

Постановка задач машииного обучения

Виктор Кантор

На этой лекции

- I. Вспоминаем стандартные задачи и методы
- II. Оптимизационные задачи
- III. Примеры

На этой лекции

- I. Вспоминаем стандартные задачи и методы
- II. Оптимационные задачи
- III. Примеры
 - исправление опечаток
 - оптимизация бюджета рекламных кампаний
 - рекомендации товаров
 - прогнозирование дефектов на производстве
 - прогнозирование оттока и удержание
 - оптимизация затрат материалов на производстве
 - очереди в магазинах
 - ранжирование
 - прогнозирование пробок
 - прогнозирование спроса
 - назначение водителей в такси

I. Вспоминаем стандартные задачи и
методы

Стандартные задачи машинного обучения

Классификация



Iris setosa



Iris versicolor



Iris virginica

Вход (обучающая выборка):

Признаки N объектов с известными классами

Выход:

Классификатор (алгоритм, прогнозирующий классы новых объектов по их признакам)

Классификация: обучающая выборка

Fisher's Iris Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.0	2.0	3.5	1.0	<i>I. versicolor</i>
6.0	2.2	5.0	1.5	<i>I. virginica</i>
6.2	2.2	4.5	1.5	<i>I. versicolor</i>
6.0	2.2	4.0	1.0	<i>I. versicolor</i>
6.3	2.3	4.4	1.3	<i>I. versicolor</i>
5.5	2.3	4.0	1.3	<i>I. versicolor</i>
5.0	2.3	3.3	1.0	<i>I. versicolor</i>
4.5	2.3	1.3	0.3	<i>I. setosa</i>
5.5	2.4	3.8	1.1	<i>I. versicolor</i>
5.5	2.4	3.7	1.0	<i>I. versicolor</i>
4.9	2.4	3.3	1.0	<i>I. versicolor</i>
6.7	2.5	5.8	1.8	<i>I. virginica</i>
5.7	2.5	5.0	2.0	<i>I. virginica</i>
6.3	2.5	5.0	1.9	<i>I. virginica</i>
6.3	2.5	4.9	1.5	<i>I. versicolor</i>
4.9	2.5	4.5	1.7	<i>I. virginica</i>

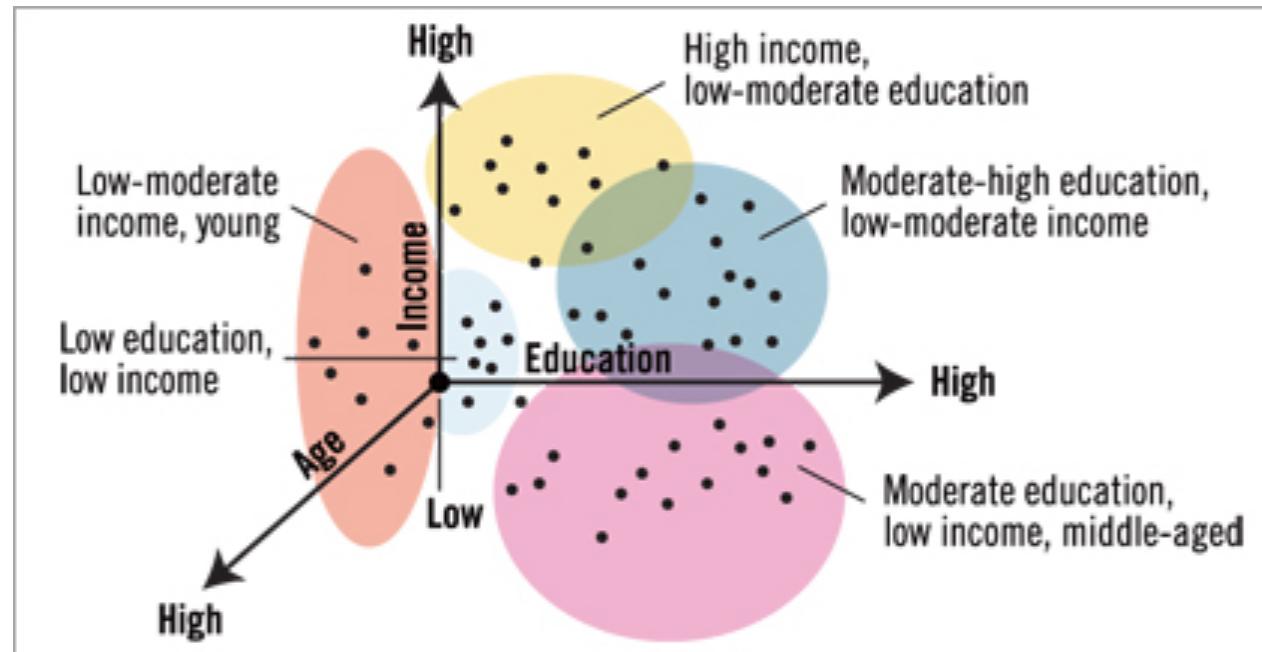
Кластеризация

Вход (обучающая выборка):

Признаки N объектов

Выход:

Найденные в выборке классы (кластеры), метки кластеров для объектов из обучающей выборки и алгоритм отнесения новых объектов к кластеру



Пример: сегментация рынка

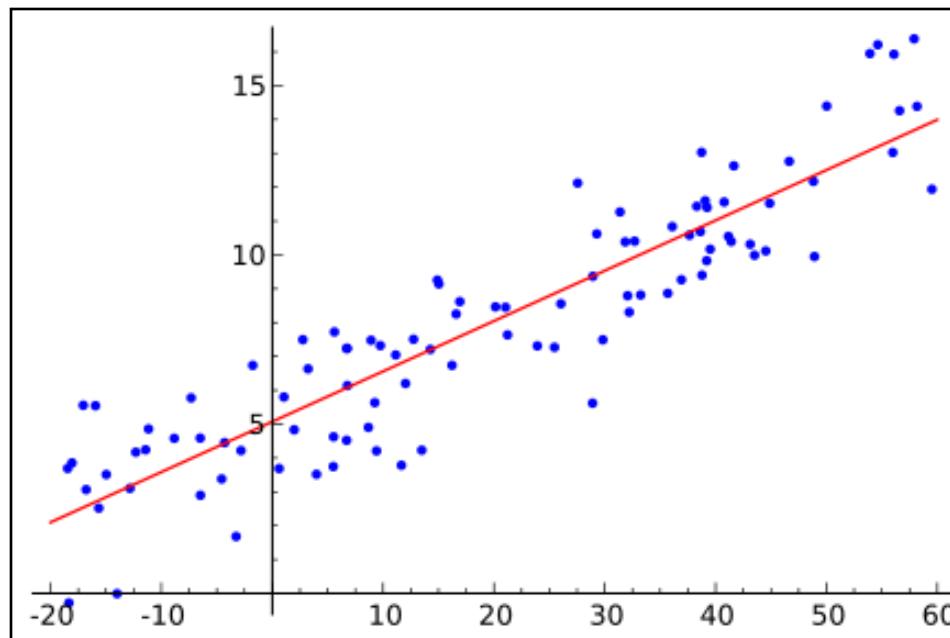
Регрессия

Вход (обучающая выборка):

Признаки N объектов с известными значениями прогнозируемого вещественного параметра объекта

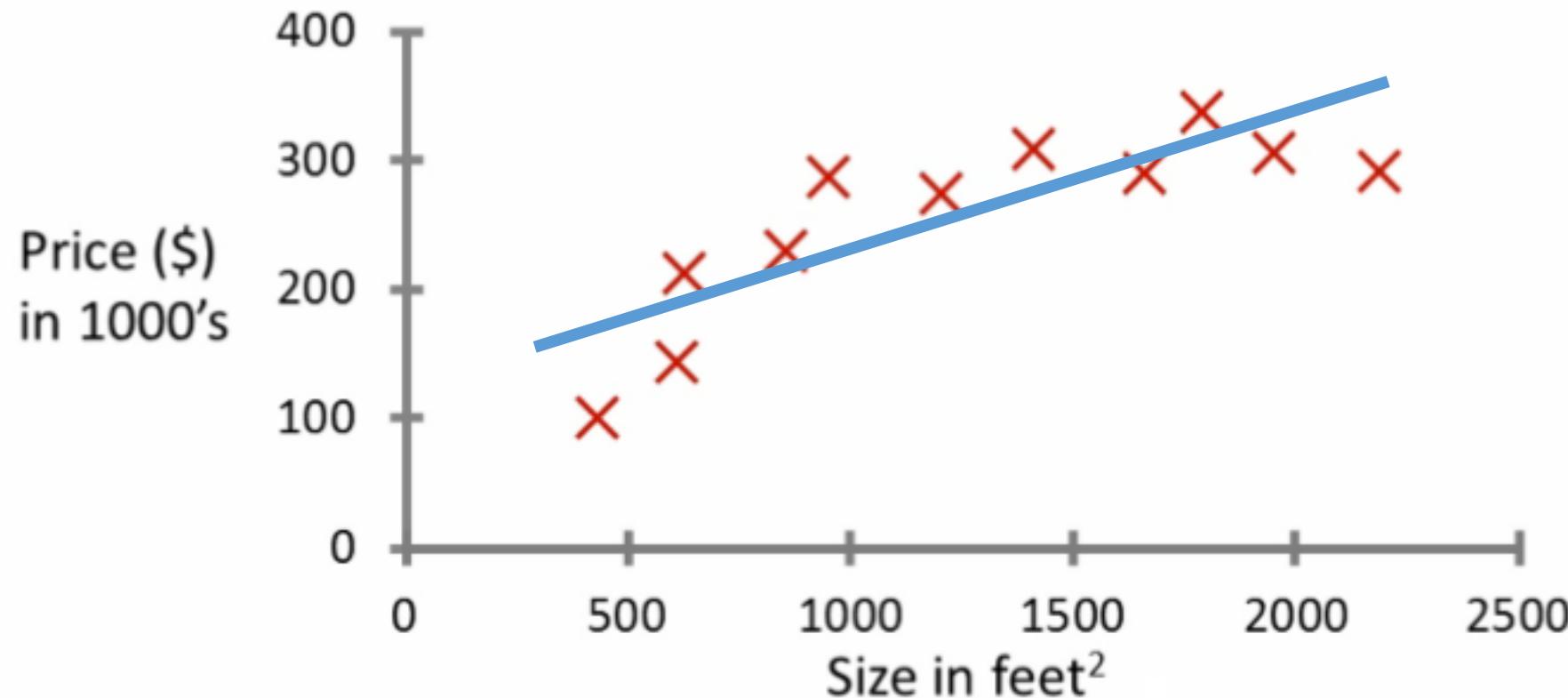
Выход:

Алгоритм, прогнозирующий значение вещественной величины по признакам объекта

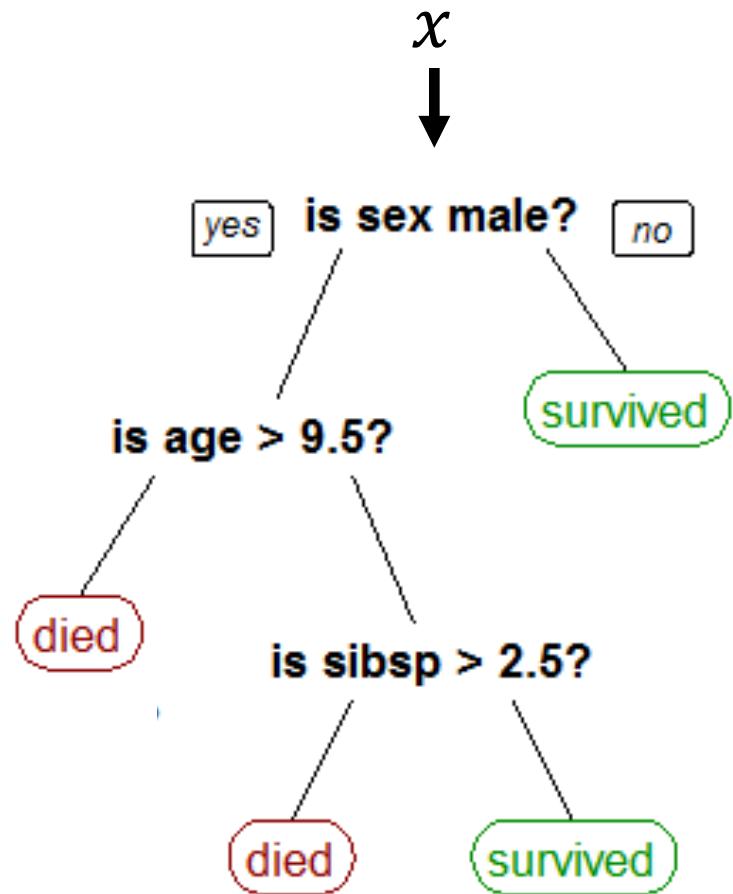


Наиболее часто используемые методы

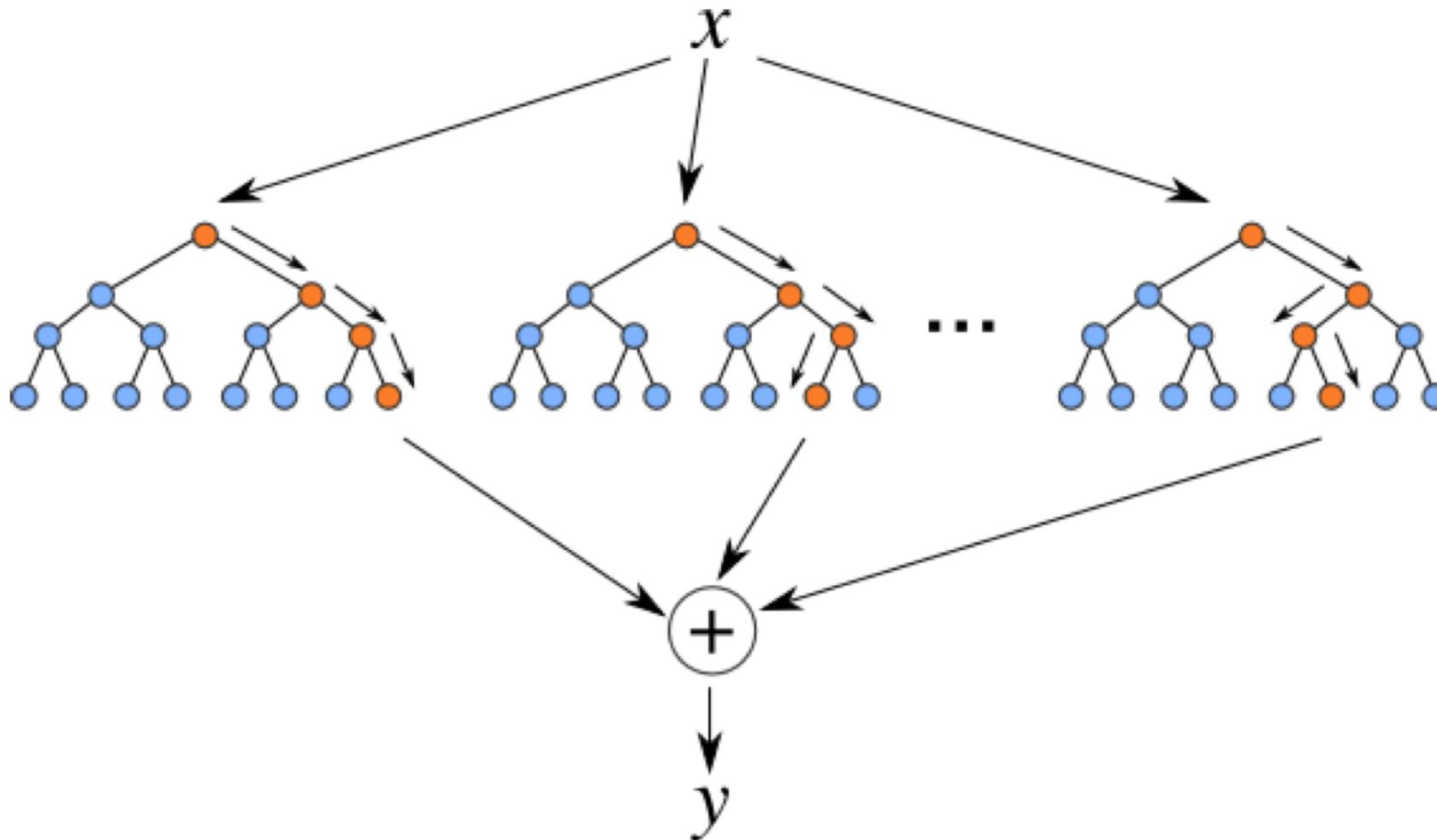
Линейные модели



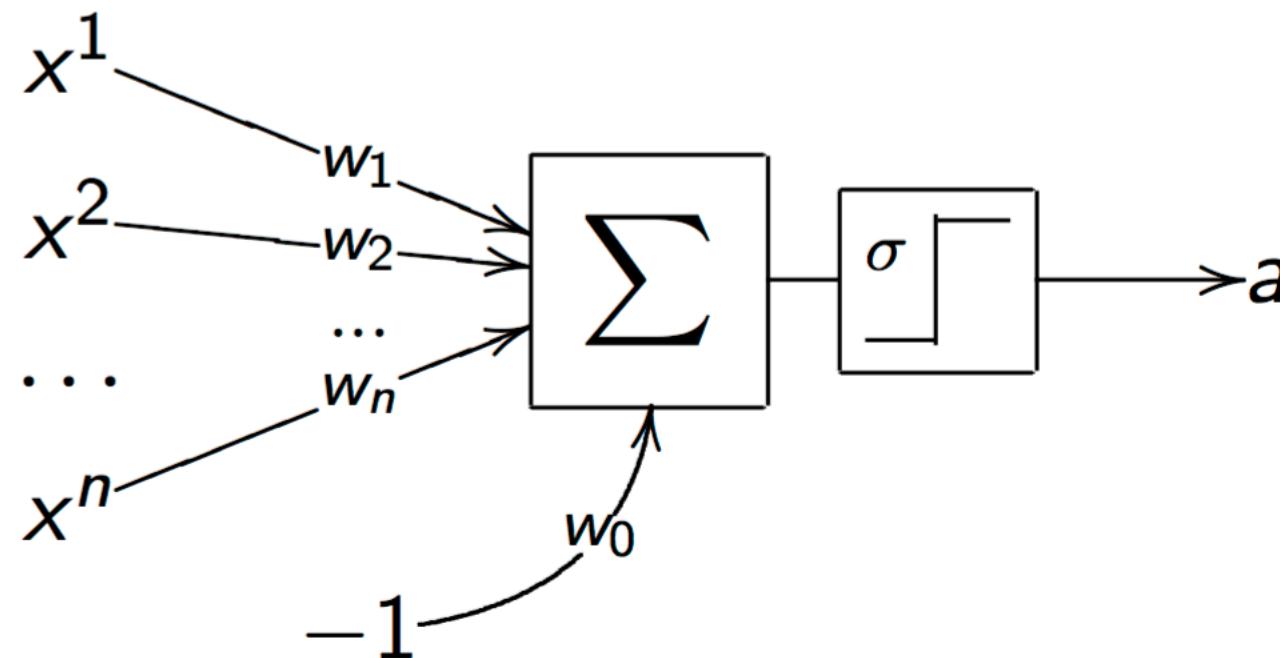
Решающие деревья



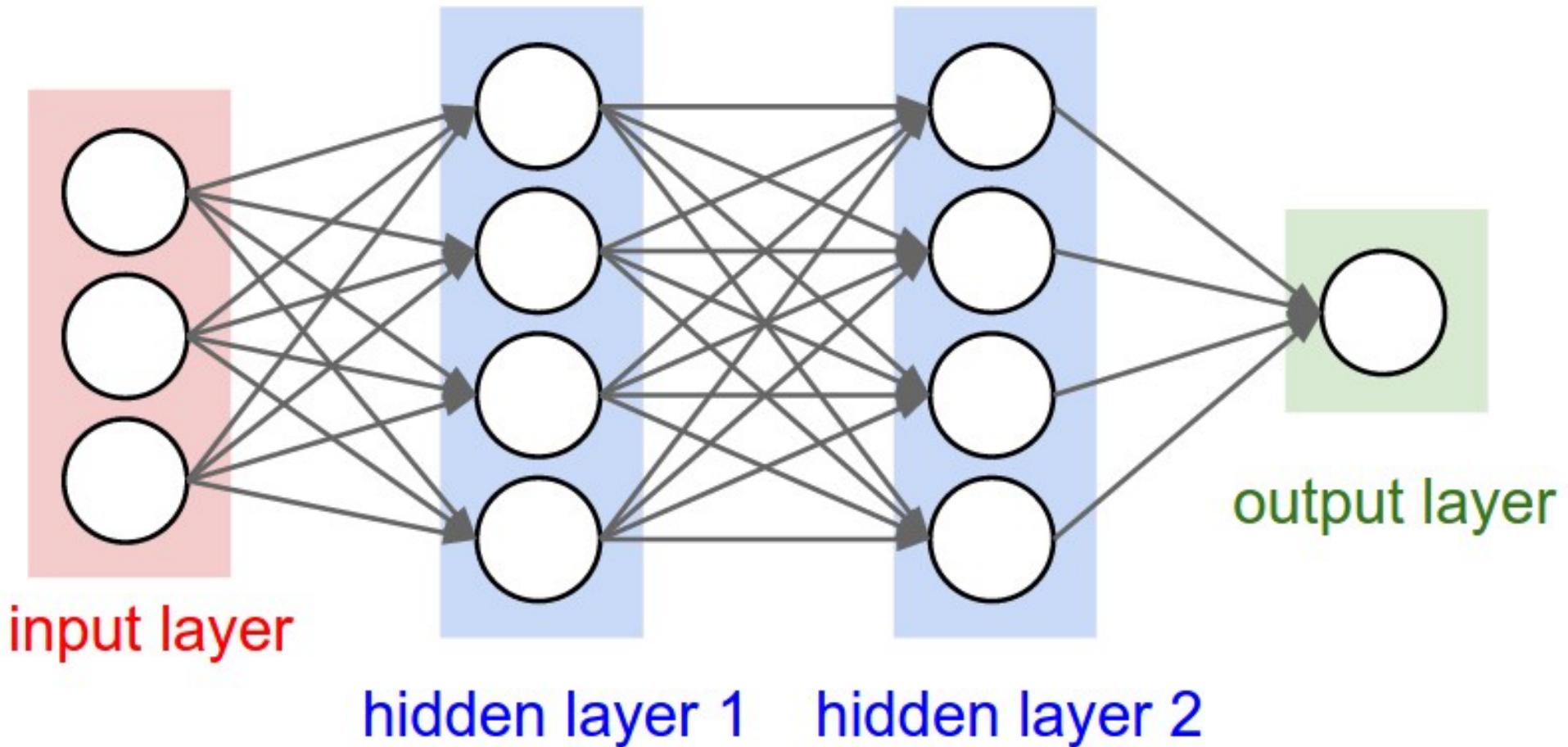
Ансамбли решающих деревьев



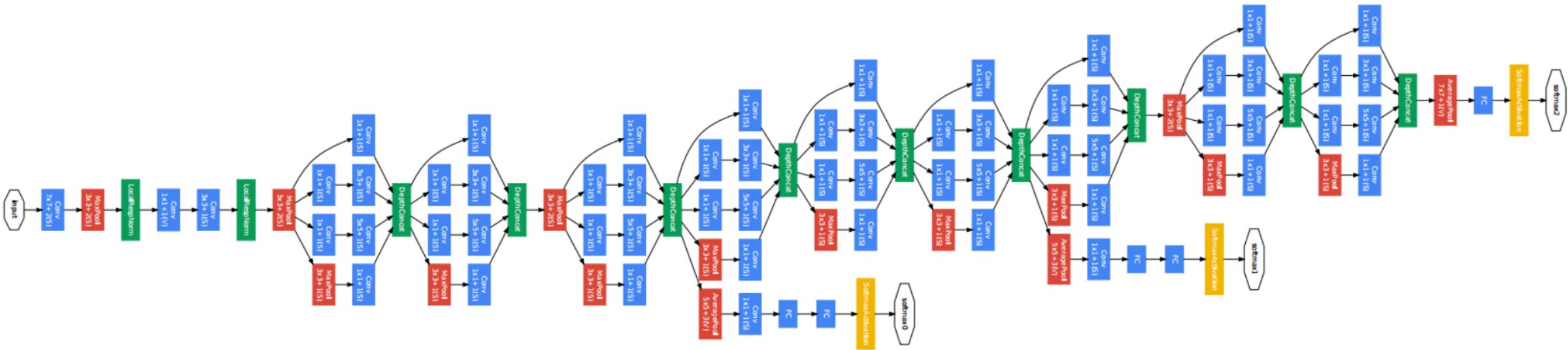
Нейронные сети



Нейронные сети



Нейронные сети



GoogLeNet

II. Оптимизационные задачи

Задача регрессии

x_1, x_2, \dots, x_l - точки, в которых известны значения некоторой величины:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i \approx a(x_i)$$

Но что значит «примерно равно»?

Задача регрессии

x_1, x_2, \dots, x_l - точки, в которых известны значения некоторой величины:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i \approx a(x_i)$$

Но что значит «примерно равно»?

Например, это:

$$\sum_{i=1}^l (y_i - a(x_i))^2 \rightarrow \min$$

Задача регрессии

x_1, x_2, \dots, x_l - точки, в которых известны значения некоторой величины:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i \approx a(x_i)$$

Но что значит «примерно равно»?

В общем случае:

$$\sum_{i=1}^l L(y_i, a(x_i)) \rightarrow \min$$

Задача классификации

x_1, x_2, \dots, x_l - объекты, для которых известны их классы:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i = a(x_i)$$

Как выразить то, что он должен угадывать класс как можно чаще?

Задача классификации

x_1, x_2, \dots, x_l - объекты, для которых известны их классы:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i = a(x_i)$$

Как выразить то, что он должен угадывать класс как можно чаще?

$$\sum_{i=1}^l [y_i \neq a(x_i)] \rightarrow \min$$

Задача классификации

x_1, x_2, \dots, x_l - объекты, для которых известны их классы (0 или 1):

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм

Как выразить то, что он должен угадывать вероятность класса 1?

Задача классификации

x_1, x_2, \dots, x_l - объекты, для которых известны их классы (0 или 1):

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм

Как выразить то, что он должен угадывать вероятность класса 1?

$$-\sum_{i=1}^l y_i \ln a(x_i) + (1 - y_i) \ln(1 - a(x_i)) \rightarrow \min$$

Задача классификации

x_1, x_2, \dots, x_l - объекты, для которых известны их классы (0 или 1):

$$y_1, y_2, \dots, y_l$$

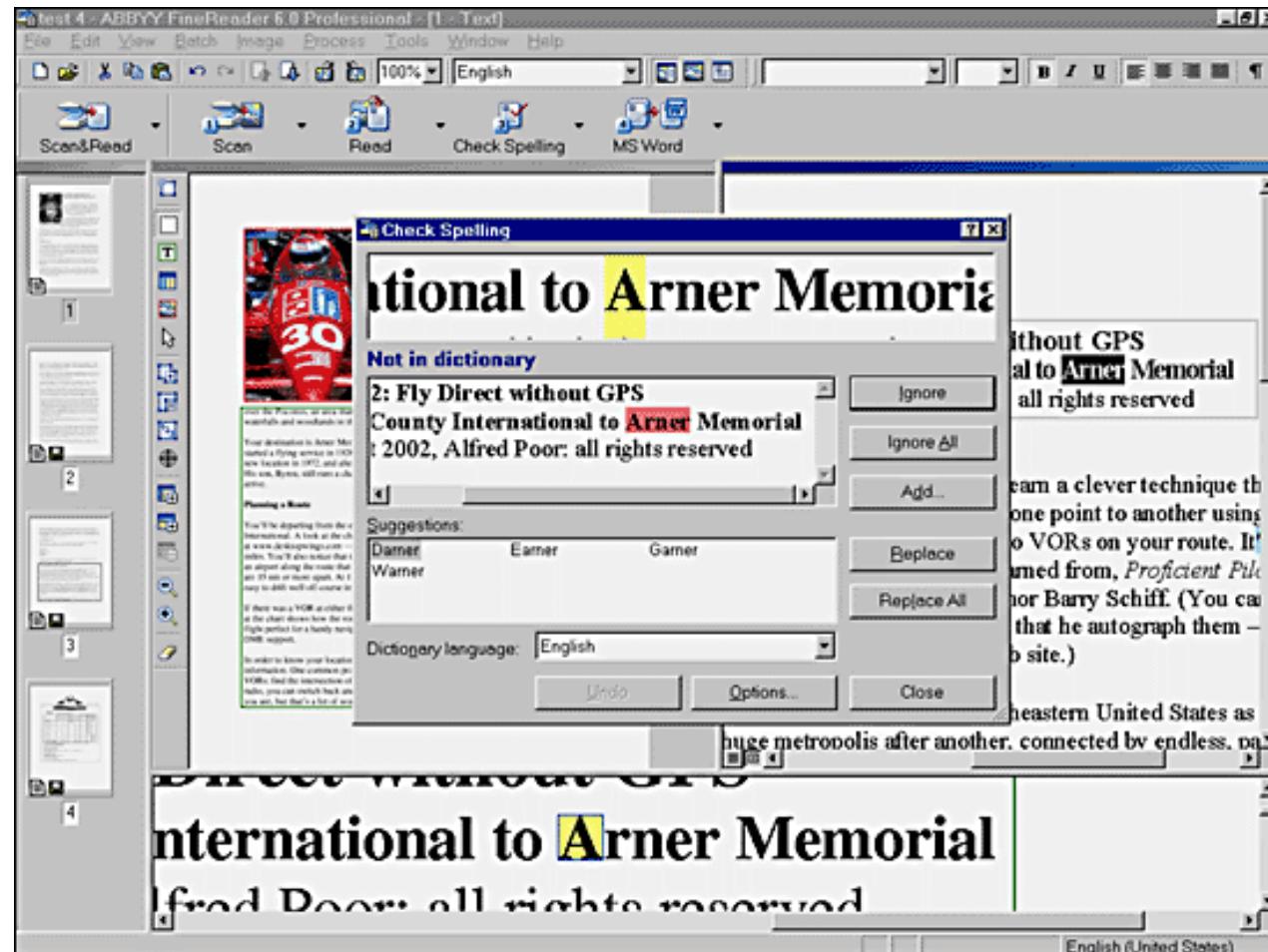
Мы строим прогнозирующий алгоритм

Как выразить то, что он должен угадывать вероятность класса 1?

$$\sum_{i=1}^l (y_i - a(x_i))^2 \rightarrow \min$$

Исправление опечаток

Исправление опечаток/ошибок распознавания



Исправление опечаток/ошибок распознавания

$$Suggest(w) = [w_1, w_2, \dots, w_k]$$

Исправление опечаток/ошибок распознавания

$$Suggest(w) = [w_1, w_2, \dots, w_k]$$

В алгоритме есть параметры, которые когда-то были заданы «вручную». Хочется настроить их так, чтобы *suggest* был как можно «адекватней».

Исправление опечаток/ошибок распознавания

$$Suggest(w) = [w_1, w_2, \dots, w_k]$$

В алгоритме есть параметры, которые когда-то были заданы «вручную». Хочется настроить их так, чтобы suggest был как можно «адекватней».

Есть выборка:

w (слово с опечаткой), cw(правильное написание)

Как сформулировать «адекватность» suggest'a,
как настроить параметры?

Сложный пример: исправление опечаток

Возможное решение:

$$\begin{aligned}Suggest(w) &= [w_1, w_2, \dots, w_k] \\ Pos(w_j, [w_1, w_2, \dots, w_k]) &= j\end{aligned}$$

$$\sum_{i=1}^l Pos(cw_i, Suggest(w_i)) \rightarrow \min$$

Оптимизация бюджета рекламных кампаний

Задача

К вам обращается крупное рекламное агентство. Они запускают рекламные кампании на ТВ и хотят лучше понимать, сколько рекламы на каждом телевизионном канале нужно заказать.

Задача

1. У каждой рекламной кампании есть своя целевая аудитория, и успешность мероприятия определяется по показателям в срезе ЦА.

Задача

1. У каждой рекламной кампании есть своя целевая аудитория, и успешность мероприятия определяется по показателям в срезе ЦА.
2. Для каждого среза считается величина «охват $k+$ » - сколько человек посмотрело рекламный ролик k и более раз и «точность» - сколько людей из данного среза посмотрели ролик

Задача

1. У каждой рекламной кампании есть своя целевая аудитория, и успешность мероприятия определяется по показателям в срезе ЦА.
2. Для каждого среза считается величина «охват $k+$ » - сколько человек посмотрело рекламный ролик k и более раз и «точность» - сколько людей из данного среза посмотрели ролик
3. Упрощение задачи: считаем, что на телеканале не покупается фиксированная реклама (когда мы точно знаем когда будет ролик показан), а только плавающие размещения (ролик будет показан в случайный момент суток)

Задача

1. У каждой рекламной кампании есть своя целевая аудитория, и успешность мероприятия определяется по показателям в срезе ЦА.
2. Для каждого среза считается величина «охват $k+$ » - сколько человек посмотрело рекламный ролик k и более раз и «точность» - сколько людей из данного среза посмотрели ролик
3. Упрощение задачи: считаем, что на телеканале не покупается фиксированная реклама (когда мы точно знаем когда будет ролик показан), а только плавающие размещения (ролик будет показан в случайный момент суток)
4. Нужно оптимизировать суммарный бюджет кампании, получив определённые значения охвата 5+ и точности ЦА (можно оптимизировать любой параметр при неуменьшении остальных, можно провести аналог со статистическими гипотезами)

Метрики качества

1. Количество сэкономленных средств
2. Какие значения охвата 5+ и точности были реально получены.

Проверка качества

1. Проверять экономию бессмысленно (это объективный параметр, мы реально тратим меньше). Надо проверять охват и точность.
2. АБ тест нельзя просто так сделать, кампания это очень долгий процесс (2-3 месяца)
3. Можно требовать, чтобы с вероятностью 95% охват и точность были не ниже, чем те, что мы предсказали.

Модель

- Есть 15 телеканалов, соответственно 15 фичей – сколько денег вложено в каждый канал
- Есть таргеты по каждой ЦА – сколько человек посмотрело хотя бы 1 раз и сколько хотя бы 5 раз
- Нужно обучить линейную модель и получить коэффициент для каждого канала – по сути отклик на вливание денег каждой аудиторией в каждый канал

Рекомендации товаров

Блок рекомендаций

Товар 1	Товар 2	Товар 3	Товар 4
---------	---------	---------	---------

Максимизация количества покупок

Товар 1	Товар 2	Товар 3	Товар 4

Вероятность:	p_1	p_2	p_3	p_4
--------------	-------	-------	-------	-------

Максимизация дохода

	Товар 1	Товар 2	Товар 3	Товар 4
Вероятность:	p_1	p_2	p_3	p_4
Цена:	c_1	c_2	c_3	c_4

Максимизация дохода



Puma
Ветровка
3 490 руб.

Crocs
Сланцы
1 990 руб.

Tony-p
Слипоны
~~1 999 руб.~~ 1 590 руб.

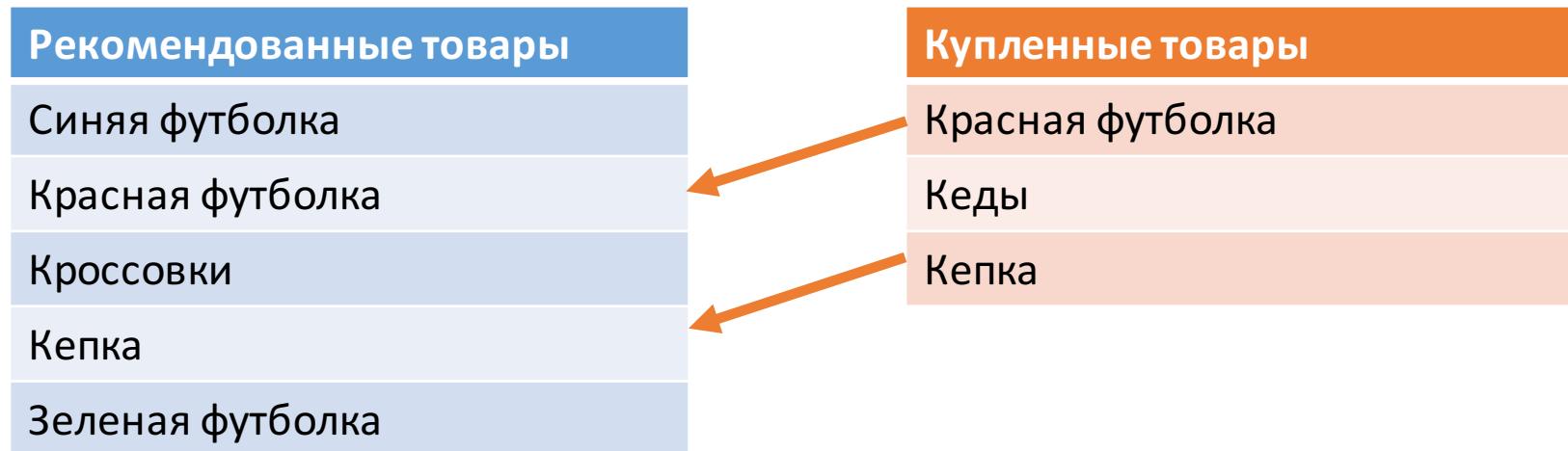
Champion
Брюки спортивные
~~3 599 руб.~~ 1 970 руб.

Вероятность:	0.05	0.02	0.015	0.009
Цена:	3490	1990	1590	1970

Прогнозирование вероятности

- Объекты: тройки (пользователь, товар, момент времени)
- Классы: 1 - товар будет куплен, 0 – товар не будет куплен
- Признаки: параметры пользователя, товара, момента времени и их «взаимодействие»

Точность (Precision@k)

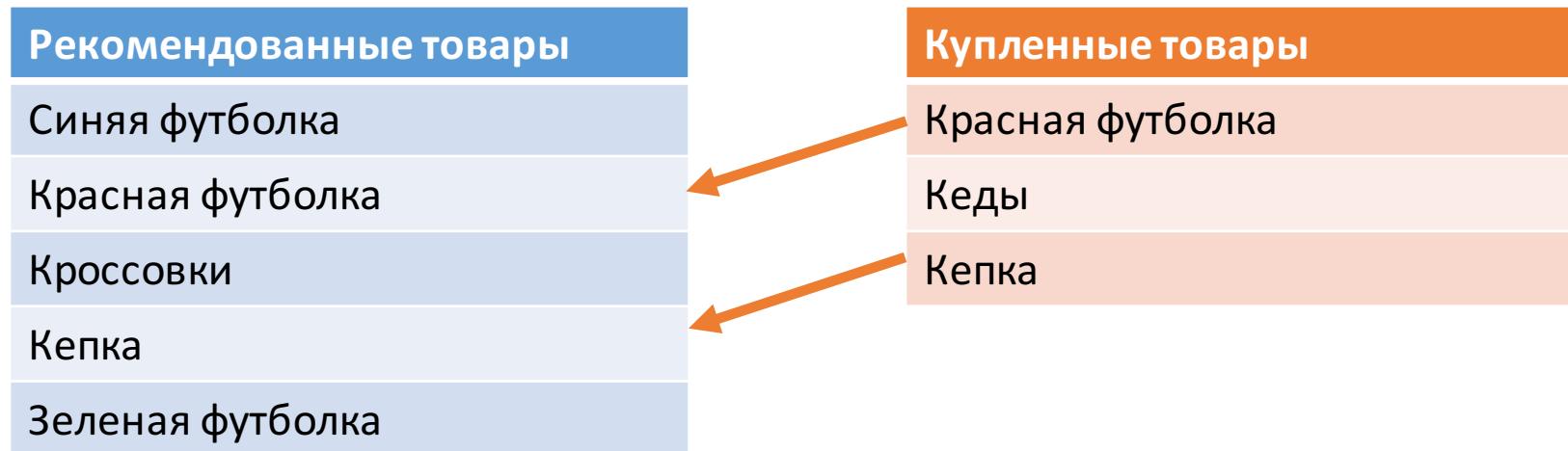


k – количество
рекомендаций

$$\text{Precision}@k = \frac{\text{купленное из рекомендованного}}{k}$$

AveragePrecision@k - усредненный по сессиям Precision@k

Полнота (Recall@k)



k – количество
рекомендаций

$$\text{Recall}@k = \frac{\text{купленное из рекомендованного}}{\text{количество покупок}}$$

AverageRecall@k - усредненный по сессиям Recall@k

Взвешенный ценами recall@k

Рекомендованные товары	Купленные товары
Синяя футболка – 1000р	Красная футболка – 1200р
Красная футболка – 1200р	Кеды – 3000р
Кроссовки – 3500р	Кепка – 900р
Кепка – 900р	
Зеленая футболка – 800р	

Взвешенный ценами Recall@k = $\frac{\text{стоимость купленного из рекомендованного}}{\text{стоимость покупок}}$

AverageRecall@k - усредненный по сессиям Recall@k

Качество классификации против качества рекомендаций

Пример – 2 решения для прогноза купит/не купит товар

	Алгоритм 1	Алгоритм 2
AUC классификатора	0.52	0.85
Recall@5	0.72	0.71

Онлайновая оценка качества

Допустим, на исторических данных качество алгоритма высокое, а будет ли оно высоким в реальности?

Идеи:

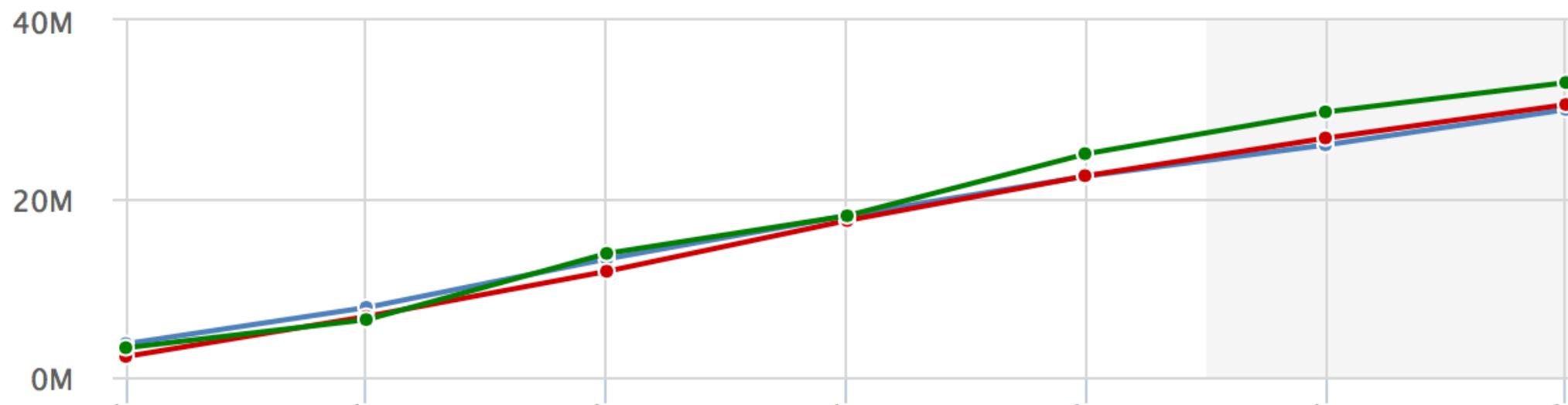
1. А/В тест
2. Оценка статзначимости результата

A/B тест

1. Случайным образом делим пользователей на равные группы
2. Измеряем целевые метрики (например, количество заказов или доход) в каждой группе за длительный период времени
3. Получаем какое-то число для каждой группы
4. Что дальше?

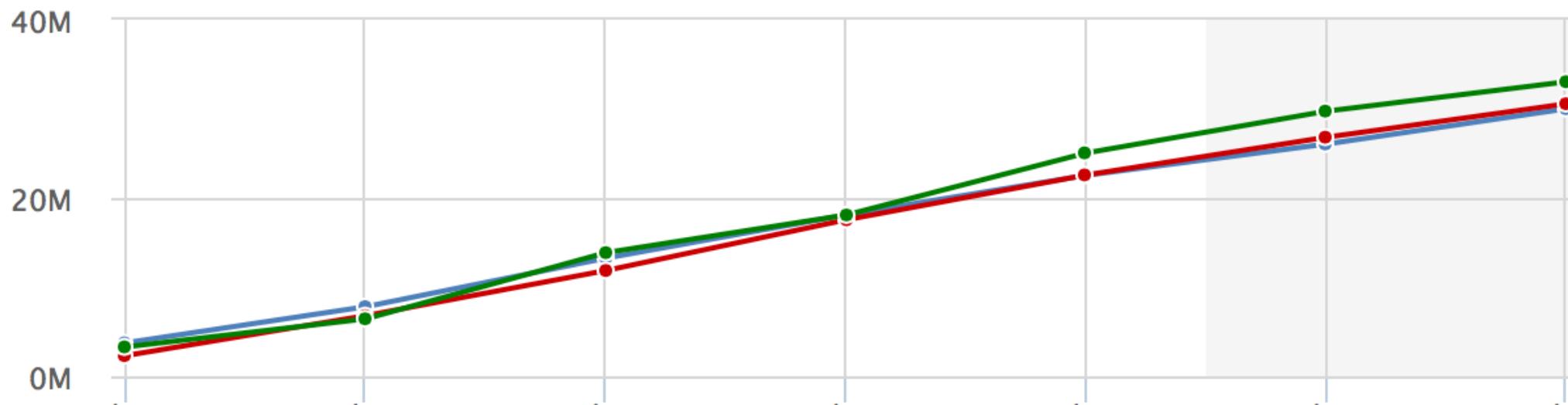
Статистическая значимость: пример

Суммарная выручка



Статистическая значимость: пример

Суммарная выручка



Одна кривая отличается от других на 10%
Но разбиение на самом деле – случайное

На какие метрики смотрят в онлайне

- Доход в группе
- Доход с пользовательской сессии
- Средняя стоимость купленного товара
- Средний чек
- Конверсия в покупку
- Клики
- Различные модели атрибуции: last click, first click

Мини-задача

Как изменится построение модели, если нам нужно максимизировать количество просмотренных пользователем товаров?

Мини-задача

Как изменится построение модели, если нам нужно максимизировать количество просмотренных пользователем товаров?

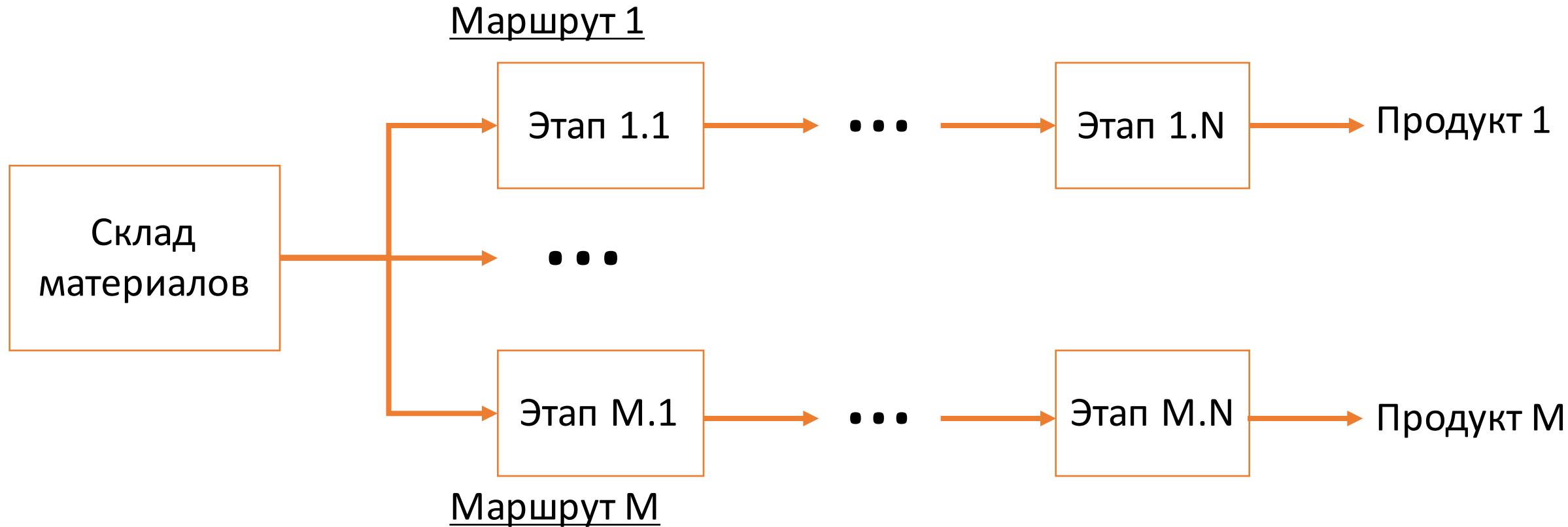
А если нужно максимизировать количество проданных товаров из категории «аксессуары»?

Итог: о чём нужно позаботиться

- Высокоуровневая постановка задачи - от экономического эффекта
- Оценка возможного экономического эффекта
- Оценка реализуемости проекта
- Оффлайновая оценка качества
- Онлайновая оценка качества
- Решение задачи – декомпозиция на подзадачи, выбор признаков, выбор моделей

Прогнозирование дефектов на производстве

Маршруты производства



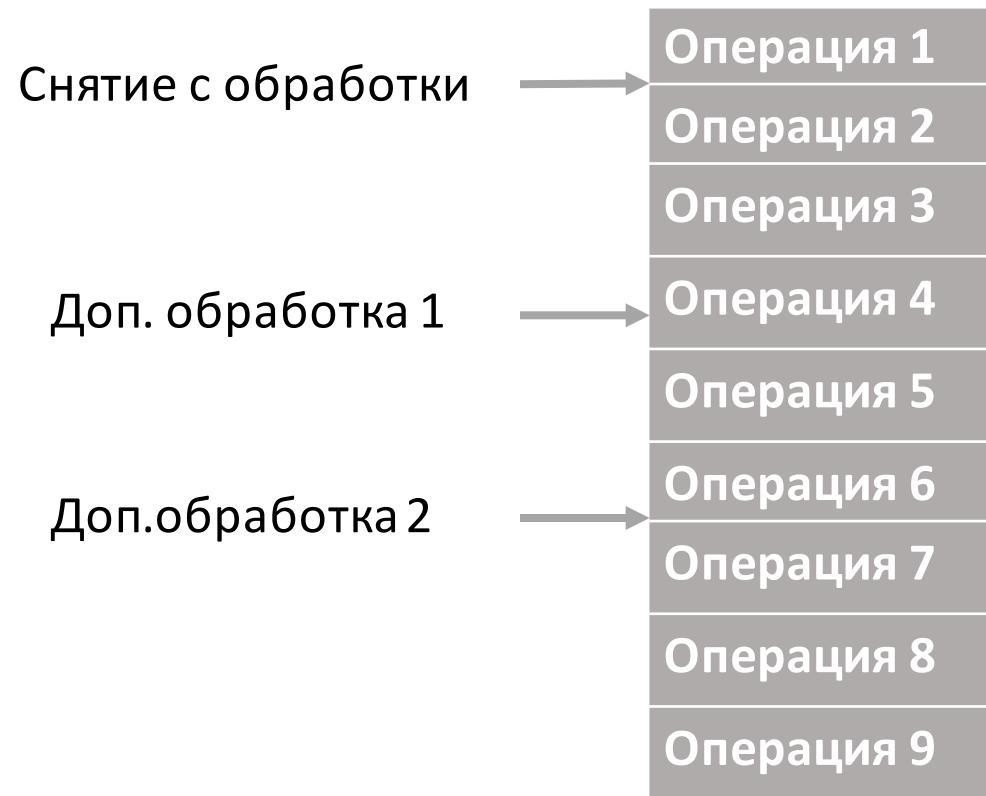
Дополнительная обработка

Между некоторыми обязательными этапами обработки часть продукции может отправляться на доработку (чтобы уменьшить вероятность брака)

Опишем процесс

Продукт 8								
Продукт 7								
Продукт 6								
Продукт 5								
Продукт 4								
Продукт 3								
Продукт 2								
Продукт 1								
Операция 1	+	+	+	+	+	+	+	+
Операция 2	+	+	+	+	+	+	+	+
Операция 3	+	+			+	+	+	+
Операция 4	+	+	+	+	+	+	+	+
Операция 5	+	+			+	+	+	+
Операция 6				+				
Операция 7					+			
Операция 8						+	+	
Операция 9						+	+	+

Место дополнительной обработки в процессе



Где здесь машинное обучение

- Учимся на этапах перед возможной доп.обработкой прогнозировать дефект в продукции
- Если дефект – бинарная величина, прогнозируем его вероятность (решаем задачу классификации), если вещественная (например, масса непригодного продукта), то прогнозируем матожидание (задача регрессии)
- Отправляем на доп.обработку продукцию, которой она больше всего нужна

Экономика процесса

- Количество потраченных денег = потери из-за производства непригодного продукта (брата) + потери на доп. обработки и снятие с производства
- Доп. обработки уменьшают первое слагаемое и увеличивают второе
- Потери из-за брака = сумма потерь на каждом произведенном продукте
- Значит, зная матожидание потерь на каждом продукте – понимаем, какие

Что еще можно обсудить

- Что включать в потери из-за снятия с производства
- Оценка того, насколько доп.обработка поможет устранить дефект
- Смещенность выборок

Прогнозирование оттока

Слайды взяты из презентации Эмели Драль для осеннего
семестра курса Data Mining in Action

Что такое “отток”?

- Отказ пользователя от некоторого продукта или услуги



Зачем прогнозировать отток?

- Больше пользователей -> больше прибыли

Откуда взять больше пользователей?

- Больше пользователей = больше новых + меньше отток существующих

Экономическая эффективность

- Удержать одного vs. привлечь
одного пользователя?

От оттока к удержанию

- Удерживать всех пользователей **дорого** -> адресное удержание
- Удержание пользователей происходит **не мгновенно**
-> прогноз с солидным горизонтом



Метрики

- Return rate
- Churn rate
- X-day retention
- Rolling retention

Возвращаемость и отток

Return rate

- RR = (current number of customers from the original set) / (number of customers at the original set) * 100

Churn rate

- CR = (number of churned customers)/(total number of customers)*100

Возвращаемость

- 1-day retention
- 7-day retention
- 28-day retention

Возвращаемость

- 1-day retention
 - 7-day retention
 - 28-day retention
- 1-day rolling retention
 - 7-day rolling retention
 - 28-day rolling retention

Анализ аудитории

- Актуальна ли проблема оттока?
- Как много пользователей оттекает?
- Сколько на этом теряется денег?

Оценка экономического потенциала удержания

- Построение экономической модели
- Сколько мы заработаем в случае реализации процесса удержания с заданным качеством?

Формализация постановки задачи

- Формальное определение оттока
- Тип модели
- Горизонт прогнозирования
- Методика оценки качества модели
- Дизайн эксперимента
- Требования к модели

Определение оттока

- Разрыв договора
- Отсутствие платных транзакций более 10/90 дней
- Отсутствие на сервисе более 14/28 дней

Горизонт прогнозирования

- Как быстро мы можем связаться с пользователем?
- Какие методики удержания мы используем?
- Сколько времени занимает процесс удержания?

Данные

- Какие данные доступны?
- За какой исторический период?
- Как объединять данные ?
- Есть ли в данных сигнал?
- Как данные следует обработать?
- Как рассчитать признаки на основе данных?

Описательная аналитика

- Ключевые характеристики пользователей?
- Различаются ли пользователи из групп отток/не отток?
- Можем ли мы решать задачу для всех сегментов?
- Однаково ли важно решать задачу для всех сегментов?

Кампания по удержанию

Разработка дизайна кампании:

- Каналы коммуникаций с пользователями?
- Время взаимодействия?
- Предложения по удержанию?

Запуск кампаний и оценка результатов

- Формирование ожиданий от результатов кампании (в том числе финансовых)
- Оценка эффективности кампании с помощью А/Б-тестирования
- Анализ результатов, планирование новых кампаний

Возможная постановка задачи

- Отток - разрыв договора подключения к сервису
- Модель - бинарная классификации
- Горизонт прогнозирования – 2 недели
- Методика оценки – метрика AUC
- Дизайн эксперимента – А/Б тестирование на 10% сегменте случайных пользователей
- Требования к модели: вероятностная модель

Вернемся к экономике

- Пусть мы удерживаем N пользователей, наиболее вероятно уходящих в отток по прогнозу нашей модели
- Тратим на удержание каждого C денег
- r – доля настоящих отточников среди удерживаемых
- N_p – удержаных пользователей, если удерживаем со 100% успехом
- N_{pr} – удержаных пользователей, если удержание успешно с вероятностью r

Вернемся к экономике

- Пусть мы удерживаем N пользователей, наиболее вероятно уходящих в отток по прогнозу нашей модели
- Тратим на удержание каждого C денег
- r – доля настоящих отточников среди удерживаемых
- Npr – удержаных пользователей, если удержание успешно с вероятностью r

Вывод:

Наша задача – максимизировать r

Экономический эффект: $ARPU * N * p * r - C * N$

Другая постановка задачи

- Отток - разрыв договора подключения к сервису
- Модель - бинарная классификации
- Горизонт прогнозирования – 2 недели
- Методика оценки – Precision@k или Lift-N% (например, 10%)
- Дизайн эксперимента – А/Б тестирование на 10% сегменте случайных пользователей
- Требования к модели: вероятностная модель

Мини-задача

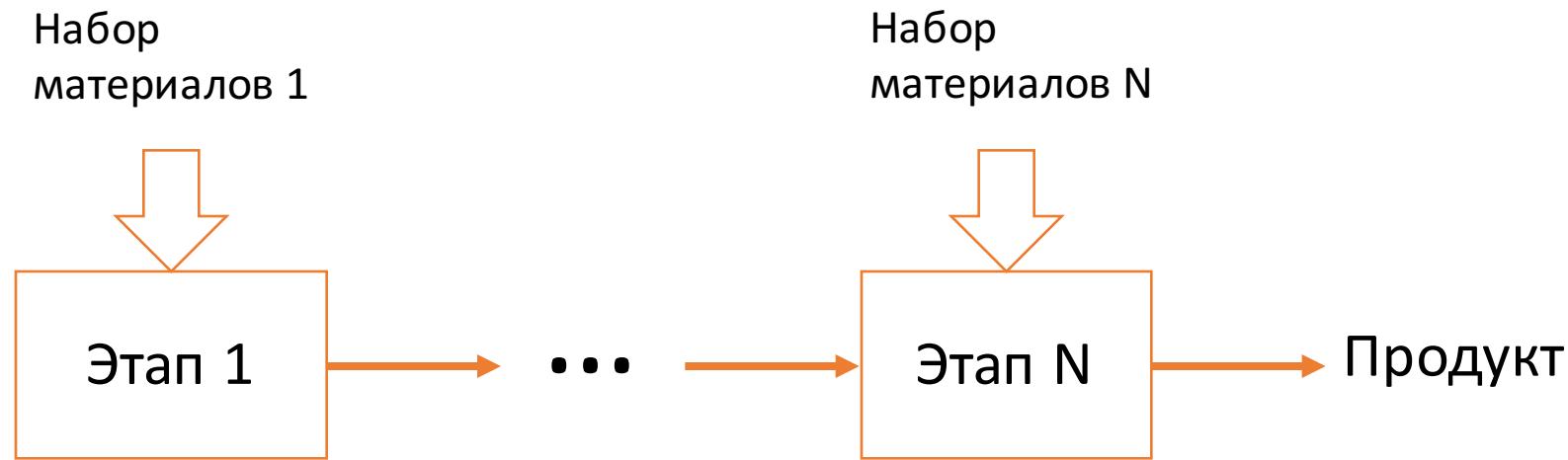
Как все поменяется, если мы не считаем пользователей одинаково ценными и имеем оценку выручки с каждого из пользователей?

Важно

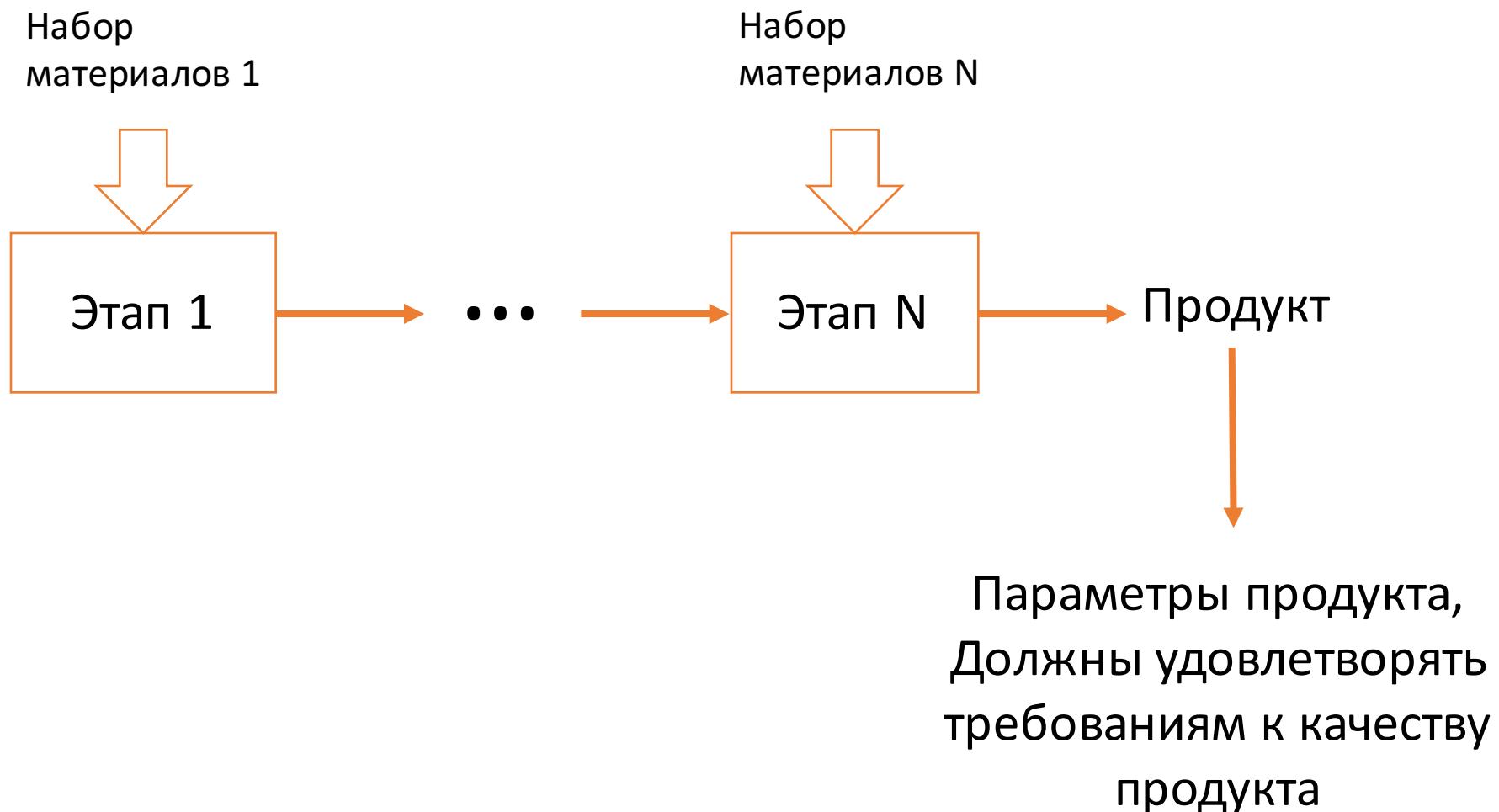
- Убедиться в обоснованности задачи с точки зрения бизнеса
- Формализовать и провалидировать постановку задачи
- Оценивать качество решения задачи на всех этапах (от постановки до результатов А/Б тестирования)

Оптимизация затрат материалов на производстве

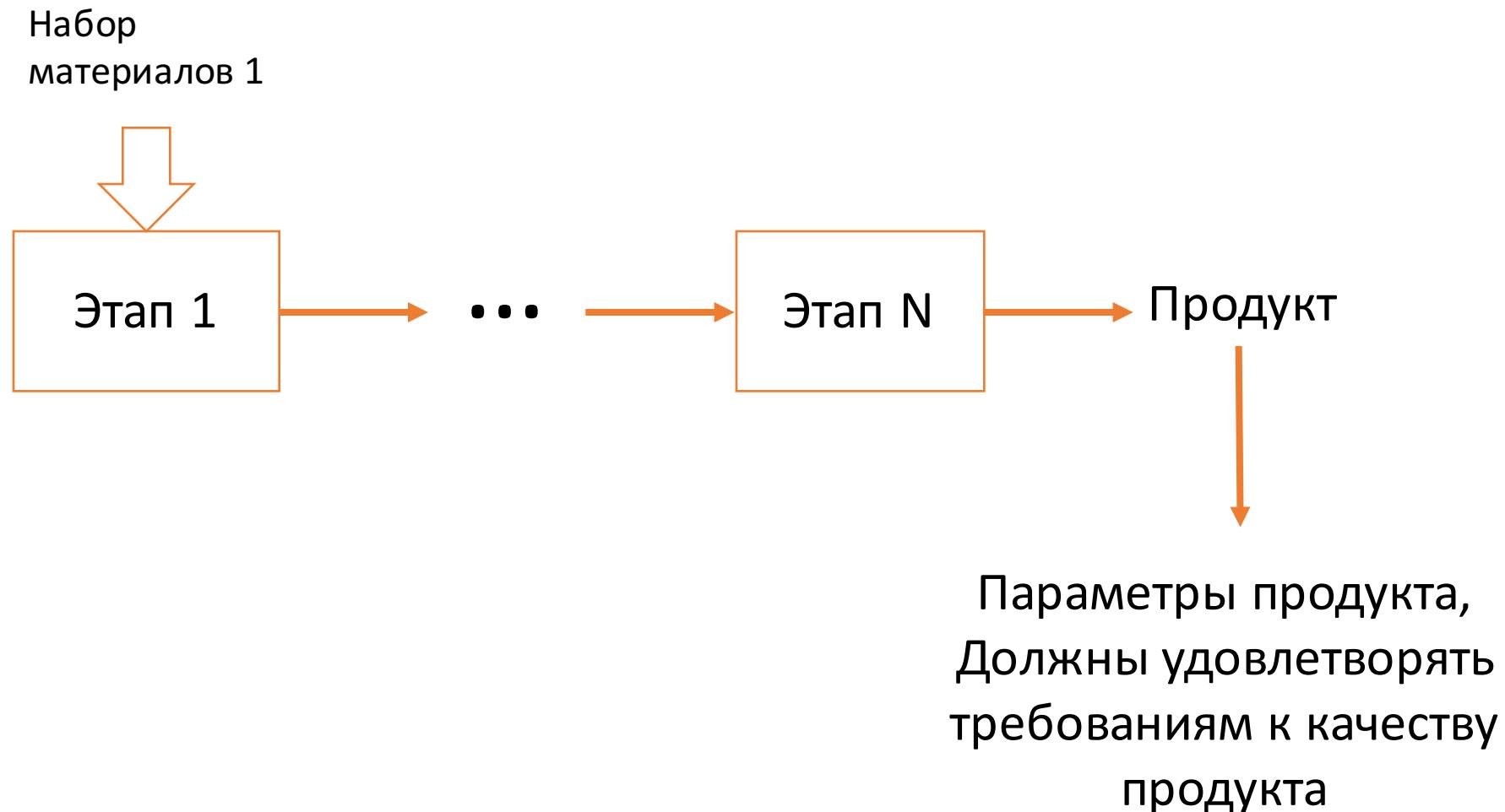
Процесс производства



Процесс производства



Процесс производства



Где здесь машинное обучение

- Учимся предсказывать параметры продукта по количеству израсходованных на него материалов

Где здесь машинное обучение

- Учимся предсказывать параметры продукта по количеству израсходованных на него материалов
- Решаем задачу минимизации затрат при условии попадания спрогнозированных параметров продукта в критерии по качеству

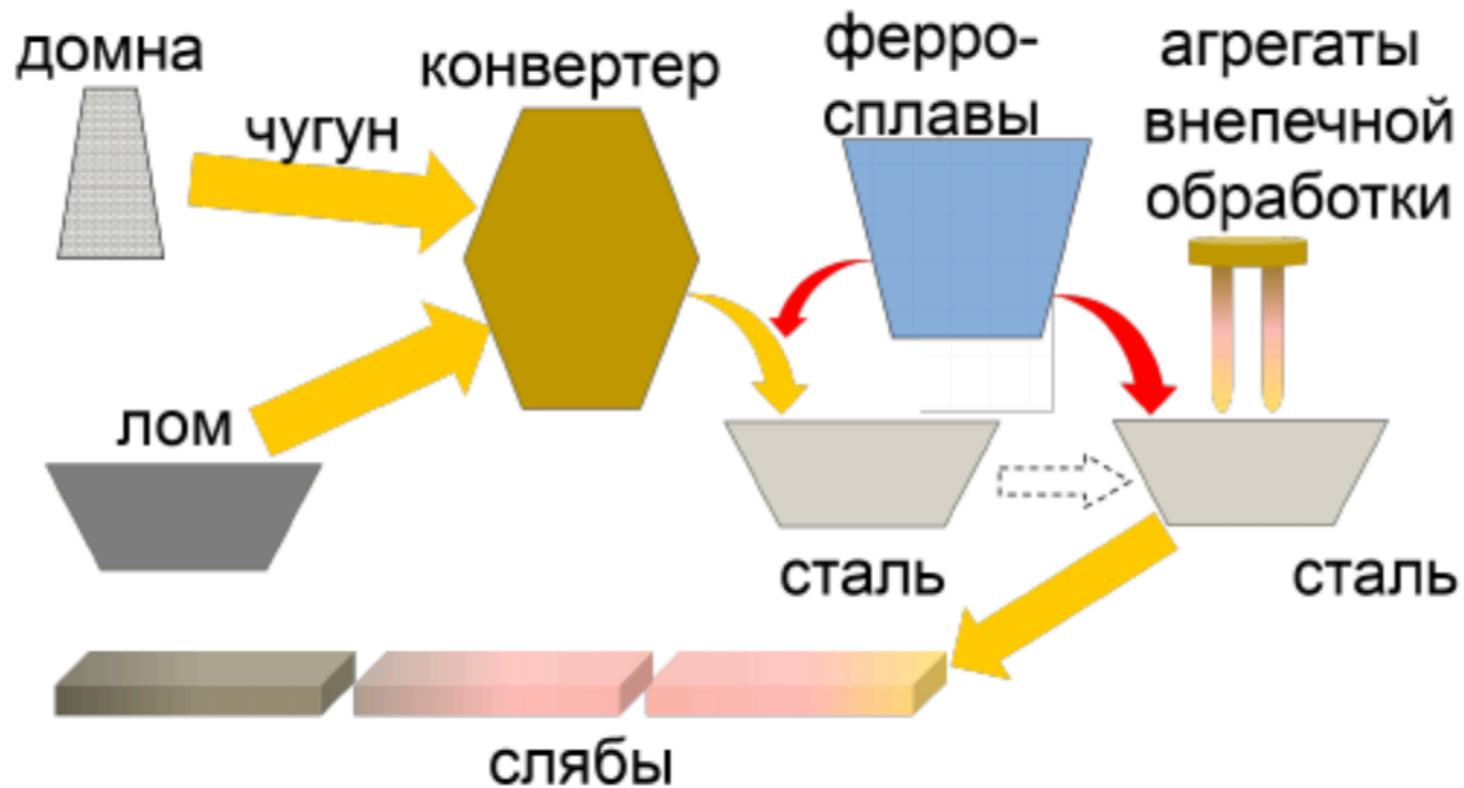
Пример задачи

Оптимизация затрат на ферросплавы при выплавке стали - кейс из практики Yandex Data Factory

Рассказ о задаче:

<https://events.yandex.ru/lib/talks/3996/>

Процесс выплавки стали



Оптимизация. Пример ограничен



- › Ограничения – условия уверенного попадания содержания элементов в заданные диапазоны по химии:

$$\int_{\alpha_i}^{\beta_i} \rho(y_i) dy_i \geq C$$

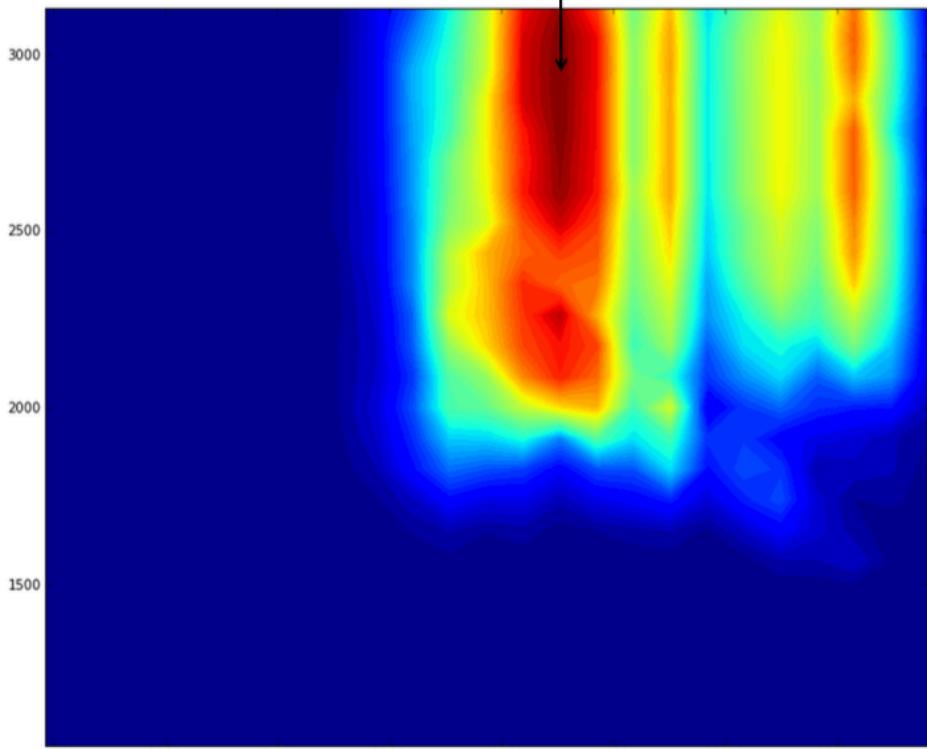
- › y_i – хим.состав; α_i, β_i – границы диапазона;
- › $\rho(y_i)$ – распределение плотности вероятности получения хим.состава y_i ;
- › C – порог уверенного попадания.

Оптимизация. Иллюстрация ограничений

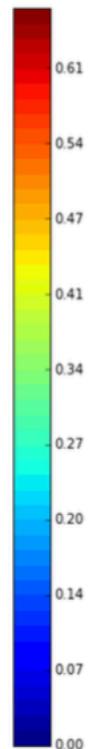


Область, удовлетворяющая ограничениям

Масса ФС65, кг



Масса СМН18, кг



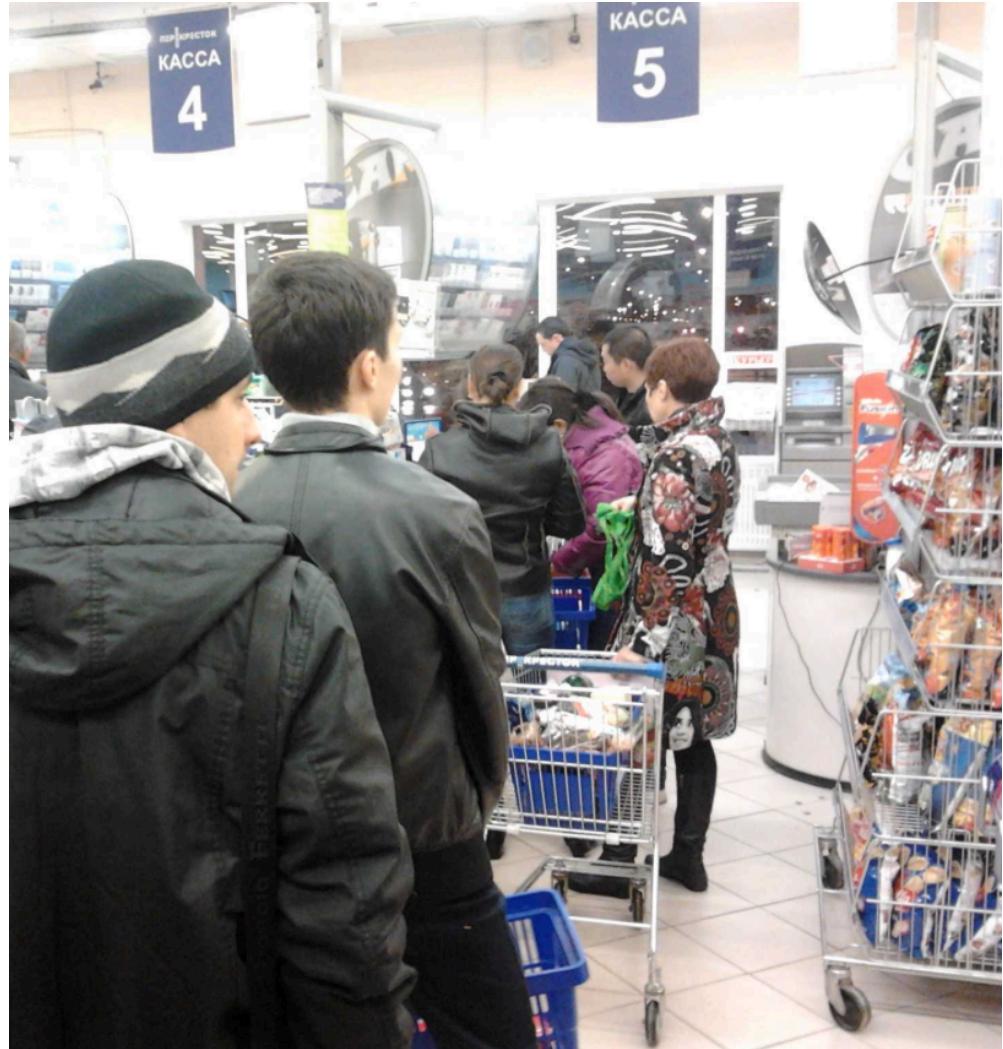
← Порог уверенного попадания
содержания элементов в
заданные

Задача про очереди

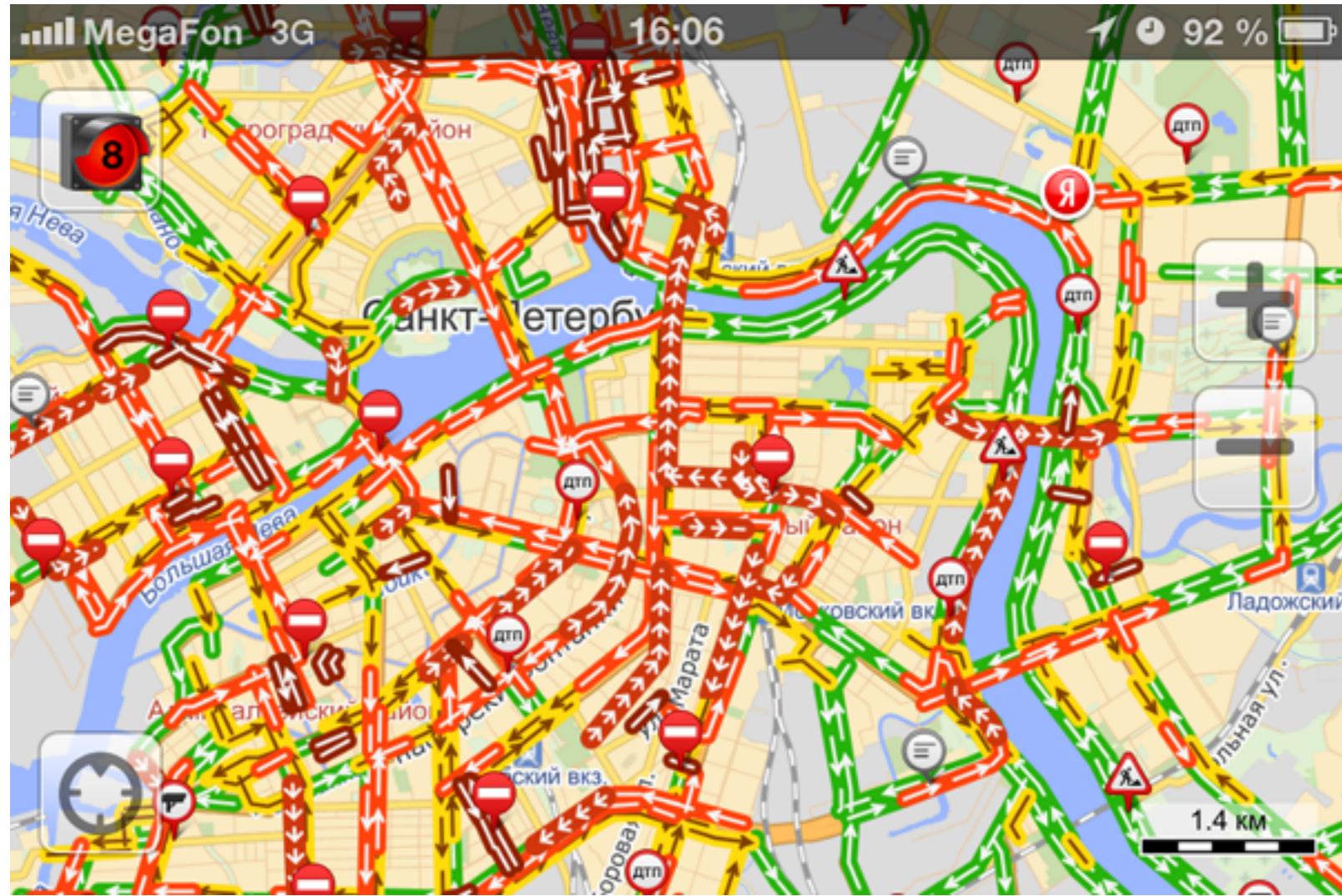
Задача про очереди

К исследователю обращается сеть магазинов с формулировкой: «Нам кажется, у нас проблемы. И нам кажется, что это очереди».

Придумайте, что может сделать исследователь: какие задачи можно решить, какие данные понадобятся.



Еще примеры для обсуждения



VK Data Mining in Action | ВКонтакте[vk.com > data_mining_in_action](#) ▾

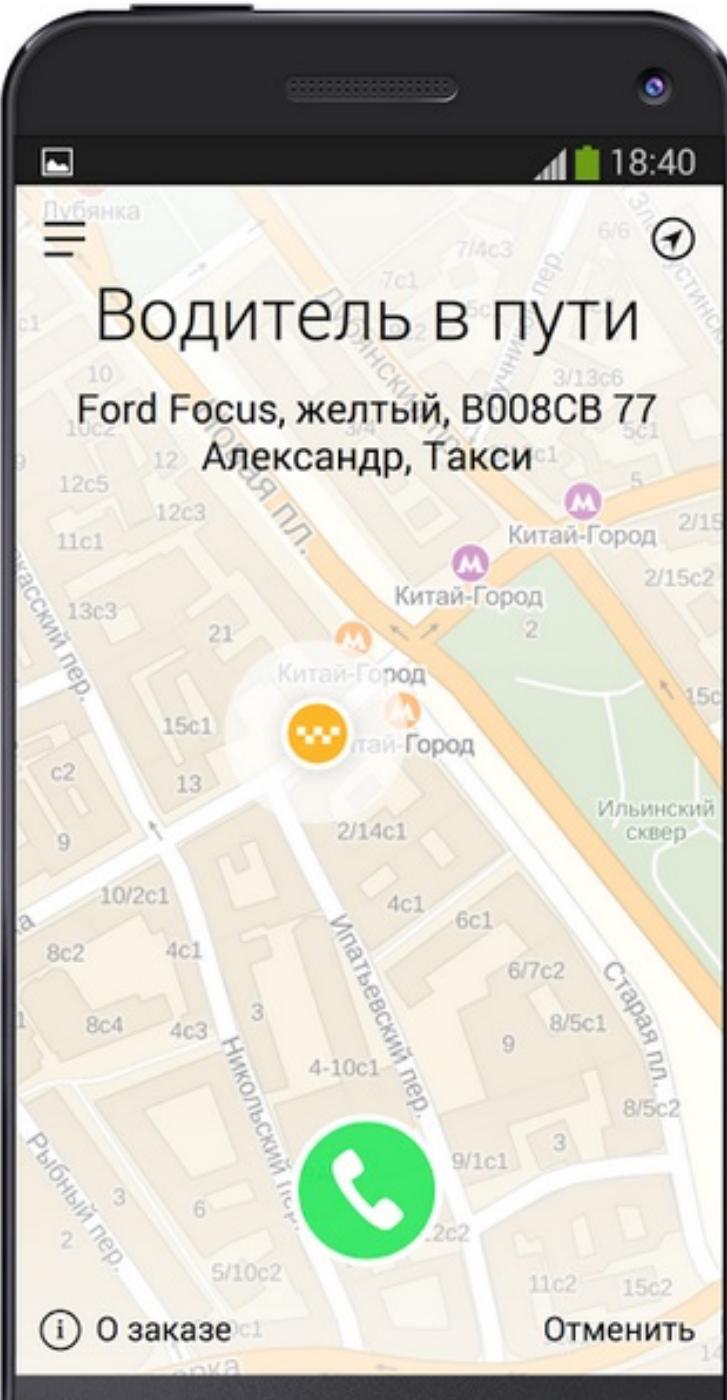
Москва, Россия Денис Семененко. Администратор сообщества. Data Mining in Action. So it begins. Местоположение: Москва, Россия. . Data Mining in Action запись закреплена. 6 мая в 23:04.

Нашлось 8 млн результатов[Добавить объявление](#) [Показать все](#)**H Process Mining: знакомство / Хабрахабр**[habrahabr.ru > post/244879](#) ▾

Статья подготовлена на основе материалов онлайн курса **Process Mining: Data Science in Action**, являющихся собственностью Технического университета Эйндховена.

Coursera Process Mining: Data science in Action... | Coursera[coursera.org > learn/process-mining](#) ▾





Так что же такое «постановка задачи»?

Ответ на вопросы:

1. Как в реальной задаче из бизнеса возникает ваша задача машинного обучения: что нужно прогнозировать?

Так что же такое «постановка задачи»?

Ответ на вопросы:

1. Как в реальной задаче из бизнеса возникает ваша задача машинного обучения: что нужно прогнозировать?
2. Как будут использоваться ваши прогнозы?

Так что же такое «постановка задачи»?

Ответ на вопросы:

1. Как в реальной задаче из бизнеса возникает ваша задача машинного обучения: что нужно прогнозировать?
2. Как будут использоваться ваши прогнозы?
3. Что нужно оптимизировать бизнесу? (выручку, количество пользователей, расходы на производство и т.д.)

Так что же такое «постановка задачи»?

Ответ на вопросы:

1. Как в реальной задаче из бизнеса возникает ваша задача машинного обучения: что нужно прогнозировать?
2. Как будут использоваться ваши прогнозы?
3. Что нужно оптимизировать бизнесу? (выручку, количество пользователей, расходы на производство и т.д.)
4. Исходя из предыдущих пунктов, что за задачу машинного обучения следует решать, что будет объектами, что будет таргетом, какие данные уместны для построения признаков, какая метрика качества должна оптимизироваться?

Так что же такое «постановка задачи»?

Ответ на вопросы:

1. Как в реальной задаче из бизнеса возникает ваша задача машинного обучения: что нужно прогнозировать?
2. Как будут использоваться ваши прогнозы?
3. Что нужно оптимизировать бизнесу? (выручку, количество пользователей, расходы на производство и т.д.)
4. Исходя из предыдущих пунктов, что за задачу машинного обучения следует решать, что будет объектами, что будет таргетом, какие данные уместны для построения признаков, какая метрика качества должна оптимизироваться?
5. Как оценивать качество работы вашей модели в процессе использования?

Спасибо за внимание!