

Машинное обучение и майнинг данных

Лекция 1

Введение в машинное обучение и разработку данных

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2017

Как перевести часы в минуты?



Как перевести часы в минуты?

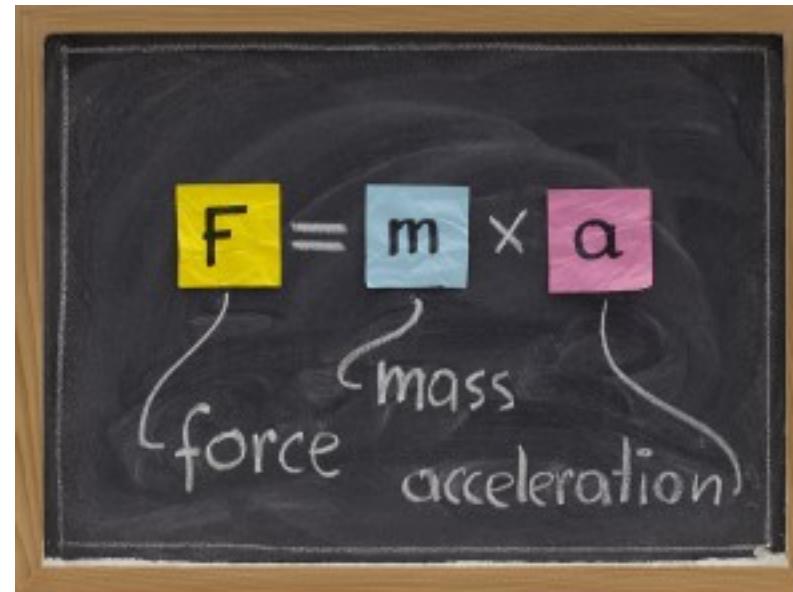
- x — часы
- $f(x) = 60x$ — преобразование в минуты, функция

Какая сила приложена к телу?

- Известны масса тела m и его ускорение a
- Чему равна сила F ?

Какая сила приложена к телу?

- Известны масса тела m и его ускорение a
- Чему равна сила F ?
- Второй закон Ньютона: $F = ma$



Как предсказать погоду?



Уравнения Навье-Стокса

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} = - \frac{\partial P}{\partial x} + Re \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right),$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} = - \frac{\partial P}{\partial y} + Re \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2} \right),$$

$$\frac{\partial w}{\partial t} + u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} + w \frac{\partial w}{\partial z} = - \frac{\partial P}{\partial z} + Re \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2} \right),$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0.$$

Уравнения Навье-Стокса

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + w \frac{\partial u}{\partial z} = - \frac{\sigma_x}{\rho} + Re \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right),$$

$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} = - \frac{\sigma_y}{\rho} + Re \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2} \right)$

$\frac{\partial w}{\partial t} + u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} + w \frac{\partial w}{\partial z} = - \frac{\sigma_z}{\rho} + Re \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2} \right)$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0.$$

Анализ тональности текста

- Какой эмоциональный окрас имеет текст?
- Варианты: позитивный, нейтральный, негативный
- Применение: автоматический анализ отзывов от пользователей

Анализ тональности текста

«Большое спасибо! Сюда по всему, это как раз то, чего не хватает всем зарубежным курсам по Machine Learning и Knowledge Discovery. Это теория, математика, объяснение того, как оно устроено “в кишках”.»

Какой окрас?

Анализ тональности текста

«Я вижу очень большой минус, что курс будет на готовой библиотеке sci-kit. Курс от Andrew лучше тем, что ученик сам пишет алгоритм и видит изнутри, как он работает.»

Какой окрас?

Анализ тональности текста

- x — текст на русском языке
 - $f(x)$ — его окрас (принимает значения -1, 0, 1)
 - Можно ли выписать формулу для $f(x)$?
-
- На входе — вовсе не числа
 - Точная зависимость может не существовать

Больше сложных задач!

- Какой будет спрос на товар в следующем месяце?
- Сколько денег заработает магазин за год?
- Вернет ли клиент кредит?
- Заболеет ли пациент раком?
- Сдаст ли студент следующую сессию?
- На фотографии гуманитарий или технарь?
- Кто выиграет битву в онлайн-игре?

Больше сложных задач!

- Везде — очень сложные неявные зависимости
- Нельзя выразить их формулой
- Но есть некоторое число примеров
 - Тексты с известным окрасом
- Будем приближать зависимости, используя примеры

Анализ данных и машинное обучение

— это про то, как восстановить сложные зависимости
по конечному числу примеров

Основные термины

Пример задачи

- Сеть ресторанов
- Хотим открыть еще один
- Несколько вариантов размещения
- Какой из вариантов принесет максимальную прибыль?

* см. [kaggle.com, TFI Restaurant Revenue Prediction](https://www.kaggle.com/tmdb/tmdb-movie-metadata)

Обозначения

- x — объект, sample — для чего хотим делать предсказания
 - Конкретное расположение ресторана
- \mathbb{X} — пространство всех возможных объектов
 - Все возможные расположения ресторанов
- y — ответ, целевая переменная, target — что предсказываем
 - Прибыль в течение первого года работы
- \mathbb{Y} — пространство ответов — все возможные значения ответа
 - Все вещественные числа

Обучающая выборка

- Мы ничего не понимаем в экономике
- Зато имеем много объектов с известными ответами
- $X = (x_i, y_i)_{i=1}^{\ell}$ — обучающая выборка
- ℓ — размер выборки

Признаки

- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x^1, \dots, x^d)$ — признаковое описание

Признаки

- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x^1, \dots, x^d)$ — признаковое описание



Вектор

Признаки

- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x^1, \dots, x^d)$ — признаковое описание



Признаки

- Про демографию:
 - Средний возраст жителей ближайших кварталов
 - Динамика количества жителей
- Про недвижимость:
 - Средняя стоимость квадратного метра жилья поблизости
 - Количество школ, банков, магазинов, заправок
 - Расстояние до ближайшего конкурента
- Про дороги:
 - Среднее количество машин, проезжающих мимо за день

Алгоритм

- $a(x)$ — алгоритм, модель — функция, предсказывающая ответ для любого объекта
- Отображает \mathbb{X} в \mathbb{Y}
- Линейная модель: $a(x) = w_1x^1 + \dots + w_dx^d$

ФУНКЦИЯ ПОТЕРЬ

- Не все алгоритмы полезны
- $a(x) = 0$ — не принесет никакой выгоды
- Функция потерь — мера корректности ответа алгоритма
- Предсказали \$10000 прибыли, на самом деле \$5000 — хорошо или плохо?
- Квадратичное отклонение: $(a(x) - y)^2$

Функционал качества

- Функционал качества, метрика качества — мера качества работы алгоритма на выборке
- Среднеквадратичная ошибка (Mean Squared Error, MSE):

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

- Чем меньше, тем лучше

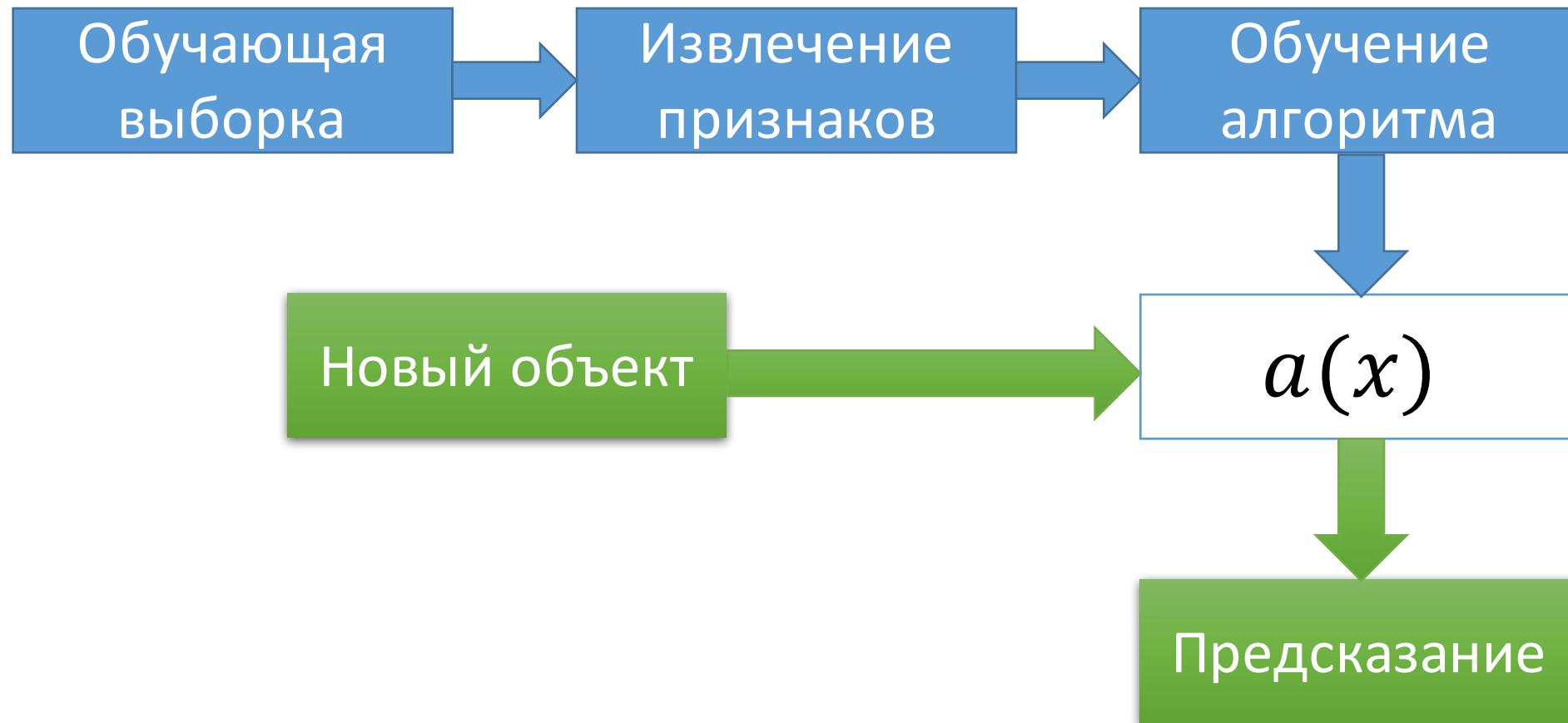
Функционал качества

- Должен соответствовать бизнес-требованиям
- Одна из самых важных составляющих анализа данных

Обучение алгоритма

- Есть обучающая выборка и функционал качества
- Семейство алгоритмов \mathcal{A}
 - Из чего выбираем алгоритм
 - Пример: все линейные модели
 - $\mathcal{A} = \{w_1x^1 + \dots + w_dx^d \mid w_1, \dots, w_d \in \mathbb{R}\}$
- Обучение: поиск оптимального алгоритма с точки зрения функционала качества

Машинное обучение

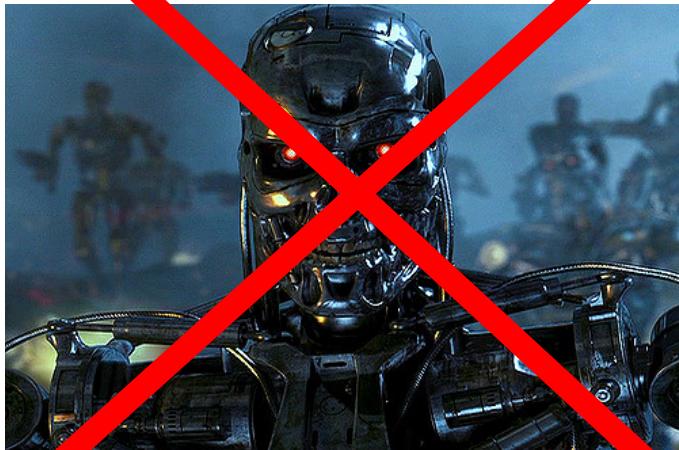


Что нужно знать

1. Как сформулировать задачу?
2. Как выделить признаки?
3. Как сформировать обучающую выборку?
4. Как выбрать метрику качества?
5. Как подготовить данные к обучению?
6. Как обучить алгоритм?
7. Как оценить качество алгоритма?

Зачем это нужно?

Искусственный интеллект



Сильный ИИ

через 20-100 лет

Яндекс

фильм где астронавту протыкают скафандр



Найти

ПОИСК КАРТИНКИ ВИДЕО КАРТЫ МАРКЕТ НОВОСТИ ПЕРЕВОДЧИК ЕЩЁ



Марсианин

The Martian, 2015 (16+)

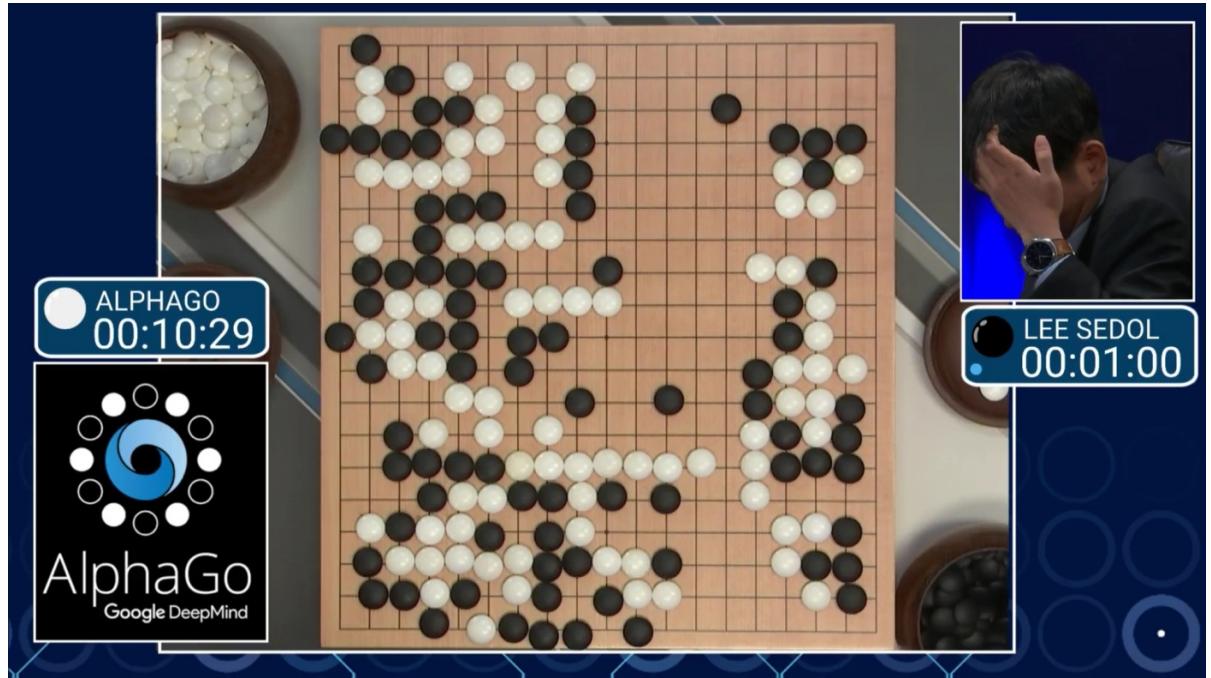
Марсианская миссия «Арес-3» в процессе работы была вынуждена экстренно покинуть планету из-за надвигающейся песчаной бури. Инженер и биолог Марк Уотни получил повреждение скафандра во время песчаной бури. Сотрудники миссии, посчитав его погибшим,...
[Читать дальше](#)

Специализированный ИИ

уже сейчас

AlphaGo

- Модель для игры в Го
- Оценивает успешность хода
- Обучалась путём игры с собой
- Победила чемпиона мира в 2016 году
- Долгое время игра в Го считалась невозможной задачей для компьютера



Аннотирование изображений



"man in black shirt is playing guitar."



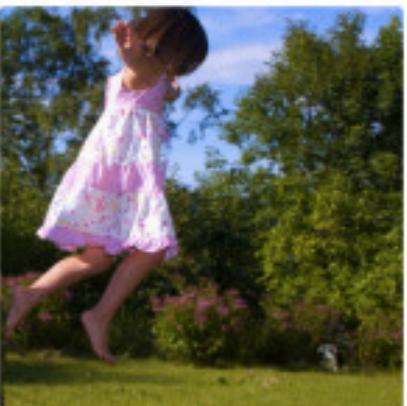
"construction worker in orange safety vest is working on road."



"two young girls are playing with legos toy."



"boy is doing backflip on wakeboard."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."

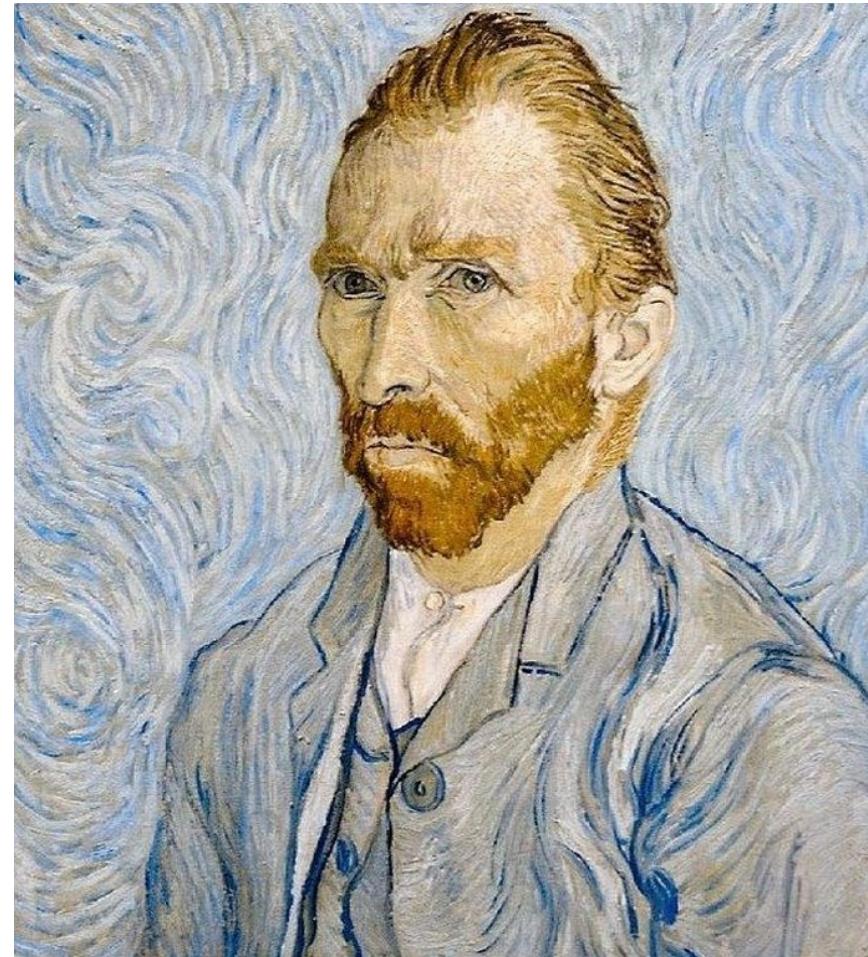


"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."

Перенос стиля



Выплавка стали

- Модель предсказывает, получится ли требуемый химический состав в результате плавки
- Сокращает расход ферросплавов на 5%
- Экономия до 23 млн. руб. в месяц
- Совместный проект Яндекса и Магнитогорского металлургического комбината



Рекомендательные системы

- Полки рекомендаций на Amazon генерируют 35% от всех покупок
- Рекомендации на основе машинного обучения и анализа больших объёмов данных

Frequently Bought Together

Price For All Three: \$86.01

Add all three to Cart Add all three to Wish List

Show availability and shipping details

This item: Machine Learning for Hackers by Drew Conway Paperback \$33.87

Machine Learning in Action by Peter Harrington Paperback \$25.75

Programming Collective Intelligence: Building Smart Web 2.0 Applications by Toby Segaran Paperback \$26.39

Customers Who Bought This Item Also Bought

Page 1 of 17

Item	Author	Type	Price
Programming Collective Intelligence: Building Smart Web 2.0 Applications	Toby Segaran	Paperback	\$26.39
Machine Learning in Action	Peter Harrington	Paperback	\$25.75
Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and More Social Networks	Matthew A. Russell	Paperback	\$26.36
Data Analysis with Open Source Tools	Philipp K. Janert	Paperback	\$24.05
R Cookbook (O'Reilly Cookbooks)	Paul Teator	Paperback	\$32.43
The Art of R Programming: Tour of Statistical Analysis, Visualization, and Data Munging with R	Norman Matloff	Paperback	\$25.06

Are any of these items inappropriate for this page? [Let us know](#)

Зачем это нужно?

- Это круто
 - Сложные задачи
 - Движение к искусственному интеллекту
- Это полезно
 - Извлечение прибыли из данных
 - Data-driven companies

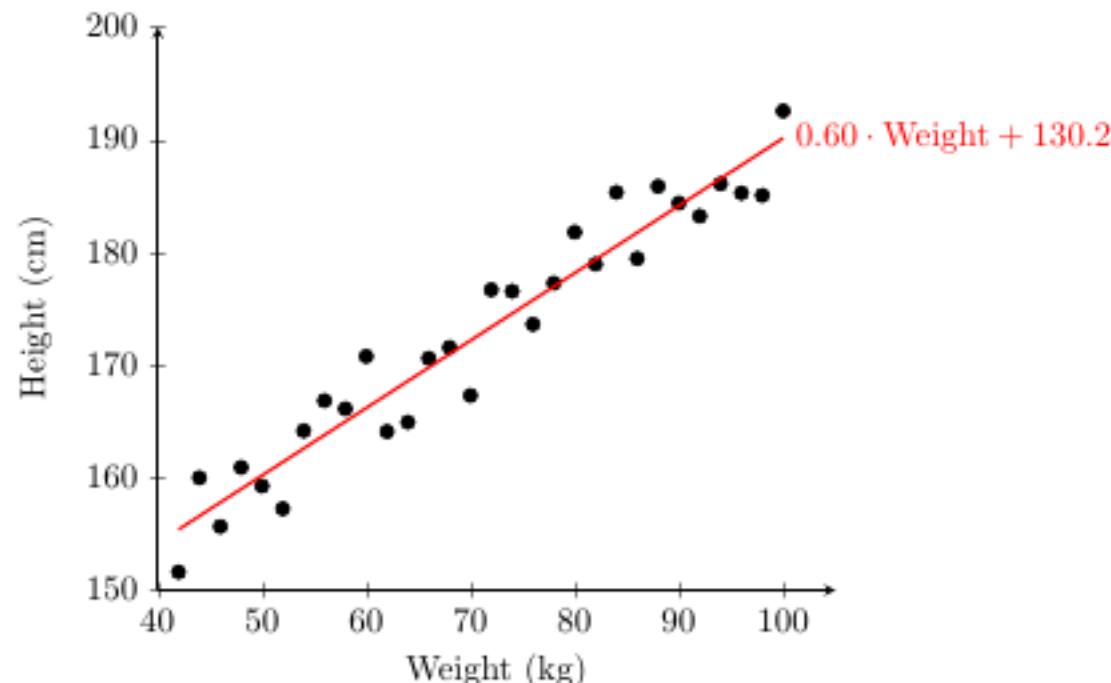
Как можно заниматься анализом данных?

- Data scientist
 - Работа с данными
 - Знание инструментов и методов
 - Опыт решения задач
- Менеджер
 - Понимание, как работает машинное обучение
 - Понимание узких мест, оценивание сроков
- Заказчик
 - Метрики качества
 - Требования к данным
 - Ограничения современных подходов

Типы ответов

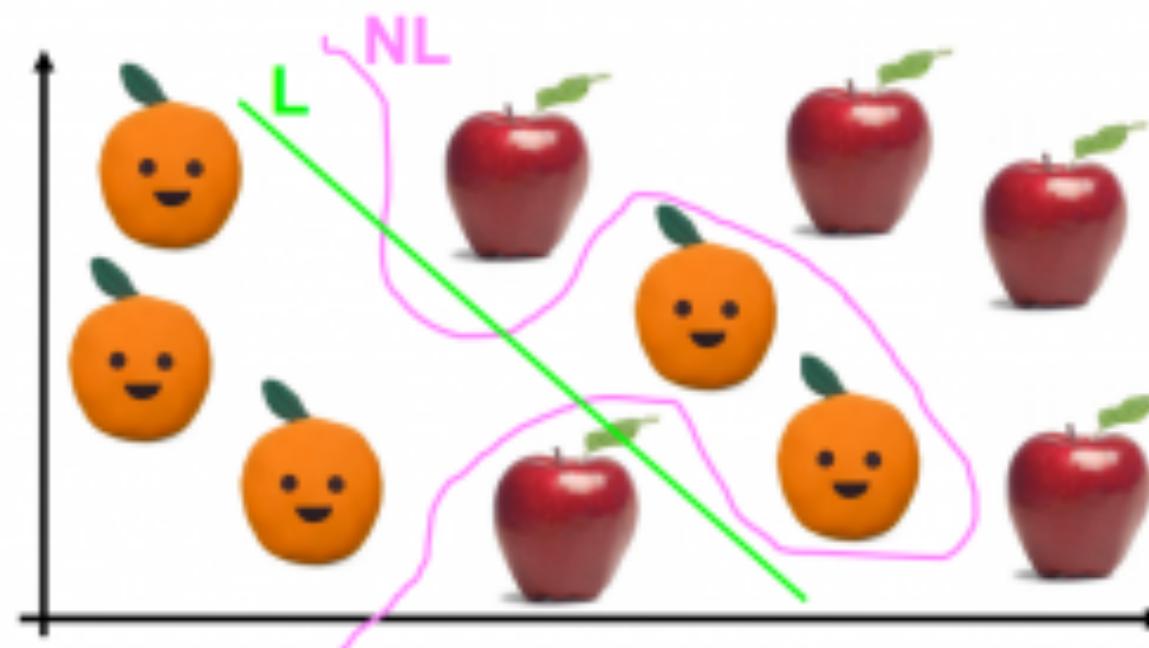
Регрессия

- Вещественные ответы: $\mathbb{Y} = \mathbb{R}$
- Пример: предсказание роста по весу



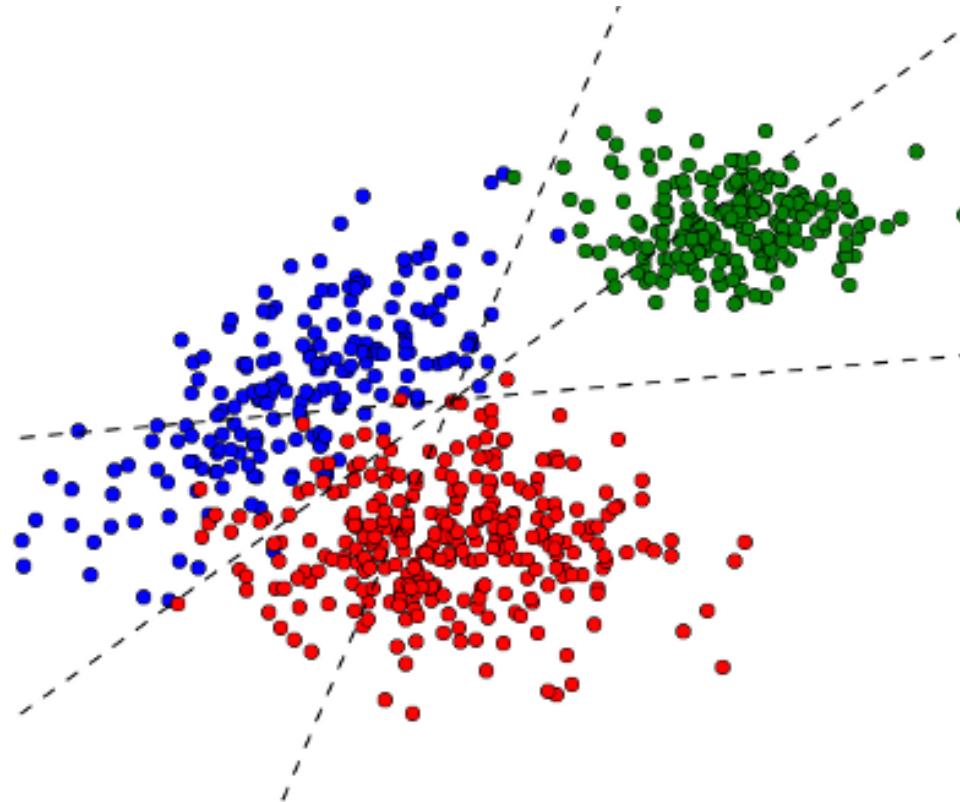
Классификация

- Конечное число ответов: $|\mathbb{Y}| < \infty$
- Бинарная классификация: $\mathbb{Y} = \{-1, +1\}$



Классификация

- Многоклассовая классификация: $\mathbb{Y} = \{1, 2, \dots, K\}$



Классификация

- Классификация с пересекающимися классами: $\mathbb{Y} = \{0, 1\}^K$
 - (multi-label classification)
- Ответ — набор из K нулей и единиц
- i -й элемент ответа — принадлежит ли объект i -му классу
- Какие темы присутствуют в статье?
- (математика, биология, экономика)

Ранжирование

- Набор документов d_1, \dots, d_n
- Запрос q
- Задача: отсортировать документы по *релевантности* запросу
- $a(q, d)$ — оценка релевантности

Ранжирование

Яндекс

картинки с котиками — 5 млн ответов



Найти

Поиск

[Картинки с кошками | Fun Cats — Забавные коты](#)

[funcats.by > pictures/](#) ▾

Картинки с кошками. Прикольные коты. 777 изображений. ... 32 изображения. Кошки

Стамбула. 41 изображение. Веселые котята.

Картинки

Видео

[Уморные котики \(57 фото\) » Бяки.нет | Картинки](#)

[byaki.net > Картинки > 14026-umornye-kotiki-57...](#) ▾

Бяки нет! . NET. Уморные котики (57 фото). 223. Коментариев:9Автор:4ertonok

Просмотров:161 395 Картинки28-10-2008, 00:03.

Карты

Маркет

Ещё

[Смешные картинки кошек с надписями | Лолкот.Ру](#)

[lolkot.ru](#) ▾

Смешные картинки для новых приколов! Сделать свой прикол очень просто. ... Котик

верит в чудеса. Он в носке подарок ищет...

[Красивые картинки и фото кошек, котят и котов](#)

[foto-zverey.ru > Кошки](#) ▾

Фото и картинки кошек и котят потрясающей красоты и нежности. Здесь мы собрали

такие изображения, которые всегда вызывают море положительных эмоций...

[Обои для рабочего стола Котята | картинки на стол Котята](#)

[7fon.ru > Чёрные обои и картинки > Обои котята](#) ▾

Картинки Котята с 1 по 15. Обои для рабочего стола Котята. ... Скачать Картинки Котята на рабочий стол бесплатно.

Прогнозирование временных рядов

- Позже — на примере

Построение рекомендательных систем

- Позже — на примере

Кластеризация

- \mathbb{Y} — отсутствует
- Нужно найти группы похожих объектов
- Сколько таких групп?
- Как измерить качество?

- Пример: сегментация пользователей мобильного оператора

Типы признаков

Типы признаков

- f_j — j -й признак
- D_j — множество значений признака

Бинарные признаки

- $D_j = \{0, 1\}$
- Доход клиента выше среднего по городу?
- Цвет фрукта — зеленый?

Вещественные признаки

- $D_j = \mathbb{R}$
- Возраст
- Площадь квартиры
- Количество звонков в колл-центр

Категориальные признаки

- D_j — неупорядоченное множество
- Цвет глаз
- Город
- Образование (может быть упорядоченным)

- Очень трудны в обращении

Порядковые признаки

- D_j — упорядоченное множество
- Воинское звание
- Роль в фильме (первого плана, второго плана, массовка)
- Тип населенного пункта

Множествозначные признаки

- (set-valued)
- D_j — множество всех подмножеств некоторого множества
- Какие фильмы посмотрел пользователь?
- Какие слова входят в текст?

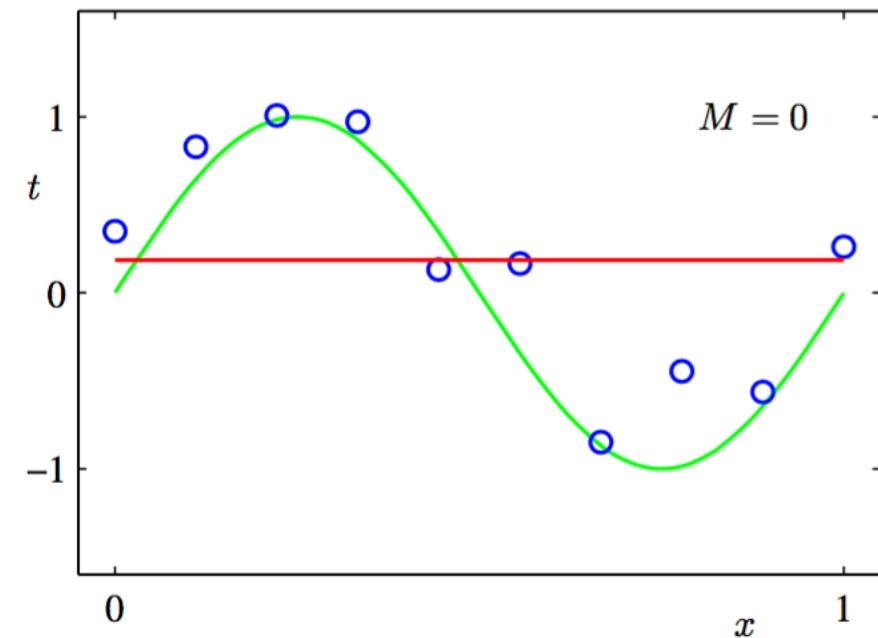
Обобщающая способность

Обобщающая способность

- Выбираем алгоритм с лучшим качеством на обучающей выборке
- Как он будет вести себя на новых данных?
- Смог ли он выразить y через x ?

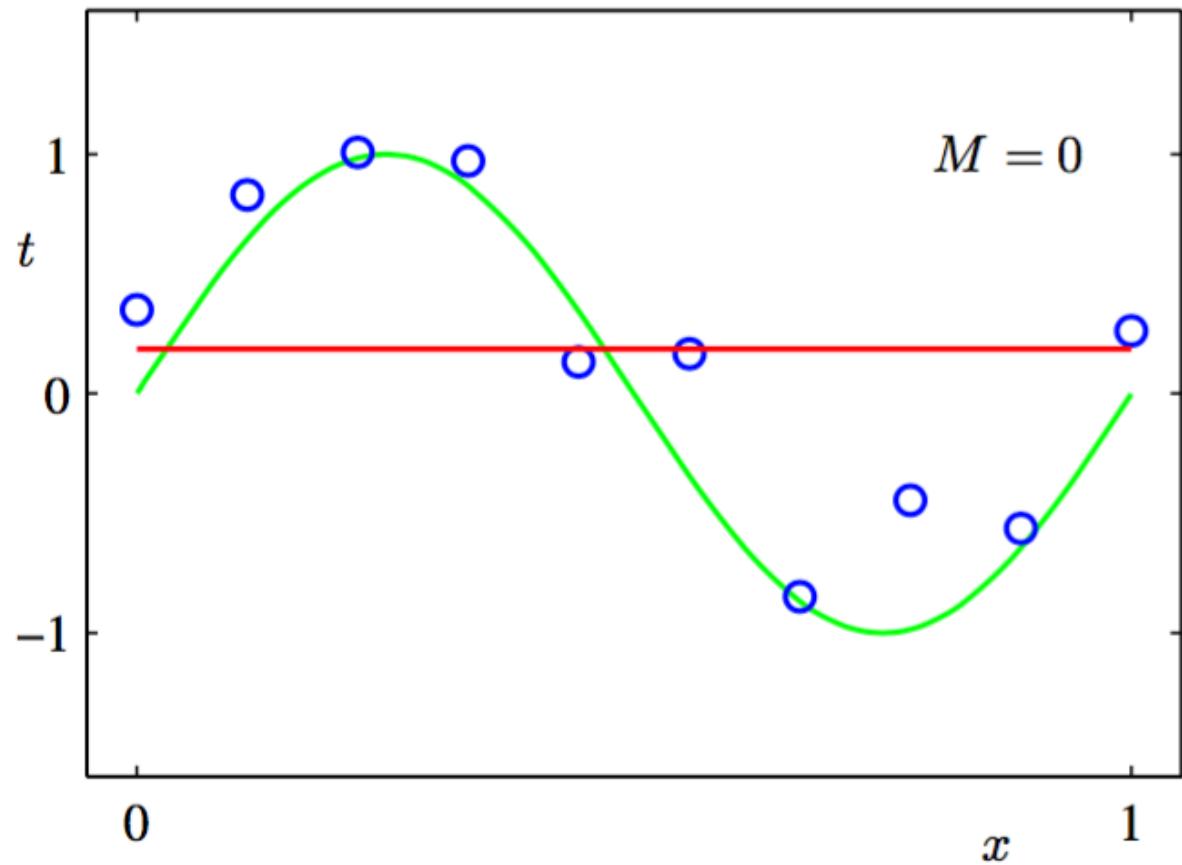
Обобщающая способность

- Зеленый — истинная зависимость
- Красный — прогноз алгоритма
- Синий — выборка
- Линейный алгоритм



Обобщающая способность

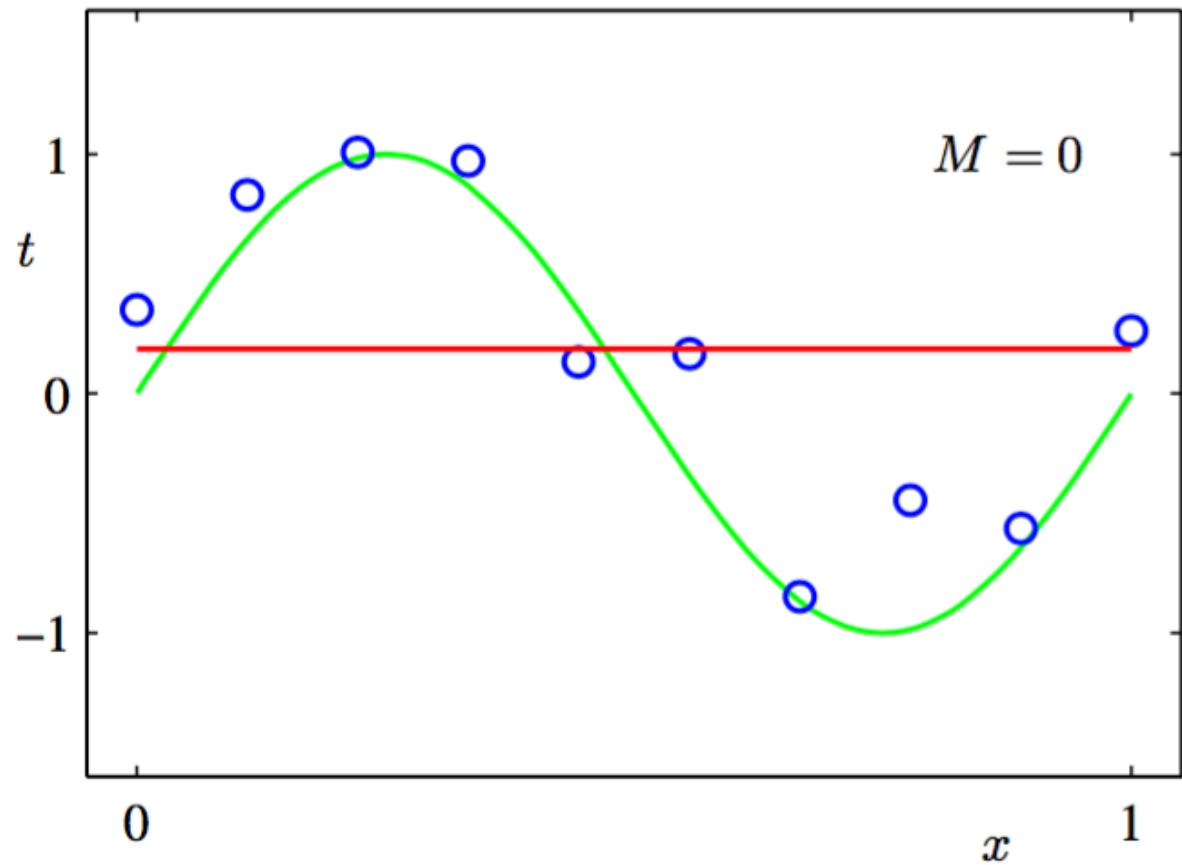
- Без признаков
- Константный алгоритм



Обобщающая способность

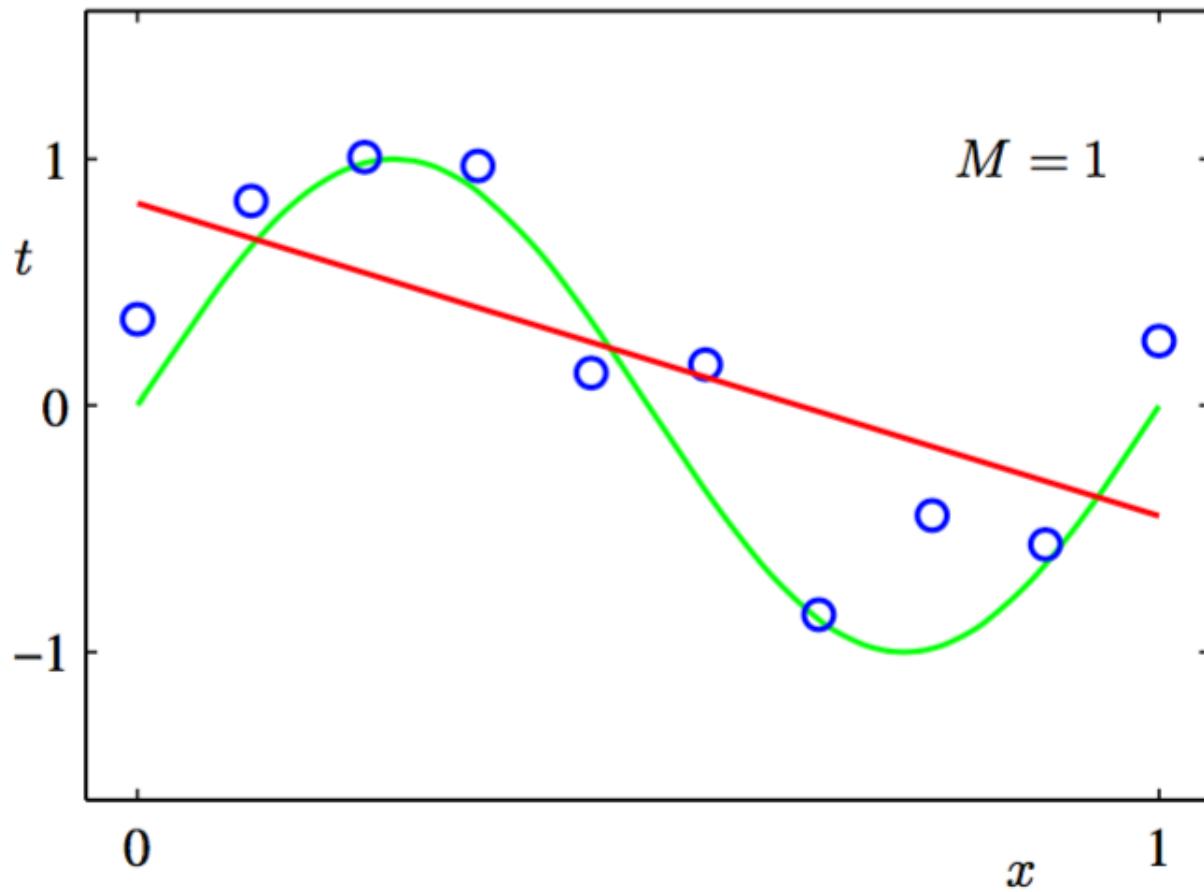
- Без признаков
- Константный алгоритм

Недообучение



Обобщающая способность

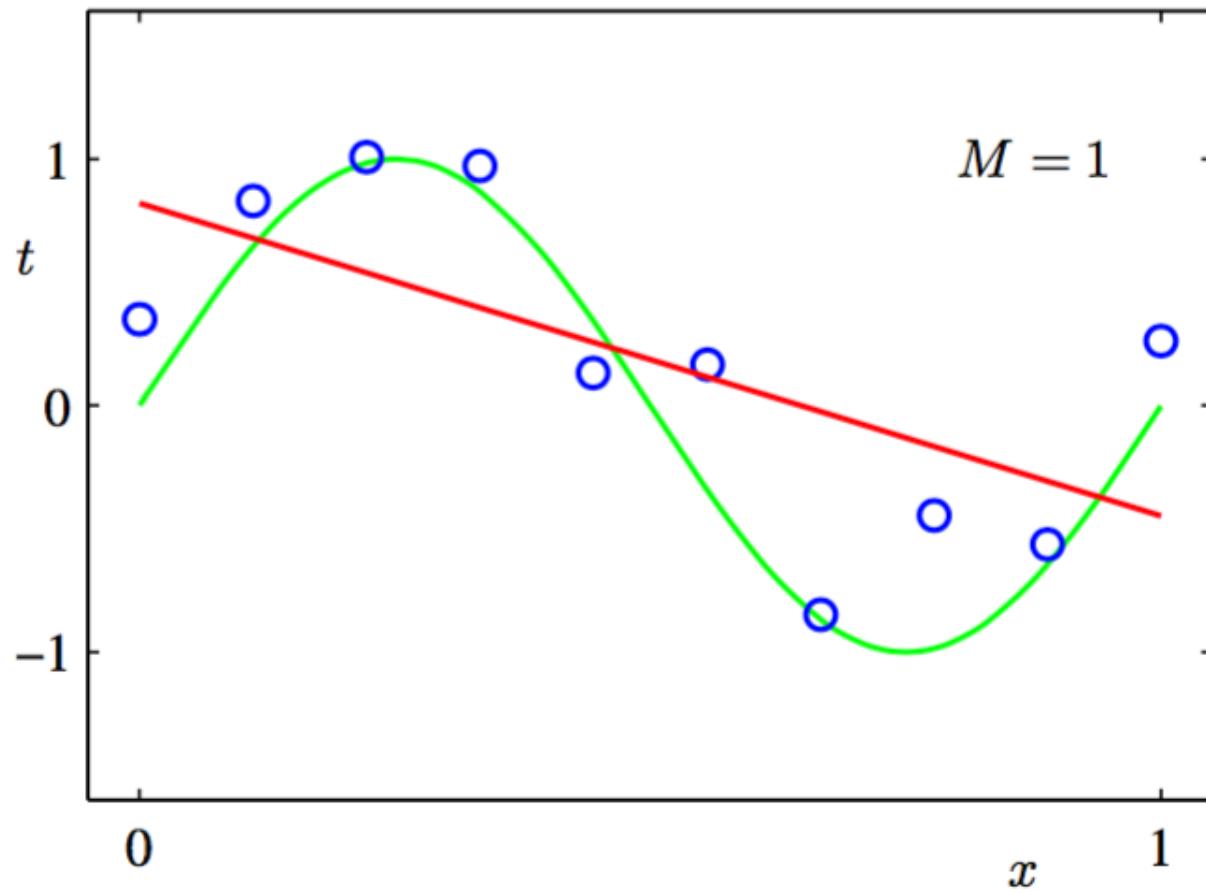
- 1 признак
- x



Обобщающая способность

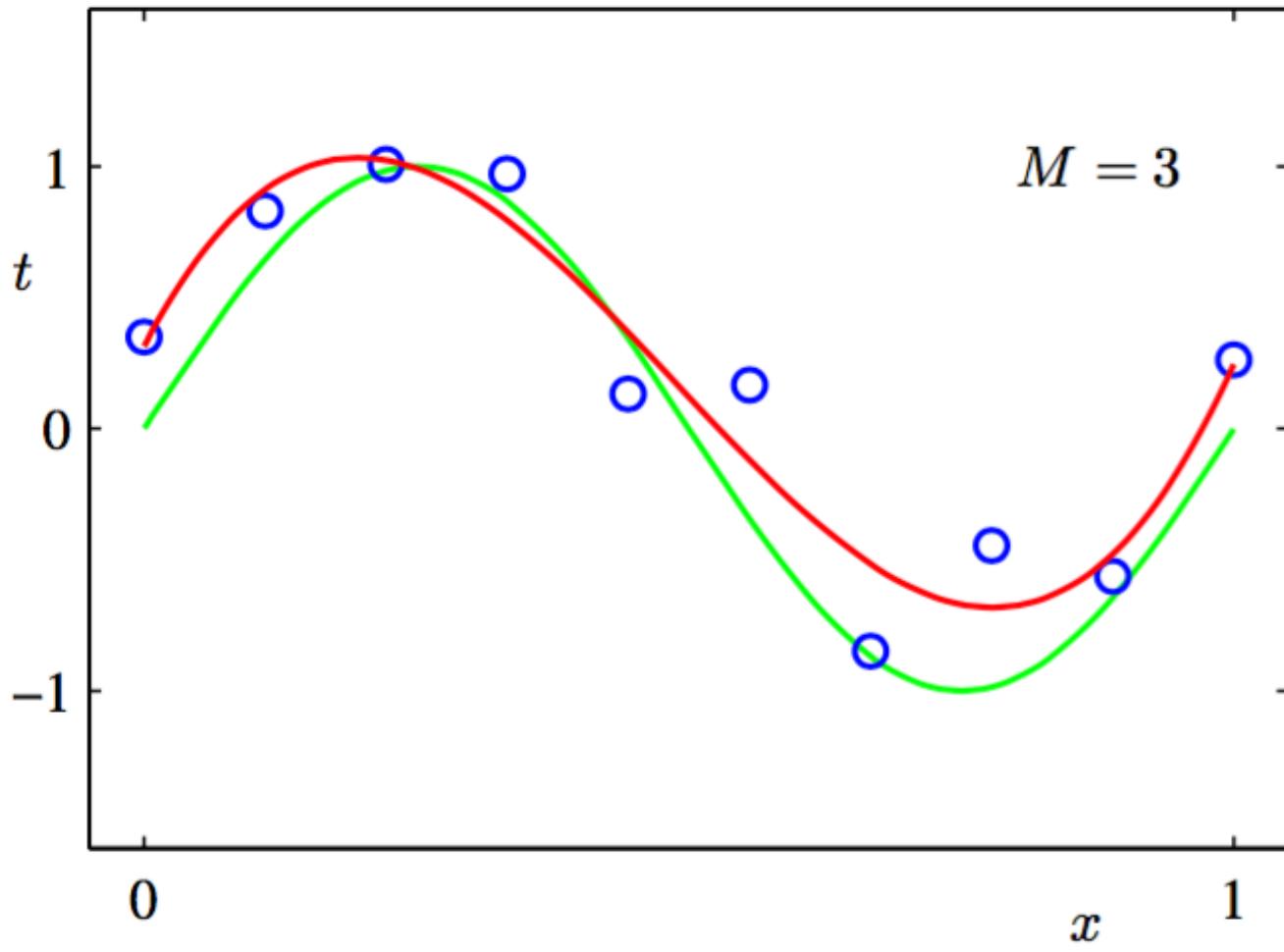
- 1 признак
- x

Недообучение



Обобщающая способность

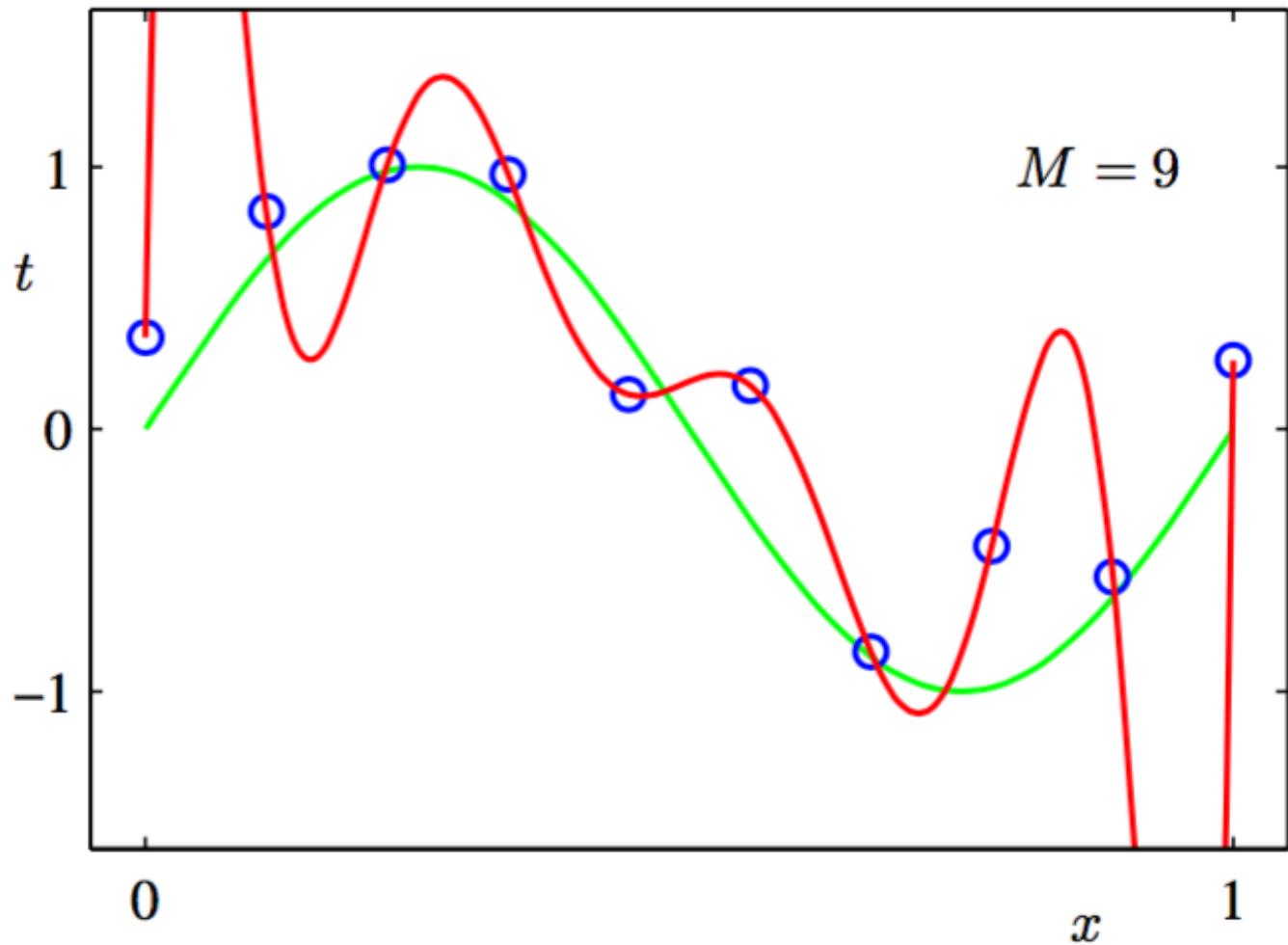
- 3 признака
- x, x^2, x^3



Обобщающая способность

- 9 признаков
- $x, x^2, x^3, x^4, \dots, x^9$

Переобучение
(overfitting)



Обобщающая способность

- Недообучение — **плохое** качество на обучении и на новых данных
- Переобучение — **хорошее** качество на обучении, **плохое** на новых данных
- Переобучение — алгоритм запоминает ответы, а не находит закономерности

Как выявить переобучение?

- Хороший алгоритм — хорошее качество на обучении
- Переобученный алгоритм — хорошее качество на обучении
- По обучающей выборке очень сложно выявить переобучение



Как выявить переобучение?

- Отложенная выборка — данные, на которых не обучались
- Кросс-валидация
- Меры сложности модели

Задачи анализа данных

Медицинская диагностика

- Объект — пациент в определенный момент времени
- Ответ — диагноз
- Классификация с пересекающимися классами

Медицинская диагностика — признаки

- Бинарные: пол, головная боль, слабость, и т.д.
- Порядковые: тяжесть состояния, желтушность, и т.д.
- Вещественные: возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т.д.

Медицинская диагностика — особенности

- Много пропусков в данных (missing data)
- Недостаточный объем данных
- Алгоритм должен быть интерпретируемым
- Нужна оценка вероятности для каждого заболевания

Кредитный скоринг

- Объект — заявка на выдачу кредита банком
- Ответ — вернет ли клиент кредит
- Бинарная классификация

Кредитный скоринг — признаки

- Бинарные: пол, наличие телефона, и т.д.
- Категориальные: место жительства, профессия, семейный статус, работодатель, и т.д.
- Порядковые: образование, должность, и т.д.
- Вещественные: возраст, зарплата, стаж работы, доход семьи, сумма кредита, и т.д.

Кредитный скоринг — особенности

- Нужно оценивать вероятность дефолта

Предсказание оттока клиентов

- Объект — абонент в определенный момент времени
- Ответ — уйдет или не уйдет в следующем месяце
- Бинарная классификация

Предсказание оттока клиентов — признаки

- Бинарные: корпоративный клиент, подключенные услуги, и т.д.
- Категориальные: регион проживания, тарифный план, и т.д.
- Вещественные: длительность разговоров, количество СМС, частота оплаты, объем трафика, и т.д.

Предсказание оттока клиентов — особенности

- Нужно оценивать вероятность ухода
- Сверхбольшие выборки
- Исходные данные — сырье логи

Стоимость недвижимости

- Объект — квартира в Москве
- Ответ — стоимость в рублях
- Регрессия

Стоимость недвижимости — признаки

- Бинарные: наличие балкона, мусоропровода, лифта, охраны, парковки, и т.д.
- Категориальные: район города, тип дома (кирпичный/блочный/панельный/монолит), ближайшая станция метро и т.д.
- Вещественные: число комнат, жилая площадь, расстояние до центра, до метро, возраст дома, и т.д.

Стоимость недвижимости — особенности

- Выборка неоднородная, меняется со временем
- Разнотипные признаки
- Нужна интерпретируемая модель

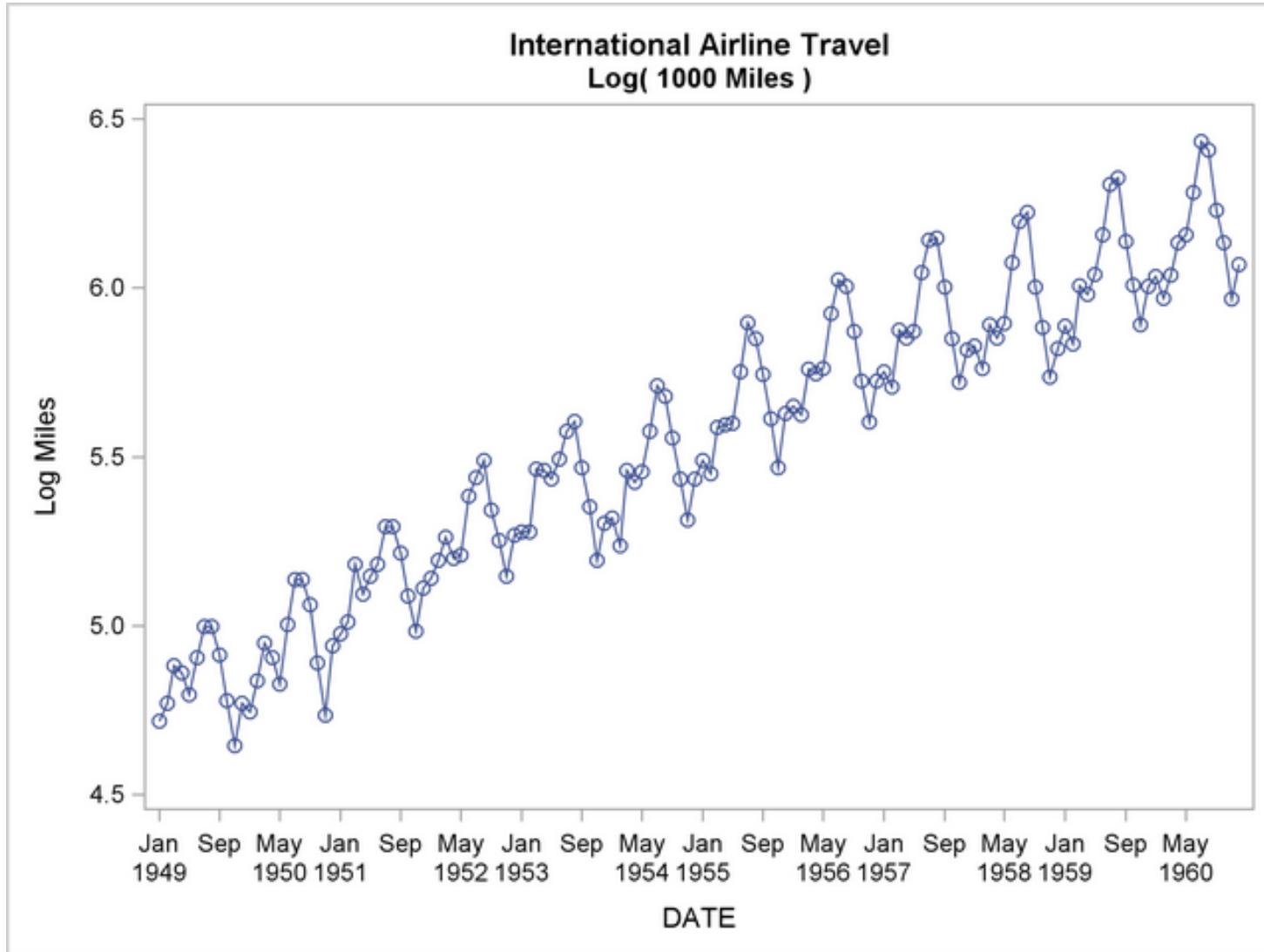
Прогнозирование продаж

- Объект — тройка (товар, магазин, день)
- Ответ — объем продаж
- Регрессия
- Прогнозирование временных рядов

Прогнозирование продаж — признаки

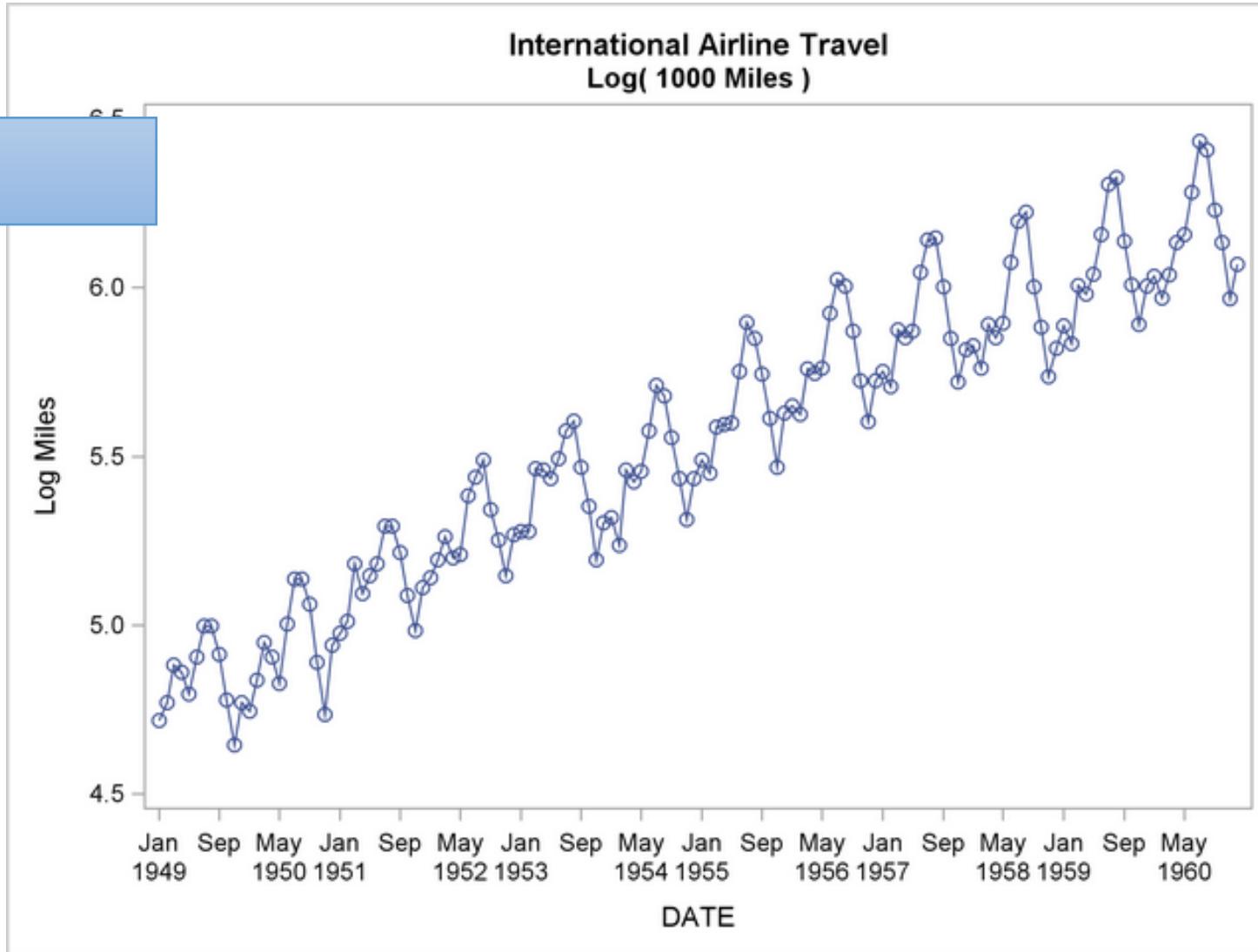
- Бинарные: выходной день, праздник, промоакция, и т.д.
- Вещественные: продажи в прошлые дни

Временные ряды



Временные ряды

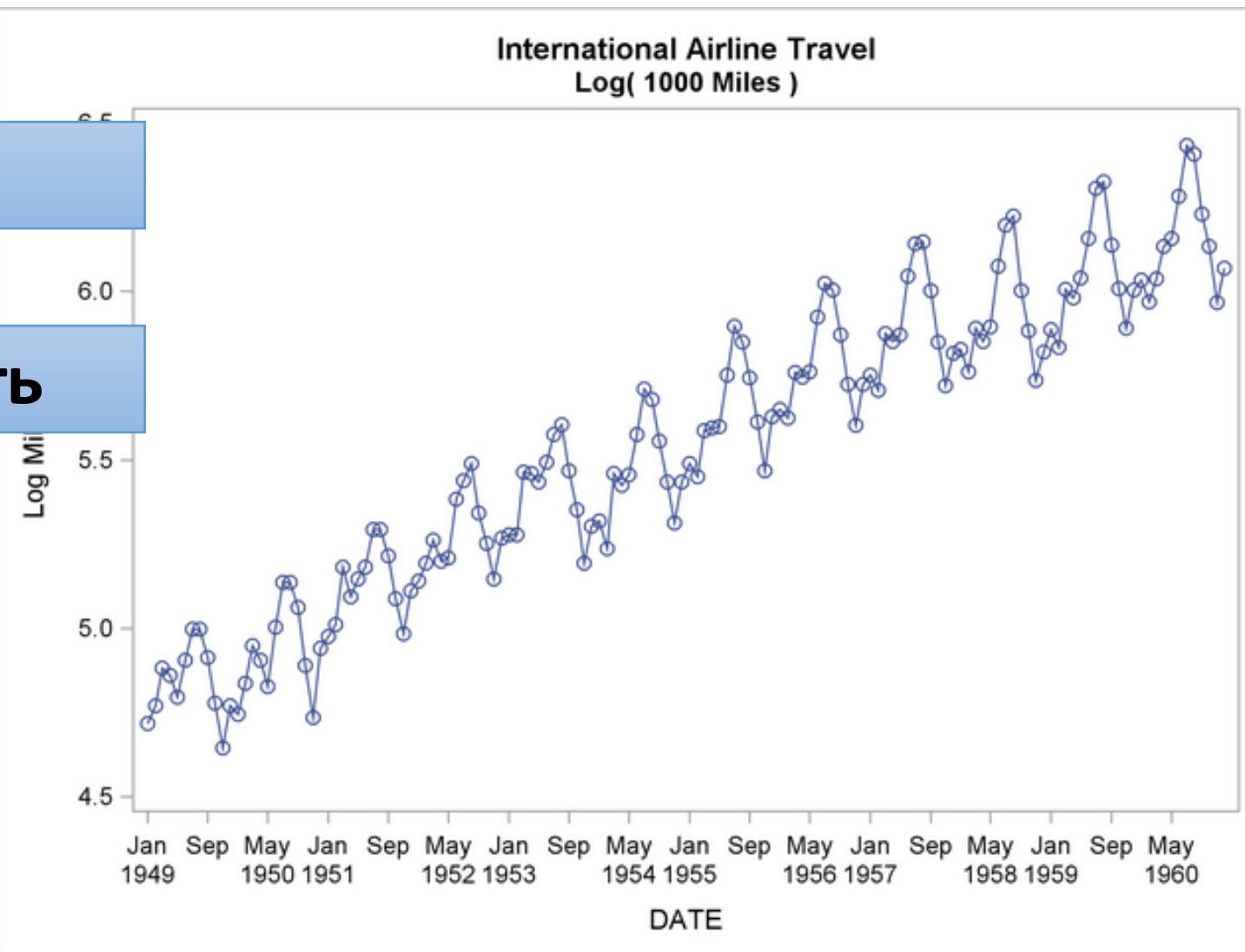
Тренд



Временные ряды

Тренд

Сезонность



Avito Context Ad Clicks Prediction

- kaggle.com
- Объект — тройка (пользователь, запрос, баннер)
- Ответ — кликнет ли пользователь по баннеру
- Классификация

Avito Context Ad Clicks Prediction — признаки

- Все действия пользователя на сайте
- Профиль пользователя (браузер, устройство, IP-адрес)
- История показов и кликов для других пользователей
- 10 таблиц с сырьими данными

Avito Context Ad Clicks Prediction — особенности

- Надо изобретать признаки
- Сотни миллионов показов
- Размер подготовленной выборки — терабайты
- Нужны технологии и алгоритмы работы с большими данными

Рекомендательная система фильмов

- Объект — пара (пользователь, фильм)
- Ответ — понравится ли пользователю фильм?
- Регрессия? Классификация?

Рекомендательная система — признаки

- Оценки фильмов от пользователей
- Возможно, профиль пользователя
- Возможно, информация о фильме

Рекомендательная система – Imhonet

Оценки фильма Любопытное стечеие обстоятельств

Een Bizarre Samenloop Van Omstandigheden, A Curious Conjunction of Coincidences

[Фильмы](#) / [Комедии](#) /
[обстоятельств»](#) /

Да, Вам стоит смотреть фильм «Любопытное
стечеие обстоятельств»

Людям, с оценками, похожими на [Ваши](#), этот фильм [нравится](#)



А ещё они рекомендуют Вам [31 фильм](#)

Ваша прогнозируемая оценка фильма после его просмотра
8.2 ★★★★★★★★☆☆

Смотрели? Оцените ★★★★★★★★☆☆☆

[Не рекомендовать](#)

[Про фильм](#) [Онлайн](#) [Скачать](#) [Отзывы](#) [Персоны](#) [Кадры](#) [Оценки](#) [Похожие](#)

Распределение оценок



Всего оценок — 9

Кому больше нравится

Кому нравится: 99% 1%

Рекомендательная система — Amazon

Frequently Bought Together



Price For All Three: \$86.01

Add all three to Cart Add all three to Wish List

Show availability and shipping details

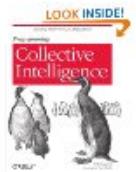
This item: Machine Learning for Hackers by Drew Conway Paperback \$33.87

Machine Learning in Action by Peter Harrington Paperback \$25.75

Programming Collective Intelligence: Building Smart Web 2.0 Applications by Toby Segaran Paperback \$26.39

Customers Who Bought This Item Also Bought

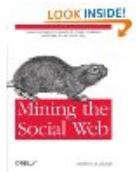
Page 1 of 17



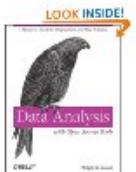
Programming Collective
Intelligence: Building ...
▶ Toby Segaran
 (84)
Paperback
\$26.39



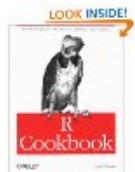
Machine Learning in Action
▶ Peter Harrington
 (10)
Paperback
\$25.75



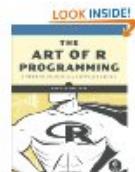
Mining the Social Web:
Analyzing Data from ...
▶ Matthew A. Russell
 (19)
Paperback
\$26.36



Data Analysis with Open
Source Tools
▶ Philipp K. Janert
 (29)
Paperback
\$24.05



R Cookbook (O'Reilly
Cookbooks)
▶ Paul Teator
 (18)
Paperback
\$32.43



The Art of R Programming: A
Tour of Statistical ...
Norman Matloff
 (29)
Paperback
\$25.06

Are any of these items inappropriate for this page? [Let us know](#)

Рекомендательная система — особенности

- Много метрик для оптимизации: число кликов по рекомендациям, число новых для пользователя товаров, разнообразие предлагаемых жанров, и т.д.
- Особый вид данных: (пользователь, фильм/товар, оценка)
- Получение оценки — явный и неявный отклик

Резюме

- Машинное обучение — восстановление зависимостей по данным
- Классы задач: регрессия, классификация, кластеризация, ранжирование, прогнозирование и т.д.
- Переобучение и обобщающая способность