Лекция 3
Методы обработки данных

Обработка пропусков в данных

Пропуски в данных

x_1	x_2	x_3	x_4	x_5	x_6	y
	?					
		?		?		
		?				
			?			

Источники пропусков

- Показания датчиков в промышленности
  - Сбои и повреждения датчиков
- Медицинская диагностика
  - Разное оборудование в разных поликлиниках
  - Нежелание пациента или невозможность проходить определенные обследования
- Социальные опросы
  - Отказ отвечать на определенные вопросы
- Статистика использования приложений
  - Программные ошибки
  - Сбои передачи статистики

Модели возникновения пропусков

Тезис: пропуски возникают случайным образом.

Модели возникновения пропусков

Тезис: пропуски возникают случайным образом.

- **Совершенно случайные пропуски**
  (missing completely at random, MCAR)
  Возникновение пропусков не связано с данными.
- **Случайные пропуски**
  (missing at random, MAR)
  Возникновение пропусков связано только с присутствующими данными.
- **Неслучайные пропуски**
  (not missing at random, NMAR)
  Возникновение пропусков связано и с присутствующими и отсутствующими данными.

Проблемы NMAR

70 000
45 000
?
250 000
?
?
60 000
?
55 000
?

Проблемы NMAR

70 000
45 000
?
250 000
?
?
60 000
?
55 000
?

70 000
45 000
270 000
250 000
40 000
180 000
60 000
150 000
55 000
200 000

Используемые модели

NMAR

- Утеряна важная информация
- Невозможно отличить от MAR и MCAR

Используемые модели

NMAR

- Утеряна важная информация
- Невозможно отличить от MAR и MCAR

MCAR и MAR

- Модель возникновения пропусков можно не учитывать
- Чаще всего используемые модели

Обозначения

$X = [x_1, \ldots, x_N]$ — объекты
x_{ij} — j-й признак i-го объекта
$j = 1, \ldots, d$

$Y = [y_1, \ldots, y_N]$ — целевая переменная

D_j — множество значений j-го признака
(например, $\mathbb{R}$, $\{0, 1\}$)

$x_{ij} \in D_j$ — нет пропуска
$x_{ij} =$ «?» — пропуск

Удаление пропусков

- Удаление объектов
- Удаление признаков

Преимущества

- Легко выполняется
- Можно применять все обычные методы

Недостатки

- Работает только при малой доле пропусков
- Плохо работает если пропуски коррелируют с ответом
- Нельзя использовать при пропусках в тестах

Заполнение средним

Для каждого признака j

- Посчитать среднее:

$$m_j = \frac{1}{|A_j|} \sum_{A_j} x_{ij}, \quad A_j = \{i \mid x_{ij} \neq \text{«?»}\}$$

- Заполнить пропуски этим значением

$$x_{ij} = \text{«?»} \quad \Rightarrow \quad x_{ij} := m_j$$

Заполнение средним

Для каждого признака j

- Посчитать среднее:

$$m_j = \frac{1}{|A_j|} \sum_{A_j} x_{ij}, \quad A_j = \{i \mid x_{ij} \neq \text{«?»}\}$$

- Заполнить пропуски этим значением

$$x_{ij} = \text{«?»} \quad \Rightarrow \quad x_{ij} := m_j$$

Недостатки

- Уменьшает вариацию в данных
- Не учитывает корреляцию между признаками

Заполнение средним по классам

(только для задачи классификации)

Для каждой пары (признак j, класс c)

- Посчитать среднее:

$$m_{jc} = \frac{1}{|A_{jc}|} \sum_{A_{jc}} x_{ij}, \quad A_{jc} = \{i \mid x_{ij} \neq \text{«?»}, y_i = c\}$$

- Заполнить пропуски этим значением (для данного класса c)

$$x_{ij} = \text{«?»} \quad \Rightarrow \quad x_{ij} := m_{jc}$$

Недостатки

- Уменьшает вариацию в данных
- Не учитывает корреляцию между признаками

Заполнение регрессией

Пусть пропуски есть только в d-ом признаке

- Новая задача
  X' — датасет без признака d
- Обучим регрессию на (X'_l, Y')
  $X'_l = [x_i \mid x_{id} \neq \text{«?»}]$
  $Y' = [x_{id} \mid x_{id} \neq \text{«?»}]$
- Применим ее к X'_t
  $X'_t = [x_i \mid x_{id} = \text{«?»}]$

Заполнение регрессией

Пусть пропуски есть только в d-ом признаке

- Новая задача
  X' — датасет без признака d
- Обучим регрессию на (X'_l, Y')
  $X'_l = [x_i \mid x_{id} \neq \text{«?»}]$
  $Y' = [x_{id} \mid x_{id} \neq \text{«?»}]$
- Применим ее к X'_t
  $X'_t = [x_i \mid x_{id} = \text{«?»}]$

Недостатки

- Уменьшает вариацию в данных (слабее)
- Добавляет лишнюю корреляцию между признаками

Заполнение ближайшим объектом

Пусть x_k — объект с пропуском в признаке l, $x_{kl} = \text{«?»}$

- Найдем ближайший объект без пропусков

$$C_l = \{i \mid x_{i1} \neq \text{«?»}, \ldots, x_{id} \neq \text{«?»}\}$$

$$\rho_l(x_s, x_t) = \sum_{j \neq l} (x_{sj} - x_{tj})^2$$

$$x_k^* = \arg \min_{x_i \in C_l} \rho_l(x_k, x_i)$$

- Используем его для заполнения

$$x_{kl} := x_{kl}^*$$

Заполнение ближайшим объектом

Пусть x_k — объект с пропуском в признаке l, $x_{kl} = \text{«?»}$

- Найдем ближайший объект без пропусков

$$C_l = \{i \mid x_{i1} \neq \text{«?»}, \ldots, x_{id} \neq \text{«?»}\}$$

$$\rho_l(x_s, x_t) = \sum_{j \neq l} (x_{sj} - x_{tj})^2$$

$$x_k^* = \arg\min_{x_i \in C_l} \rho_l(x_k, x_i)$$

- Используем его для заполнения

$$x_{kl} := x_{kl}^*$$

Недостатки

- Не учитываются глобальные свойства датасета

Заполнение с помощью KNN

Пусть x_i — объект с пропуском в признаке j, $x_{ij} = \text{«?»}$

- Найдем k ближайших объектов

$$\{v_1, v_2, \ldots, v_k\}$$

- Используем их для заполнения

$$x_{ij} := \frac{1}{k}(v_{1j} + v_{2j} + \ldots + v_{kj})$$

  - Среднее для числовых признаков
  - Мода для категориальных признаков
  - Возможны взвешенные варианты

Заполнение с помощью KNN

Пусть x_i — объект с пропуском в признаке j, $x_{ij} = \text{«?»}$

- Найдем k ближайших объектов

$$\{v_1, v_2, \ldots, v_k\}$$

- Используем их для заполнения

$$x_{ij} := \frac{1}{k}(v_{1j} + v_{2j} + \ldots + v_{kj})$$

  - Среднее для числовых признаков
  - Мода для категориальных признаков
  - Возможны взвешенные варианты

Недостатки

- Высокая вычислительная сложность

Расстояние между объектами с пропусками

(heterogeneous euclidean overlap metric, HEOM)

$$D(x_a, x_b) = \sqrt{\sum_{i=1}^{N} d_j(x_{aj}, x_{bj})^2}$$

- Числовые признаки

$$d_j(x_{aj}, x_{bj}) = \begin{cases} 1, & \text{если } x_{aj} = \text{«?»} \text{ или } x_{bj} = \text{«?»}, \\ \frac{|x_{aj} - x_{bj}|}{M_j - m_j}, & \text{иначе} \end{cases}$$

M_j — максимум, m_j — минимум j-го атрибута.

- Категориальные признаки

$$d_j(x_{aj}, x_{bj}) = \begin{cases} 1, & \text{если } x_{aj} = \text{«?»} \text{ или } x_{bj} = \text{«?»}, \\ [x_{aj} \neq x_{bj}], & \text{иначе} \end{cases}$$

Заполнение с помощью модели плотности

Распределение на возможных данных

$$x_i \sim p(x \mid \theta)$$

θ — неизвестные параметры модели

Заполнение с помощью модели плотности

Распределение на возможных данных

$$x_i \sim p(x \mid \theta)$$

θ — неизвестные параметры модели

$X_m = \{x_{ij} \mid x_{ij} = \text{«?»}\}$ — пропуски
(тоже должны подчиняться распределению)

Заполнение с помощью модели плотности

Распределение на возможных данных

$$x_i \sim p(x \mid \theta)$$

θ — неизвестные параметры модели

$X_m = \{x_{ij} \mid x_{ij} = \text{«?»}\}$ — пропуски
(тоже должны подчиняться распределению)

Восстановление данных и параметров

- Инициализация
  $X_m :=$ случайные значения (или среднее и т.п.)
- Повторять до сходимости
  - Оценить $p(x \mid \theta)$ по X
  - Восстановить X_m по $p(x \mid \theta)$

Категориальные признаки

Категориальные признаки

Признаки, которые нельзя интерпретировать как числа.

- {красный, зеленый, синий}
- {Москва, Новосибирск, Владивосток, . . . }
- Тексты

Допустимы	**Недопустимы**
Решающие деревья	Линейные модели
Метрические методы*	*Нейронные сети*
Байесовский классификатор	

* нужно правильно выбрать метрику

One-hot encoding

Каждое значение кодируется бинарным признаком.

Цвет		**красный?**	**зеленый?**	**синий?**
красный		1	0	0
зеленый		0	1	0
красный	$\longrightarrow$	1	0	0
синий		0	0	1
зеленый		0	1	0

Недостатки

- Очень много признаков
- Признаки очень разреженные

Кодирование Bag-of-Words

- Признаки = число вхождений каждого слова
- Порядок не учитывается

Тексты

- John likes to watch movies
- Mary likes movies too
- John also likes football

Признаки

John	like	to	watch	movie	Mary	too	also	football
1	1	1	1	1	0	0	0	0
0	1	0	0	1	1	1	0	0
1	1	0	0	0	0	0	1	1

Кодирование в несколько признаков

Предметная область $\Rightarrow$ Новые признаки

Город		**На Западе**	**Миллионнер**
Москва	$\longrightarrow$	1	1
Новосибирск		0	1
Владивосток		0	0

Кодирование в несколько признаков

Предметная область ⇒ Новые признаки

Город		**На Западе**	**Миллионнер**
Москва	⟶	1	1
Новосибирск		0	1
Владивосток		0	0

Преимущества

- Возможно, дополнительная информация

Недостатки

- Нужны знания в предметной области

Хеширование признаков

Задача

- Кодирование в несколько признаков
- Без знаний предметной области

Хеширование признаков

Задача

- Кодирование в несколько признаков
- Без знаний предметной области

Идея:

- One-hot encoding (Bag-of-Words)
- Случайная группировка признаков
- Сумма признаков внутри группы

1	**2**	**3**	**4**	**5**	**6**
0	1	0	0	1	0
1	1	1	0	0	1
1	0	1	0	0	0
0	0	1	1	1	1
1	0	0	1	1	0
0	1	1	0	1	0
0	1	1	0	0	0

1	2	3	4	5	6
0	1	0	0	1	0
1	1	1	0	0	1
1	0	1	0	0	0
0	0	1	1	1	1
1	0	0	1	1	0
0	1	1	0	1	0
0	1	1	0	0	0

1	2	3	4	5	6
0	1	0	0	1	0
1	1	1	0	0	1
1	0	1	0	0	0
0	0	1	1	1	1
1	0	0	1	1	0
0	1	1	0	1	0
0	1	1	0	0	0

245	136
2	0
1	3
0	2
2	2
2	1
1	2
1	1

Хеширование признаков

Свойство:
Похожесть векторов приблизительно сохраняется.

Эмпирическое свойство:
Достаточно хорошо работает.

Хеширование признаков

Свойство:
Похожесть векторов приблизительно сохраняется.

Эмпирическое свойство:
Достаточно хорошо работает.

Преимущества

- Хорошо сокращает размерность

Недостатки

- Что происходит?!

Хеширование признаков

Свойство:
Похожесть векторов приблизительно сохраняется.

Эмпирическое свойство:
Достаточно хорошо работает.

Преимущества

- Хорошо сокращает размерность

Недостатки

- Что происходит?!

См. также: Locality-sensitive hashing.

Счетчики: пример

Задача: предсказание вероятности клика на рекламу.

Признаки: пользователь, сайт, реклама.
Целевая переменная: вероятность клика.

- Мало признаков
- У каждого признака много значений

Счетчики: пример

Задача: предсказание вероятности клика на рекламу.

Признаки: пользователь, сайт, реклама.
Целевая переменная: вероятность клика.

- Мало признаков
- У каждого признака много значений

Идея:

- Для каждого значения признака найти вероятность
- Подставить эту вероятность вместо признака

Счетчики для отдельных признаков

Счетчики

User	**p**
Alice	2.3%
Bob	0.5%
...	...

Site	**p**
example.com	20%
qwerty.org	4%
...	...

Ad	**p**
examples	0.2%
keyboards	1.3%
...	...

Трансформированная выборка

p-User	**p-Site**	**p-Ad**	**p**
2.3%	20%	4%	0.015%
0.5%	15%	33%	0.022%
...	...	...	...

Счетчики для отдельных признаков

Счетчики для каждого признака

- Легко считать, мало признаков на выходе
- Расширенный наивный байесовский подход
- Не учитываются связи между признаками

Счетчики для отдельных признаков

Счетчики для каждого признака

- Легко считать, мало признаков на выходе
- Расширенный наивный байесовский подход
- Не учитываются связи между признаками

Идея:

- Для каждого значения признака найти вероятность
- Для каждого значения пары признаков найти вероятность (совместную)
- ...
- Подставить эти вероятности вместо всех признаков

Счетчики для комбинаций признаков

Счетчики

User-Site	**p**
Alice, example.com	0.11%
...	...
Bob, qwerty.org	0.03%
...	...

User-Ad	**p**
Alice, examples	0.02%
...	...
Bob, keyboards	0.47%
...	...

Трансформированная выборка

p-User	...	**p-User-Site**	...	**p-User-Site-Ad**	**p**
2.3%	...	0.11%	...	0.035%	0.015%
0.5%	...	0.03%	...	0.02%	0.022%
...	...	...	...	...	...

Счетчики для комбинаций признаков

Проблема: какие-то тестовые комбинации могли не встречаться при обучении.

Счетчики для комбинаций признаков

Проблема: какие-то тестовые комбинации могли не встречаться при обучении.

Идея: добавить еще признаки

- Признак $\langle$ была ли комбинация в обучении $\rangle$
- Число кликов, число показов

Итог

- Признаков и так немного
- Больше информации

Счетчики: общая схема

Рассматриваемая задача

- Мало категориальных признаков
- У каждого признака много значений

Счетчики

- Набор комбинаций признаков (м.б. не все)
- Для каждой подсчет статистик ответов

Новые признаки

- Статистики для разных комбинаций
- Доп. признаки

Источники

- **Пропуски:** García-Laencina et al. Pattern classification with missing data: a review (2010). [pdf]
- **Хеширование:** Attenberg J. et al. Collaborative spam filtering with the hashing trick (2009). [pdf]
- **Счетчики:** Microsoft Blog. Big Learning Made Easy with Counts. [link]
- Pyle, D. Data preparation for data mining.