# Машинное обучение 1
# 2017

Евгений Елтышев

# Основные понятия

- Объект (sample)
- Признак (feature)
- Метка (label, ground truth)
- Выборка (dataset)

# Основные понятия

- Обучение с учителем (Supervised learning)
- Обучение без учителя (Unsupervised learning)

- Классификация ( labels.nunique() < ∞)
- Регрессия ($labels \subset \mathbb{R}$)
- Кластеризация

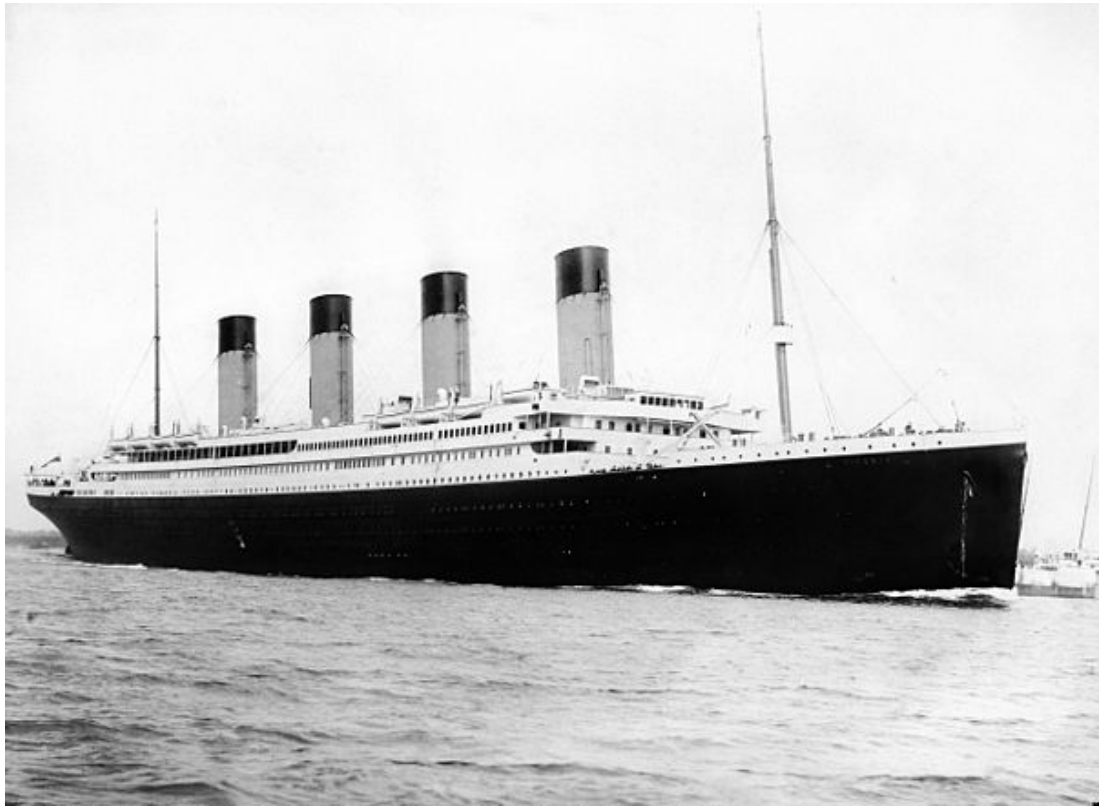
# Признаки

- Категориальные
  - Марка автомобиля
  - Регион пользователя
- Вещественные
  - Пробег автомобиля
  - Возраст пользователя

# Задача: Титаник

Задача: предсказать выжил ли человек?

| | |
|---|---|
| Погибших | 1496[1] |
| Пострадавших | 712 |



Какие признаки?

# Оценка качества

# Оценка качества

- Обучение с учителем
  - Сравниваем наши ответы с истинными метками
- Обучение без учителя
  - Менее формальные метрики, например минимальное расстояние между кластерами

# Оценка качества классификации

- Accuracy
- Logloss

# Оценка качества регрессии

- Mean squared error

# Окей, а на каких данных измерять качество?

# Окей, а на каких данных измерять качество?

- На обучающей выборке
  - Переобучение!
- На отложенной выборке
  - Уже лучше, но можем переобучиться под нее
- K-fold кросс-валидация
  - Совсем хорошо, но долго

# Линейные модели

- Регрессия: $\hat{y} = \vec{w} \cdot \vec{x}$
- Классификация: $\hat{y} = sign(\vec{w} \cdot \vec{x})$

- Легко интерпретировать
- Легко обучать
- Требуется мало данных

- Не могут моделировать сложные зависимости

# Метод ближайших соседей (kNN)

- $\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_{(i)}$

- Просто устроен
- Не требует долгого обучения
- При разных k работает «по-разному»

- Плохо работает в больших размерностях