Категориальные признаки

Признаки, на которых нельзя ввести порядок: тэги, города, цвета и т.д.

Dummy-кодирование:

- Признак x^j принимает значения из множества $U = \{u_1, \ldots, u_m\}$
- Создадим m новых признаков-индикаторов $x^{j1}, \ldots, x^{jm}$:

$$x^{jk} = [x^j = u_k]$$

Пример:

- $U = \{\text{Москва}, \text{Санкт-Петербург}, \text{Екатеринбург}\}$
- Кодируем тремя бинарными признаками
- Москва $\to (1, 0, 0)$
- Санкт-Петербург $\to (0, 1, 0)$
- Екатеринбург $\to (0, 0, 1)$

Проблемы dummy-кодирования

Уникальные категории:

- Что, если Екатеринбург встречается в выборке лишь один раз?
- Один из кодирующих признаков примет значение 1 лишь на одном объекте
- Такой признак не имеет смысла

Решение:

- Объединить редкие категории в одну
- Категория u редкая, если $\sum_{i=1}^{\ell}[x_i^j = u] <= r$
- r — параметр

Проблемы dummy-кодирования

Задача: предсказать, кликнет ли пользователь по рекламному баннеру

Признаки:

- Идентификатор пользователя
- Идентификатор баннера
- Идентификатор сайта, на котором показан баннер
- Идентификатор категории баннера

При dummy-кодировании мы получим миллионы признаков!

Счетчики

Идея:

- Пусть на баннер u_1 в среднем кликают чаще, чем на баннер u_2
- Это важный признак!
- Заменим категории на вероятности кликов

Счетчики:

- Задача классификации, $Y = \{0, 1\}$
- Оценим вероятность первого класса при условии значения признака:

$$c(u_k) = p(y = 1 \,|\, x^j = u_k) = \frac{\sum_{i=1}^{\ell} [x_i^j = u_k][y_i = 1]}{\sum_{i=1}^{\ell} [x_i^j = u_k]}$$

- Заменяем категориальный признак x^j на числовой $\tilde{x}^j$:

$$\tilde{x}_i^j = c(x_i^j)$$

- Для борьбы с переобучением можно вычислять счетчики с помощью кросс-валидации
  - выборка разбивается на k частей
  - для i-й части используются оценки вероятностей, полученные по остальным частям
  - для контрольной выборки используются оценки, полученные по всей обучающей выборке

Пример:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
Город	Мск	Екб	Мск	СПб	СПб	Мск	Екб
y	1	1	0	0	0	1	1

Оценки вероятностей:

$$p(y = 1 \,|\, \text{Мск}) = 2/3 = 0.67$$
$$p(y = 1 \,|\, \text{СПб}) = 0/2 = 0$$
$$p(y = 1 \,|\, \text{Екб}) = 2/2 = 1$$

Новая выборка:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
Город	0.67	1	0.67	0	0	0.67	1
y	1	1	0	0	0	1	1

Текстовые признаки

Значение признака x^j — последовательность слов $(w_1, w_2, \dots)$.

Мешок слов:

- Выводы о тексте можно даже по перемешанным словам
- Слова из текста принадлежат словарю $W = \{w_1, \dots, w_m\}$
- Создадим m новых признаков-индикаторов $x^{j1}, \dots, x^{jm}$:

$$x^{jk} = n_{w_k},$$

  n_{w_k} — число вхождения слова w_k в документ
- Аналог dummy-кодирования, но теперь несколько признаков могут быть больше нулю

Пример:

- $U = \{\text{ночь}, \text{улица}, \text{фонарь}, \text{аптека}\}$
- Кодируем четырьмя бинарными признаками
- ночь улица аптека ночь $\to (2, 1, 0, 1)$
- улица фонарь $\to (0, 1, 1, 0)$

Идея: вычислять не количество вхождений слов, а оценки их важности для текста

- чем чаще слово встречается в документе, тем оно важнее
- чем реже слово встречается в остальных документах, тем оно важнее

n_{iw} (term frequency) — число вхождений слова w в текст x_i^j;
N_w (document frequency) — число текстов, содержащих w;

Важность слова w для документа x_i^j:

$$\text{TF-IDF}(i,w) = \underbrace{n_{dw}}_{\text{TF}(i,d)} \underbrace{\log(\ell/N_w)}_{\text{IDF}(w)}.$$

$\text{TF}(i,d) = n_{iw}$ — term frequency;
$\text{IDF}(w) = \log(\ell/N_w)$ — inverted document frequency.

N-граммы

Иногда важны не только слова, но и словосочетания:

- "рекомендую" и "не рекомендую"
- "разработчик" и "старший разработчик"

N-граммы:

- Добавим в словарь W все возможные пары слов
- Добавим признаки-индикаторы для пар слов:

$$x_i^{jks} = [(w_k, w_s) \in x_i^j]$$

- Многие пары ни разу не встречаются — выбросим их

Примеры биграмм:
"ночь улица фонарь аптека" $\rightarrow$ (ночь, улица), (улица, фонарь), (фонарь, аптека)

Заключение

- Для категориальных и текстовых признаков можно делать dummy-кодирование
- Для категориальных признаков могут быть полезны счетчики
- Для текстовых признаков можно вычислять не количество слов, а TF-IDF — меру важности