Введение в анализ данных

Лекция 1

Введение

Евгений Соколов

sokolov.evg@gmail.com

НИУ ВШЭ, 2016

Как перевести часы в минуты?

Как перевести часы в минуты?

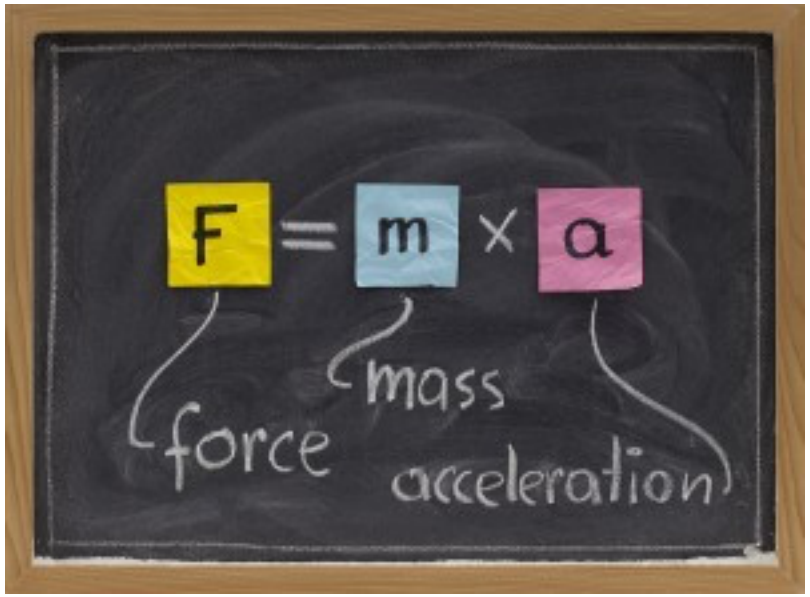
- x — часы
- $f(x) = 60x$ — преобразование в минуты, функция

Какая сила приложена к телу?

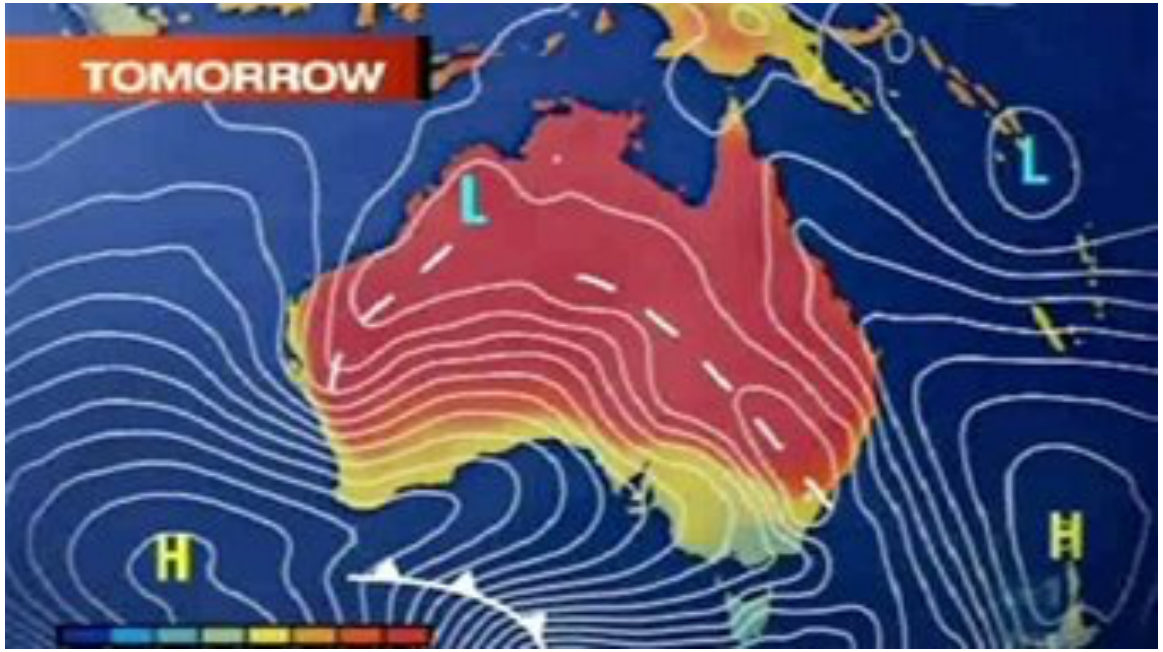
- Известны масса тела m и его ускорение a
- Чему равна сила F?

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

Дифференциальные уравнения

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

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} +$$

$$\frac{\partial w}{\partial t} + u\frac{\partial w}{\partial x} + v\frac{\partial w}{\partial y} +$$

Позволяют найти скорость воздуха и давление в любой точке

Очень тяжело решать

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

Анализ тональности текста

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

Анализ тональности текста

«Большое спасибо! Сюдя 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?
- 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

Про лектора

- Евгений Соколов
- Почта: sokolov.evg@gmail.com
- ФКН ВШЭ, преподаватель
- Yandex Data Factory, руководитель группы

Про курс

- wiki: http://wiki.cs.hse.ru/Майнор_Интеллектуальный_анализ_данных/Введение_в_анализ_данных
- Два модуля
- Домашние задания (7-8 штук)
- Проверочные работы
- Проект по анализу данных
- 3й модуль — зачет
- 4й модуль — экзамен

Про оценку

- $О_{\text{текущая 1 модуль}} = 0.8 * О_{\text{дз}} + 0.2 * О_{\text{проект}}$
- $О_{\text{текущая 2 модуль}} = 0.6 * О_{\text{дз}} + 0.4 * О_{\text{проект}}$
- $О_{\text{накопленная}} = 0.8 * О_{\text{текущая}} + 0.2 * О_{\text{ауд}}$
- $О_{\text{промежуточная}} = 0.5 * О_{\text{накопленная 1 модуль}} + 0.5 * О_{\text{зачет}}$
- $О_{\text{накопленная итог}} = 0.5 * О_{\text{промежуточная}} + 0.5 * О_{\text{накопленная 2 модуль}}$
- $О_{\text{итог}} = 0.6 * О_{\text{накопленная итог}} + 0.4 * О_{\text{экзамен}}$

Про план курса

- Введение
- Математика для анализа данных
- Линейные и логические методы, кластеризация
- Композиции алгоритмов
- Оценивание качества алгоритмов
- Работа с реальными данными

Про семинары

- Формируем данные: http://goo.gl/forms/EpBjdfWbZU

Про литературу

- Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014.
- Boris Mirkin. Core Concepts in Data Analysis: Summarization, Correlation, Visualization. 2010.
- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning. 2013.

Про литературу

- Курс К.В. Воронцова
- https://github.com/esokolov/ml-course-msu
- https://www.coursera.org/learn/**machine-learning**
- https://www.coursera.org/learn/**introduction-machine-learning**
- Специализация «Машинное обучение и анализ данных», Coursera

Пример задачи

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

* см. kaggle.com, TFI Restaurant Revenue Prediction

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

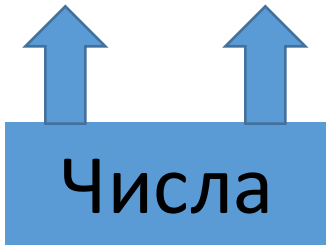
- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x^1, \dots, x^d)$ — признаковое описание

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

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: $a(x) = w_1 x^1 + \cdots + w_d x^d$

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

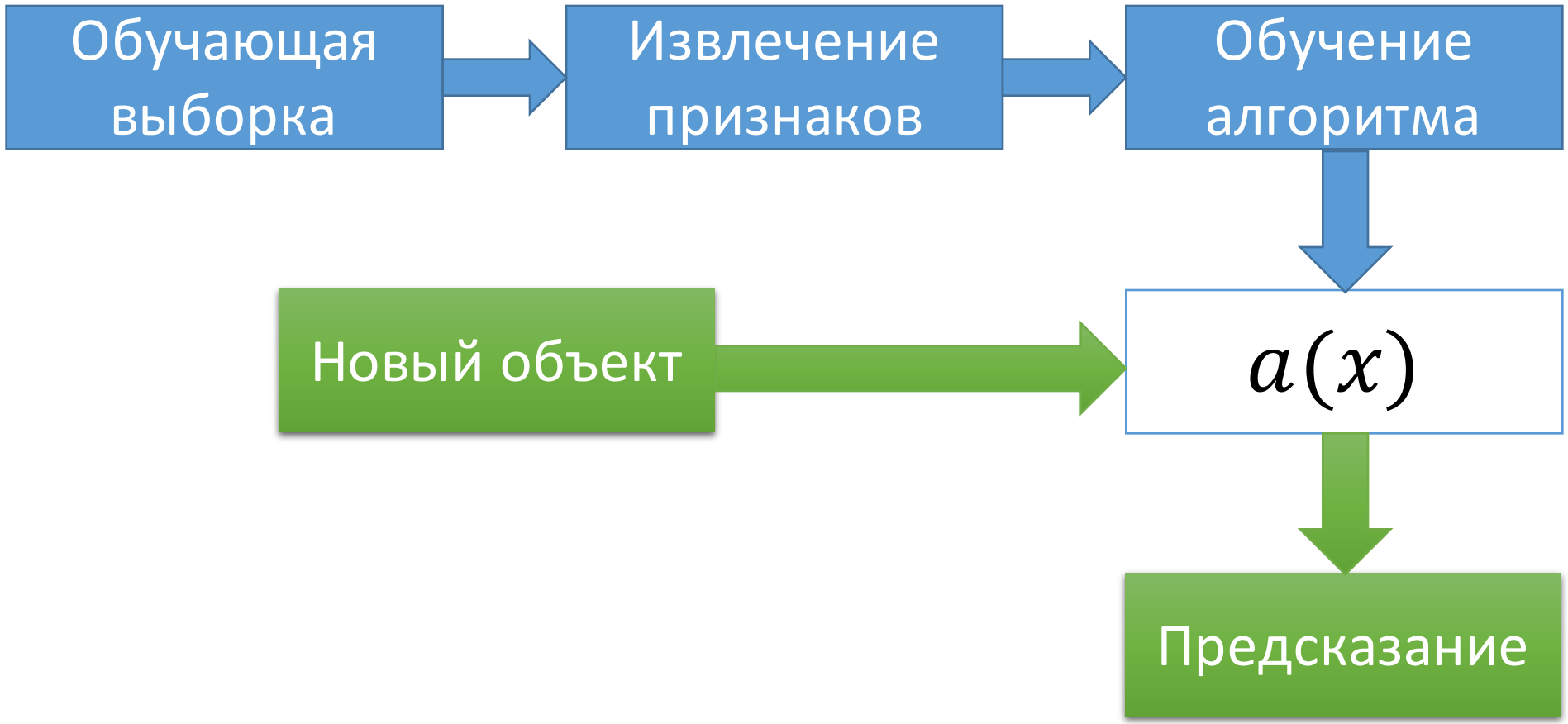
Функционал качества

- 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
  - $\mathcal{A} = \{w_1 x^1 + \cdots + w_d x^d \mid w_1, \ldots, w_d \in \mathbb{R}\}$
- 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?

На следующей лекции

- Типы задач в машинном обучении
- Типы признаков
- Примеры прикладных задач