Числовые признаки

Возраст, доход, широта, год и т.д.

- уже готовы к использованию
- но можно улучшить!

Изобретение признаков

Пример:

- Нужно предсказать, потратит ли пользователь больше 10,000 рублей в интернет-магазине в следующем месяце
- Ответ существенно зависит от суммы трат в прошлом месяце
- Среди признаков есть количество покупок c_i в прошлом месяце для каждого из 50,000 товаров, а также их цены p_i
- Можно добавить суммарные траты как признак, повысив качество: $\sum_{i=1}^{50000} p_i c_i$

[картинка: 50к одних чисел, 50к других, перемножаем, суммируем]

Масштабирование признаков

- Разные признаки могут иметь разный масштаб
- Пример: годовой доход в рублях и количество детей
- При измерении евклидова расстояние количество детей будет слабо влиять на результат
- Клиент x_1: 100,000 рублей, 0 детей
- Клиент x_2: 120,000 рублей, 5 детей
- Новый клиент z: 105,000 рублей, 4 детей
- Расстояния: $\rho(z, x_1) \approx 5000$, $\rho(z, x_2) \approx 15000$
- Нужно приводить признаки к одному масштабу!

Важно для метрических, линейных, нейросетевых моделей.
Не имеет значения для логических методов.

Масштабирование признаков

Способы масштабирования:

- на среднее и дисперсию

$$\tilde{x}^j = \frac{x^j - \mu}{\sigma}$$

- на отрезок $[0, 1]$:

$$\tilde{x}^j = \frac{x^j - \min(x^j)}{\max(x^j) - \min(x^j)}$$

Трансформация признаков

Признаки могут иметь тяжелые хвосты:
[картинка с тяжелым хвостом; по оси X количество звонков пользователя в колл-центр; по оси Y число пользователей с таким количеством звонков]

Наблюдения:

- Качество выше, если распределение признаков близко к нормальному
- Простая эвристика для неотрицательных признаков: логарифмирование

$$\tilde{x}^j = \log(x^j + 1)$$

Заключение

- Придумывать новые признаки — это искусство
- Признаки нужно масштабировать
- Может быть полезно приводить распределение признака к нормальному