выборки x_i . В задачах с двумя классами число соседей берут нечётным.
- 03** Метод взвешенных ближайших соседей — в задачах с числом классов 3 и более нечётность уже не помогает и ситуации неоднозначности всё равно могут возникать. Тогда i -му соседу приписывается вес w_i , как правило, убывающий с ростом ранга соседа i . Объект относится к тому классу, который набирает больший суммарный вес среди k ближайших соседей.

kNN



Выборка: $X_m = \{(x_1, y_1), \dots, (x_m, y_m)\}$

Функция расстояния: $\rho(x, x')$

Нумерация объектов: $\rho(u, x_{1;u}) \leq \rho(u, x_{2;u}) \leq \dots \leq \rho(u, x_{m;u})$

Задача: присвоить объекту u лейбл

Формальный алгоритм:

$$a(u) = \operatorname{argmax}_{y \in Y} \sum_{i=1}^m [y_{i;u} = y] w(i, u)$$

где $w(i, u)$ – заданная весовая функция, которая оценивает степень важности i -го соседа для классификации объекта u .

Метрики расстояния

kNN – метрики расстояния

$$\{A \mid \rho(A, O) \leq 1\}$$

- 1. Расстояние Минковского

$$\rho_p(A, B) = (|x_2 - x_1|^p + |y_2 - y_1|^p)^{1/p}$$

- 2. Манхэттенское расстояние

$$\rho_1(A, B) = |x_2 - x_1| + |y_2 - y_1|$$

- 3. Евклидово расстояние

$$\rho(A, B) = (|x_2 - x_1|^2 + |y_2 - y_1|^2)^{1/2}$$

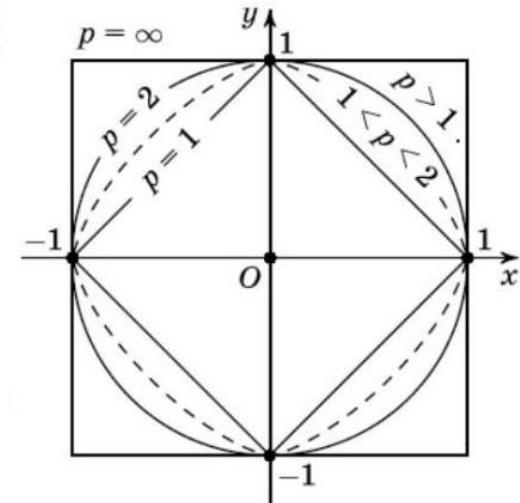
- 4. Расстояние Чебышева

$$\rho' = \max(|x_2 - x_1|, |y_2 - y_1|)$$

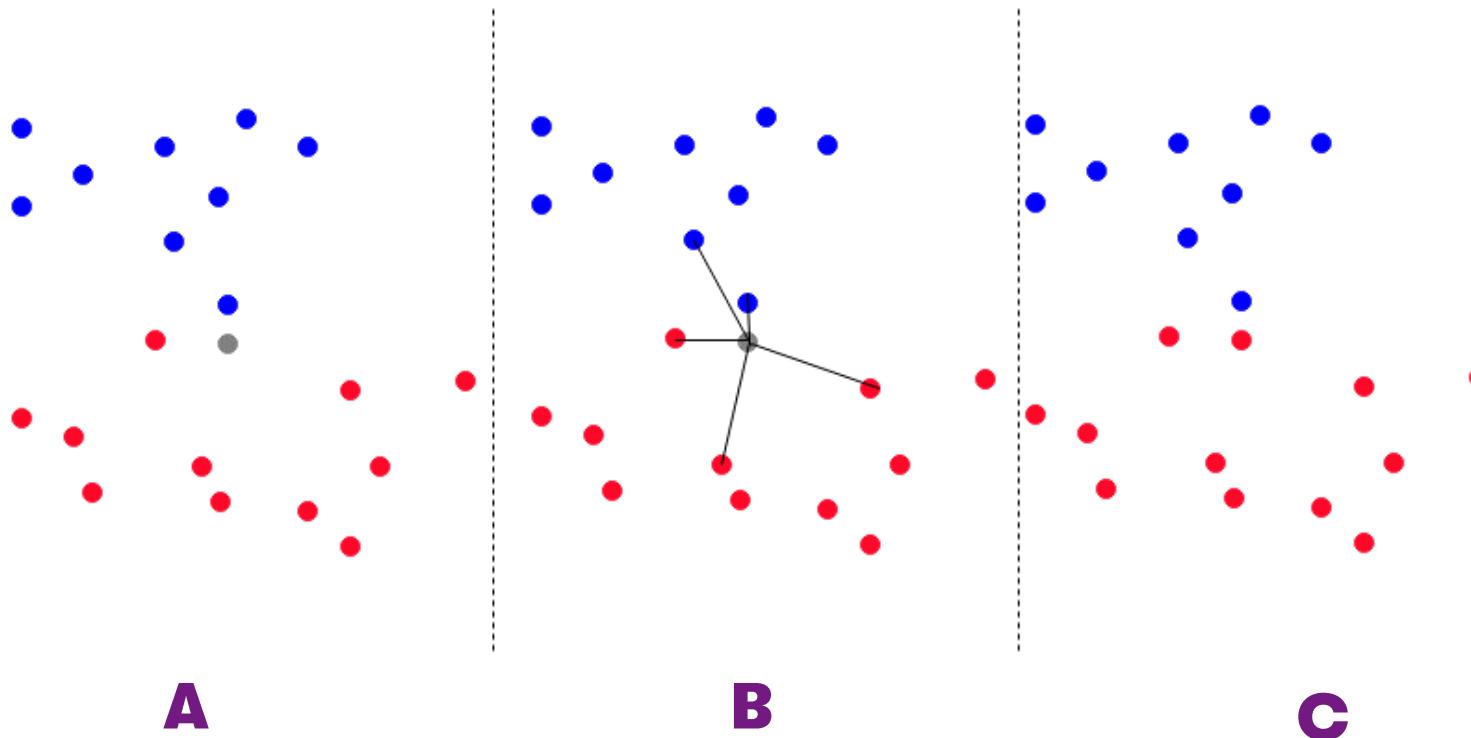
- 5. Косинусная мера

- 6. Коэффициент корреляции

...



kNN – пример работы алгоритма



A

B

C

kNN – основные недостатки

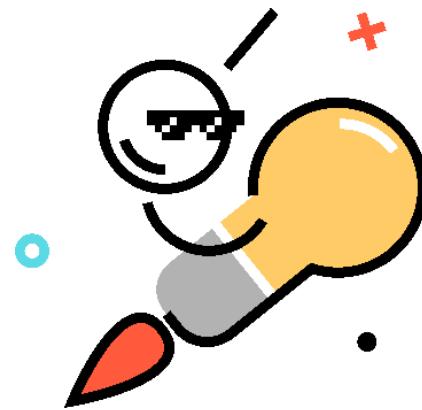
- Необходимость хранить обучающую выборку **целиком**
- Бедный набор параметров
- Затраты в производительности велики

kNN - плюсы

- Алгоритм устойчив к аномальным выбросам
- Простая программная реализация
- Результат легко поддаётся интерпретации



kNN – область применения



1. **Baseline** в решении какой-либо задачи
2. Метод **kNN** часто используется **как составная часть** более сложного алгоритма классификации (композиции алгоритмов)
3. **Рекомендательные системы** – простым начальным решением может быть рекомендация какого-то товара (или услуги), популярного среди **ближайших соседей** человека, которому хотим сделать рекомендацию.
4. **Предсказание отклика клиентов** – можно определить отклик новых клиентов по данным из прошлого.

Итог



К следующей лекции:

