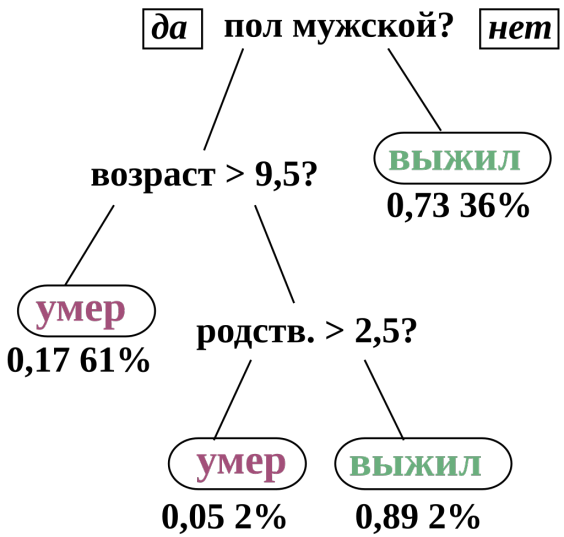
Лекция 2. Признаки

Модальность данных

- Таблицы
- Текст
- Изображения
- Видео
- Аудио
- etc

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

Что подавать на вход модели?

Хотим классифицировать изображения.

1. Как подавать изображение на вход модели?
2. Сработают ли эти модели?

Глубокое и неглубокое обучение

- **Классическое обучение** — ручное проектирование признаков. Модель ищет закономерности в заранее заданных признаках.
- **Глубокое обучение** — автоматическое создание признаков.

Задача предсказания оценки студента

- Пол
- Дата рождения
- Школа (город и номер)
- Средний школьный балл
- Мотивационное письмо
- Ссылка на github
- Профили в социальных сетях
- Расстояние от дома до университета
- Пиво/неделя
- Наличие ноутбука
- Ряд в аудитории
- Доля пропущенных лекций
- Периметр головы
- Оценка по мнению бабушки
- Оценка по мнению одногруппников
- Любимая книга

Признаки

Признаки (факторы, features, attributes, etc):

- необходимо преобразовывать в $\mathbb{R}^n$;
- для разных задач важны разные признаки.

Задачи на признаки

- Извлечение признаков
- Конструирование признаков
- Подготовка признаков (масштабирование и нормализация, кодирование категориальных признаков и т.п.)
- Отбор признаков (Feature Selection)
- Анализ важности признаков (Feature Importance)

Числовые признаки

Непосредственно из $\mathbb{R}^n$

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

Категориальные признаки

Номинальные

- Пол
- Наличие ноутбука

Порядковые

- Ряд в аудитории
- Оценка по мнению бабушки
- Оценка по мнению одногруппников

Кодирование номинальных признаков

- Label encoding
- One-hot encoding
- Frequency encoding
- Target encoding
- etc

Временные признаки

Дата рождения (знак гороскопа, возраст).

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

Географические признаки

Школа (город):

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

На подумать

Как закодировать следующие признаки:

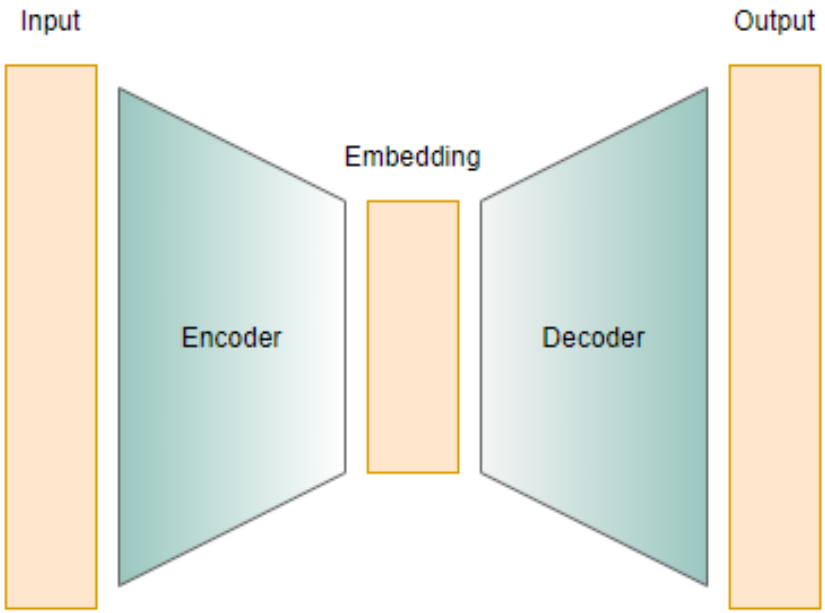
- Школа (номер)
- Любимая книга

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
- Конструирование новых признаков
- 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?

Ещё вопросы

- А если фактор преобразовать, может его станет проще использовать?
- Если есть похожие факторы, наверное это можно учесть?
- Стот ли рассмотреть комбинации нескольких факторов?
- Что делать, если фактор посчитать нельзя?

Важность признаков

- Какие признаки наиболее влияют на качество?
- Можно ли удалить некоторые признаки без потери качества?
- Почему модель приняла такое решение?
- Соответствуют ли важные признаки предметной логике?

Для разных моделей могут быть важны разные признаки!