НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Разработка управленческих решений в маркетинге

ПРОДАЖИ

Или все про регрессию

Элен Теванян
Филипп Ульянкин

Москва, 2018

Основные понятия машинного обучения

Есть данные!
Хотим извлечь знания из них

Или нет данных, но знания извлечь хотим…….

ОСНОВНЫЕ ПОНЯТИЯ

- $\boldsymbol{x}$ **(sample)** – объекты, с которыми мы хотим что-то делать. В нашем случае – потребители. Обязательно есть всегда.

- $\boldsymbol{y}$ **(target)** – ответ, целевая переменная. То, что свойственно объекту и то, что мы хотим научиться прогнозировать/объяснять. Не всегда есть.

- $(\boldsymbol{x_i}, \boldsymbol{y_i})_{\boldsymbol{i=1}}^{\boldsymbol{\ell}}$ – обучающая выборка, прецеденты, т.е. все объекты, для которых известны значения целевого признака
- $\boldsymbol{\ell}$ – размер выборки.

- Объекты характеризуются признаками (фичами, features)

ОСНОВНЫЕ ПОНЯТИЯ

- Т.е. Если совсем просто, то у нас есть таблица

	f_1	f_2	f_3	y
x_1				
x_2				
x_3				

ОСНОВНЫЕ ПОНЯТИЯ

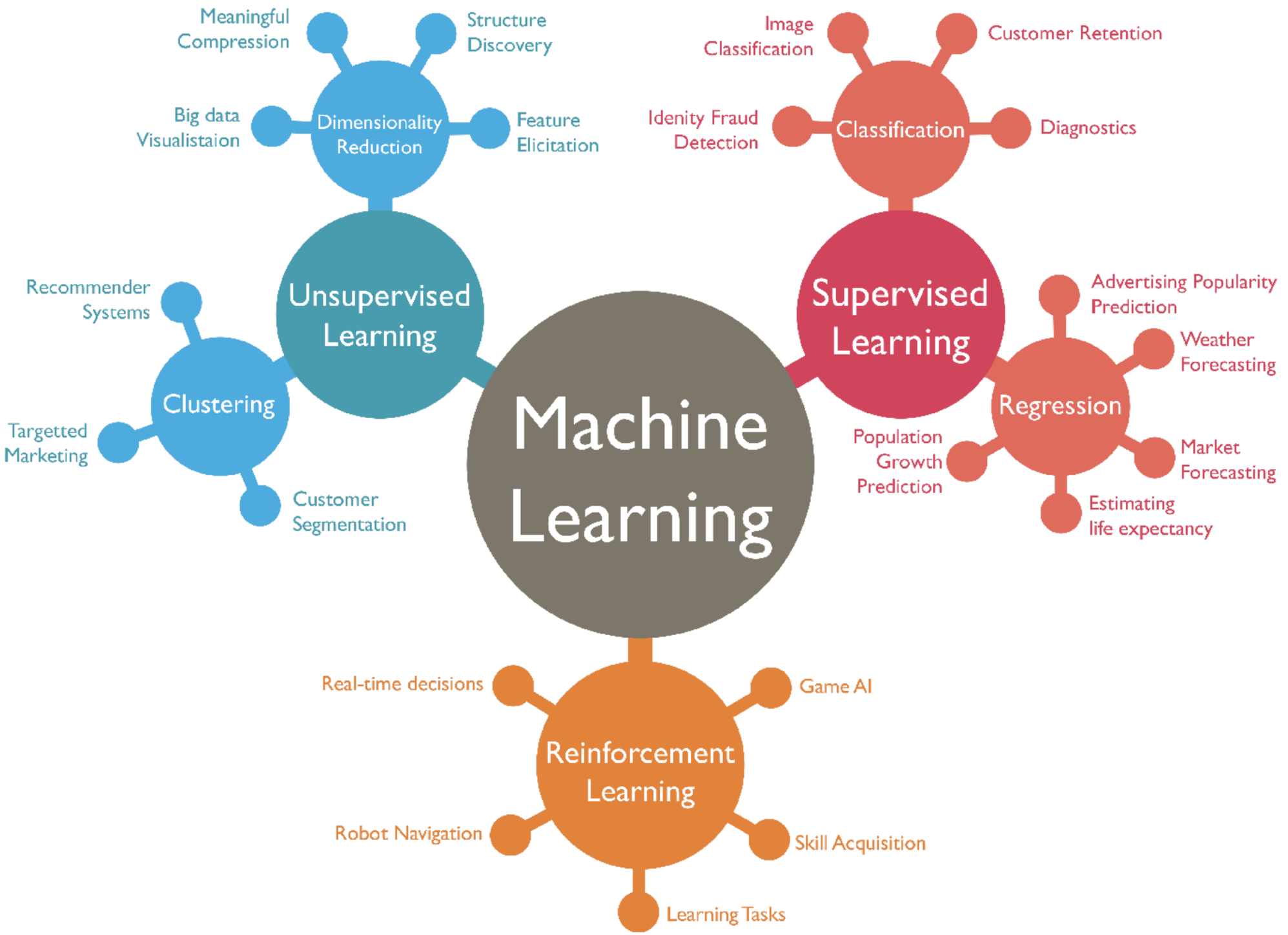
- Например, у нас есть три потребителя с корзинами в интернет-магазине.

	Свитер	Юбка	Поясная сумка	Совершена покупка
Катя	3	1	2	Да
Лера	1	1	1	Нет
Витя	0	0	2	Да

объекты

признаки

целевая переменная


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

ПОДХОДЫ К ОБУЧЕНИЮ

- Обучение с учителем

❑Классификация

❑Регрессия

❑Ранжирование

- Обучение без учителя

❑Кластеризация

❑Уменьшение размерности

- Обучение с частичным привлечением учителя
- Обучение с подкреплением

ОБУЧЕНИЕ С УЧИТЕЛЕМ

- Есть вектор с целевой переменной

	Свитер	Юбка	Поясная сумка	Совершена покупка
Катя	3	1	2	Да
Лера	1	1	1	Нет
Витя	0	0	2	Да

объекты

признаки

целевая переменная

ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ

- Нет вектора с целевой переменной

	Свитер	Юбка	Поясная сумка	
Катя	3	1	2	
Лера	1	1	1	
Витя	0	0	2	

объекты

признаки

пусто ☹

ЧТО ДЕЛАЕМ?

- $\boldsymbol{x_i}$ – объект (*потребитель*)
- $\boldsymbol{y_i}$ – целевая переменная (*сегмент*)
- $(\boldsymbol{x_i}, \boldsymbol{y_i})$ – прецедент
- **Обучающая выборка** – набор всех прецедентов

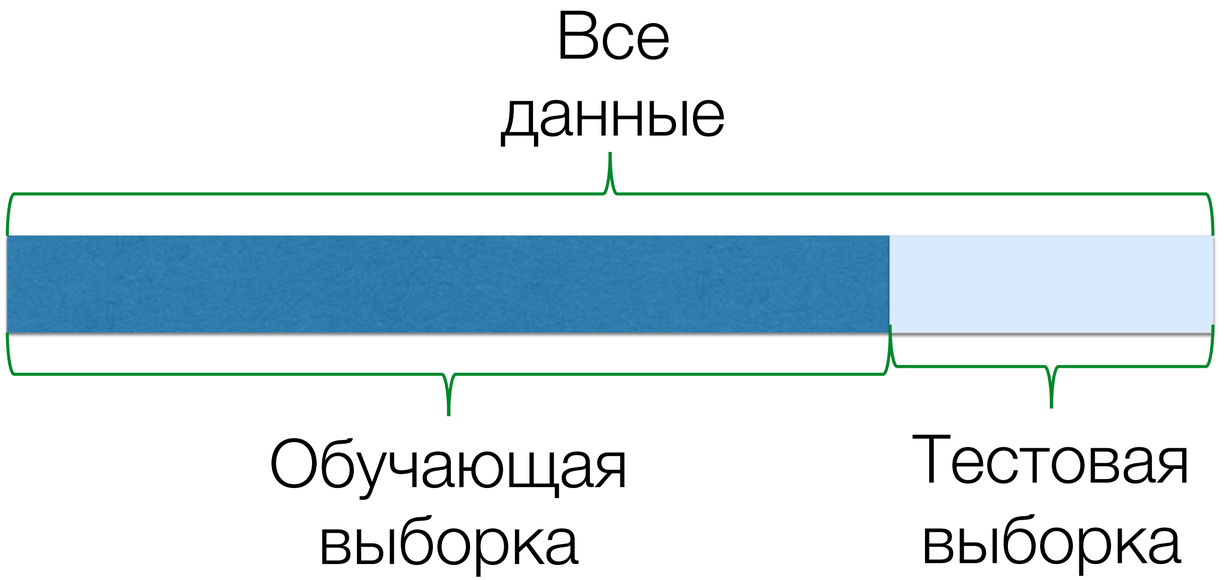
Как решить эту задачу?

Найти алгоритм $\boldsymbol{a(x)}: \boldsymbol{a(x_i)} \approx \boldsymbol{y_i}$

ЧТО ДЕЛАЕМ?

- Алгоритм (модель) – это формула, учитывающая характеристики объекта
- Формулы могут быть любыми

КАК НАЙТИ ЛУЧШЕЕ РЕШЕНИЕ?



- Алгоритм обучается на обучающей выборке
- Алгоритм тестируется на тестовой выборке (валидационной)

ФЛЭШБЭК
из математики

Производная????

ПРОИЗВОДНАЯ

- Пусть дана функция $f(x)$
- Производной в точке называется:

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x_0 + \Delta x) - f(x_0)}{x_0 + \Delta x \ - x_0}$$

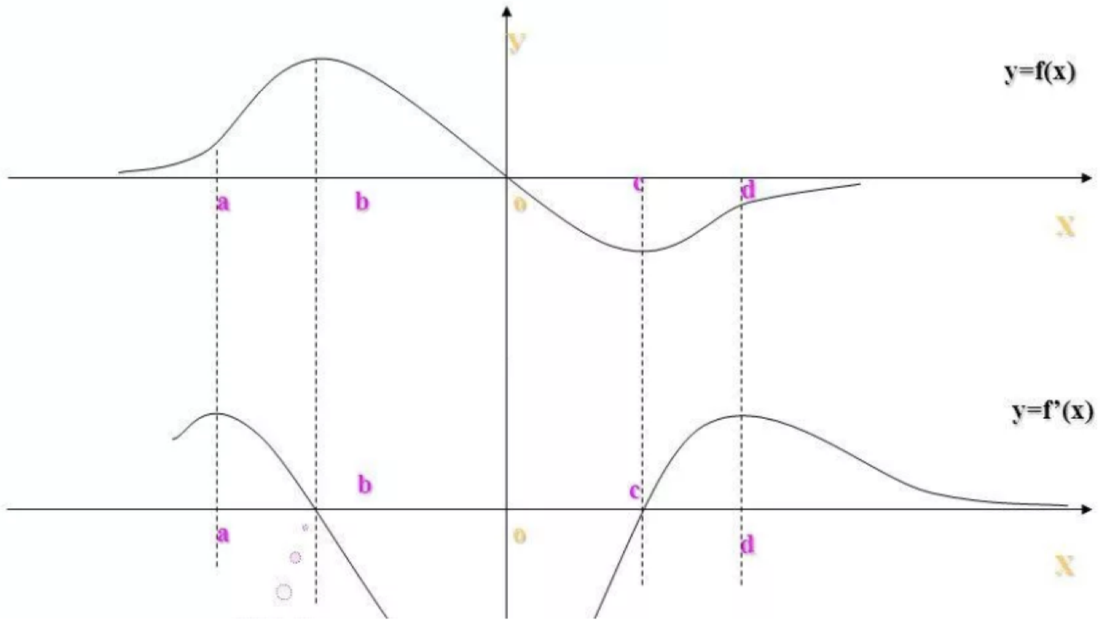


ТАБЛИЦА ПРОИЗВОДНЫХ

Таблица производных

1. $\left(u^{\alpha}\right)' = \alpha u^{\alpha-1}u'$, $\alpha = const$	
2. $\left(\frac{1}{u}\right)' = -\frac{1}{u^2}u'$	3. $\left(\sqrt{u}\right)' = \frac{1}{2\sqrt{u}}u'$
4. $\left(e^{u}\right)' = e^{u}u'$	5. $\left(a^{u}\right)' = a^{u}\ln a\, u'$
6. $\left(\ln u\right)' = \frac{1}{u}u'$	7. $\left(\log_a u\right)' = \frac{1}{u\ln a}u'$
8. $\left(\sin u\right)' = \cos u \cdot u'$	9. $\left(\arcsin u\right)' = \frac{1}{\sqrt{1-u^2}}u'$

Функция многих переменных

Задача регрессии

ПОСТАНОВКА ЗАДАЧИ

- **Задача**: найти алгоритм по прецедентам, который будет для каждого нового объекта делать предсказания

- $\boldsymbol{x_i}$ – объект (*потребитель*)
- $\boldsymbol{y_i}$ – целевая переменная (объем его заказа)
- $(\boldsymbol{x_i}, \boldsymbol{y_i})$ – прецедент
- **Обучающая выборка** – набор всех прецедентов

- $\boldsymbol{y_i} \in \mathbb{R}$ - вещественное число, т.е. число, в том числе и с дробной частью

РЕГРЕССИЯ. ПРИМЕРЫ

- Прогнозирование цены дома
- Прогнозирование заработной платы по описанию вакансии
- Прогнозирование спроса на товар в ближайшую неделю
- Прогнозирование уровня экспрессии гена
- Прогнозирование температуры воздуха
- Прогнозирование суммы компенсаций по страховке
- Прогнозирование объема потребления электроэнергии

РЕГРЕССИЯ. МОДЕЛЬНАЯ ЗАДАЧА НА СЕГОДНЯ

- Нам упала задача! Спрогнозировать продажи отдела фермерских продуктов крупной сети супермаркетов

- Как поставим задачу машинного обучения?

РЕГРЕССИЯ. МОДЕЛЬНАЯ ЗАДАЧА НА СЕГОДНЯ

- x – это один отдел одного супермаркета
- y – его продажи
- Есть следующая информация

Кол-во жителей в радиусе 3 км от супермаркета	Кол-вол фитнес-клубов и других спортивных комплексов в районе расположения супермаркета	Количество конкурирующих магазинов в радиусе 3 км от супермаркета	Выручка за 2017 год
100 000	7	2	5 000 000 рублей
570 000	3	0	3 000 000 рублей
400 000	19	6	7 000 000 рублей

РЕГРЕССИЯ. МОДЕЛЬНАЯ ЗАДАЧА НА СЕГОДНЯ

- Как спрогнозировать?

Кол-во жителей в радиусе 3 км от супермаркета	Кол-вол фитнес-клубов и других спортивных комплексов в районе расположения супермаркета	Количество конкурирующих магазинов в радиусе 3 км от супермаркета	Выручка за 2017 год	Прогноз
100 000	7	2	5 000 000 рублей	
570 000	3	0	3 000 000 рублей	
400 000	19	6	7 000 000 рублей	

РЕГРЕССИЯ. МОДЕЛЬНАЯ ЗАДАЧА НА СЕГОДНЯ

- Как спрогнозировать?
- Предположим, прогнозом для каждого отдела будет среднее за весь период
- Хорошо ли или плохо?

Кол-во жителей в радиусе 3 км от супермаркета	Кол-вол фитнес-клубов и других спортивных комплексов в районе расположения супермаркета	Количество конкурирующих магазинов в радиусе 3 км от супермаркета	Выручка за 2017 год	Прогноз
100 000	7	2	5 000 000 рублей	5 000 000 рублей
570 000	3	0	3 000 000 рублей	5 000 000 рублей
400 000	19	6	7 000 000 рублей	5 000 000 рублей

РЕГРЕССИЯ. МОДЕЛЬНАЯ ЗАДАЧА НА СЕГОДНЯ

- Как спрогнозировать?
- Предположим, прогнозом для каждого отдела будет среднее за весь период
- Хорошо ли или плохо?

Кол-во жителей в радиусе 3 км от супермаркета	Кол-вол фитнес-клубов и других спортивных комплексов в районе расположения супермаркета	Количество конкурирующих магазинов в радиусе 3 км от супермаркета	Выручка за 2017 год	Прогноз	Ошибка
100 000	7	2	5 000 000 рублей	5 000 000 рублей	0 рублей
570 000	3	0	3 000 000 рублей	5 000 000 рублей	- 2 000 000 рублей
400 000	19	6	7 000 000 рублей	5 000 000 рублей	2 000 000 рублей

РЕГРЕССИЯ. МОДЕЛЬНАЯ ЗАДАЧА НА СЕГОДНЯ

- Суммарная ошибка: 0 + (- 2 000 000) + 2 000 000 = 0 руб.
- ИДЕАЛЬНО! Но нет

Кол-во жителей в радиусе 3 км от супермаркета	Кол-вол фитнес-клубов и других спортивных комплексов в районе расположения супермаркета	Количество конкурирующих магазинов в радиусе 3 км от супермаркета	Выручка за 2017 год	Прогноз	Ошибка
100 000	7	2	5 000 000 рублей	5 000 000 рублей	0 рублей
570 000	3	0	3 000 000 рублей	5 000 000 рублей	- 2 000 000 рублей
400 000	19	6	7 000 000 рублей	5 000 000 рублей	2 000 000 рублей

РЕГРЕССИЯ. МОДЕЛЬНАЯ ЗАДАЧА НА СЕГОДНЯ

- Пусть $\boldsymbol{y_i}$ - это действительные значения целевой переменной
- $\boldsymbol{\hat{y}_i}$ - это прогноз, который мы сделали
- $\boldsymbol{e_i = y_i - \hat{y}_i}$ - ошибка предсказания для одного наблюдения
- $\boldsymbol{e_i^2 = (y_i - \hat{y}_i)^2}$ - квадрат ошибки

- Хотим сделать так, что:

- $\sum_{i=1}^{l} \boldsymbol{e_i^2} \rightarrow min$

КАК СТРОИТЬ ПРОГНОЗ?

- $\boldsymbol{x_i} = (\boldsymbol{f_1^i}, \ldots . \boldsymbol{f_n^i})$ – каждый объект описан признаками
- Можно поставить универсальный прогноз, как сделали раньше
- Можно придумать формулу, которая учтет все признаки
- $\widehat{\boldsymbol{y_i}}$= F($\boldsymbol{f_1^i}, \ldots . \boldsymbol{f_n^i}$)
- В зависимости от типа формулы модели могут быть линейные или нелинейные

- Мы очень близко познакомимся с линейными

МОДЕЛЬ ЛИНЕЙНОЙ РЕГРЕССВИИ

- $\boldsymbol{x} = \left(\boldsymbol{f_1^i}, \ldots . \boldsymbol{f_n^i}\right)$ – каждый объект описан признаками
- $\hat{y} = \boldsymbol{w_0} + \boldsymbol{w_1 f_1} + \cdots + \boldsymbol{w_n f_n}$ – модель линейной регрессии.
- Как найти $\boldsymbol{w_0}, \ldots . , \boldsymbol{w_n}$?

- Давайте начнем с простого. Пусть мы хотим строить прогноз только по одной характеристике.
- $\hat{y} = \boldsymbol{w_0} + \boldsymbol{w_1 f_1}$
- Помните про $\sum_{i=1}^{l} \boldsymbol{e_i^2} \to min$?

МОДЕЛЬ ЛИНЕЙНОЙ РЕГРЕССИИ

- $\hat{y} = \boldsymbol{w_0} + \boldsymbol{w_1 f_1}$
- $\sum_{i=1}^{l} \boldsymbol{e_i^2} = \sum_{i=1}^{l} (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\boldsymbol{2}} = \sum_{i=1}^{l} \left(\boldsymbol{y_i} - \boldsymbol{w_0} - \boldsymbol{w_1 f_1}\right)^{\boldsymbol{2}} \to \min$ - прекрасная оптимизационная задача, которую мы можем решить

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ