# Машинное обучение

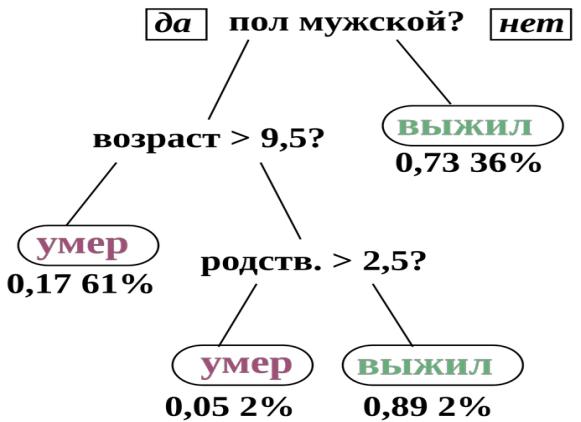
Лекция 2. Признаки

# Модальность данных

- Таблицы
- *Текст*
- *Изображения*
- *Видео*
- *Аудио*
- *Графы*

# Простые модели

- **Линейная регрессия** $y = Xw$
- **Метод ближайших соседей (knn)** – выбираем наиболее распространённый класс среди k ближаших соседей данного элемента, классы которых уже известны.
- **Дерево решений**



# Мотивация

Хотим классифицировать изображения.

- Как подавать изображение на вход модели?
- Сработают ли эти модели?

Глубокое обучение

# Задача предсказания оценки студента

- Пол
- Дата рождения
- Школа (город и номер)
- Средний школьный балл
- Эссе или код
- Социальный граф
- Расстояние от дома до универа
- Пиво/неделя
- Наличие ноутбука
- Ряд в аудитории
- Доля пропущенных лекций
- Периметр головы
- Оценка по мнению родителей
- Любимая книга

# Признаки

- Факторы, признаки, features, attributes, etc
- Для разных задач важны разные признаки
- Признаки необходимо преобразовать в $\mathbb{R}^n$

# Числовые признаки

Из $\mathbb{R}^n$

- Средний школьный балл
- Расстояние от дома до универа
- Пиво/неделя
- Периметр головы
- Доля пропущенных лекций

# Категориальные признаки

- Номинальные
  - Наличие ноутбука
  - Пол
- Порядковые
  - Ряд в аудитории
  - Оценка по мнению родителей

# Кодирование номинальных признаков

- Label encoding
- One-hot encoding
- Frequency Encoding
- etc

# Временные признаки

Дата рождения (знак зодиака, возраст).

- Периодические – день недели, месяц, год, etc.
- Разность между моментами времени (также до или после события)

# Географические признаки

Школа город:

- Местный/неместный.
- Регион.
- Расстояние до СПб.
- Большой/маленький город.

# На подумать

Как закодировать следующие признаки:

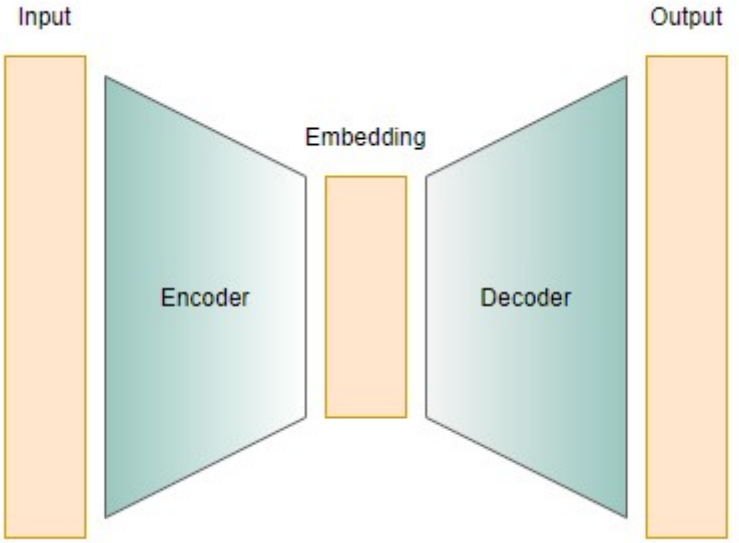
- Школа (номер)
- Любимая книга

# Кодирование текста или графов

- Пусть для каждого студента есть вступительное эссе и информация из социальных сетей. Можем ли мы как-нибудь её использовать?

# Эмбеддинги

## Преобразование объектов в вектора



# Предобработка признаков

- Постоянные признаки
- Пропущенные значения
- Выбросы
- Масштабирование

# Масштабирование

- Стандартизация, нормализация
- Модели на деревьях не чувствительны к масштабу признаков
- Линейные модели – регуляризация, сходимость методов оптимизации

# Вопросы для анализа

- Мало данных или много факторов?
  - Все ли факторы одинаково хороши?
  - Может их можно скомбинировать?
  - Стоит ли одинаково верить всем факторам?
- Может быть в данных что-то нечисто?
  - Все ли мы можем объяснить?
  - А набирали данные правильно?
  - Не подсматриваем ли мы в ответ?
  - Все ли важные примеры представлены в данных и репрезентативно ли это представление?

# Не все факторы одинаково полезны

- Можем ли мы обойтись без какого-нибудь фактора?
- А если фактор преобразовать, может его станет проще использовать?
- Если есть похожие факторы, наверное это можно учесть.
- Стот ли рассмотреть комбинации нескольких факторов?
- Что мы делаем, если фактор посчитать нельзя?