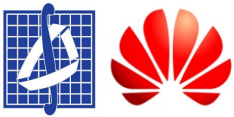
Введение в искусственный интеллект. Машинное обучение

Тема: Категориальные признаки. Пропущенные значения.

Бабин Д.Н., Иванов И.Е., Петюшко А.А.

кафедра Математической Теории Интеллектуальных Систем



План семинара

1. Категориальные признаки
2. Пропущенные значения

Примеры категориальных данных

- Пол
- Страна, город
- Образование
- Категория товаров
- Тарифный план
- Профессия
- ...
- Любой признак, который имеет небольшое количество значений, независимо от типа данных

Проблема категориальных данных

Проблема

Большинство алгоритмов машинного обученя не может работать с нечисловыми признаками

Проблема категориальных данных

Проблема

Большинство алгоритмов машинного обученя не может работать с нечисловыми признаками

Решение

Поэтому возникает необходимость закодировать все нечисловые признаки

Простейшее кодирование

Идея

Заменить категории на некоторые числовые значения, согласно заранее определенному соотвествию

Простейшее кодирование

Идея

Заменить категории на некоторые числовые значения, согласно заранее определенному соотвествию

Недостатки

- Непонятно, как делать это соотвествие
- Возникает отношение порядка, которое может никак не реализовываться в реальном мире

CompanyName	Categoricalvalue	Price
VW	1	20
Acura	2	10011
Honda	3	50000
Honda	3	10000

Dummy кодирование (one hot encoding)

Идея

Для каждой категории добавить бинарный признак

CompanyName	Categoricalvalue	Price
VW	1	20
Acura	2	10011
Honda	3	50000
Honda	3	10000

One hot encoding

VW	Acura	Honda	Price
1	0	0	20000
0	1	0	10011
0	0	1	50000
0	0	1	10000

Умные способы кодирования

- Замена категории на некоторое агрегированное значение
  - Если речь идет о категории товара, то средняя цена товара хорошо подойдет
  - Замена категории на количество объектов, входящих в нее

Причины пропусков в данных

- Данные теряются
- Данные хранятся в разных системах с различными интерфейсами
- Люди часто не заполняют необязательные поля
- Измерительные устройства выходят из строя

Классификация пропусков

В зависимости от причин появления пропусков возникают следующие типы пропусков:

Полностью случайные пропуски

Вероятность пропуска не зависит ни от наблюдаемых данных, ни от пропущенных

Классификация пропусков

В зависимости от причин появления пропусков возникают следующие типы пропусков:

Полностью случайные пропуски

Вероятность пропуска не зависит ни от наблюдаемых данных, ни от пропущенных

Пропуски зависят от наблюдаемых значений

Вероятность пропуска зависит от наблюдаемых данных, но не зависит от пропущенных значений

Классификация пропусков

В зависимости от причин появления пропусков возникают следующие типы пропусков:

Полностью случайные пропуски

Вероятность пропуска не зависит ни от наблюдаемых данных, ни от пропущенных

Пропуски зависят от наблюдаемых значений

Вероятность пропуска зависит от наблюдаемых данных, но не зависит от пропущенных значений

Пропуски не случайны

Вероятность пропуска зависит как от наблюдаемых данных, так от пропущенных

Методы обработки пропущенных данных: удаление данных

- Удаление признаков
- Удаление объектов

Методы обработки пропущенных данных: удаление данных

- Удаление признаков
- Удаление объектов

Замечание

Хорошо работает, когда данных достаточно и пропуски полностью случайные

Методы восстановления пропущенных данных

- Замена специальным значением

Методы восстановления пропущенных данных

- Замена специальным значением
- Замена средним
- Замена медианой
- Замена модой

Методы восстановления пропущенных данных

- Замена специальным значением
- Замена средним
- Замена медианой
- Замена модой

Замечание

При восстановлении данных рекомендуется добавлять бинарный признак, помечающий объекты, где было применено восстановление

Методы восстановления пропущенных данных: kNN

Идея

Поиск ближайших соседей по наблюдаемым данным и замена пропущенного значения на значения из похожих объектов

Методы восстановления пропущенных данных: кластеризация

Задача

$$V = \sum_{i=1}^{k} \sum_{x \in S_i} (x - \mu_i)^2 \to \min_{S_i},$$

где k — число кластеров, S_i — полученные кластеры, μ_i — центр масс S_i кластера.

Алгоритм

1. Случайно выбираются k элементов из выборки и объявляются центроидами
2. Для фиксированных центроидов каждый элемент выборки относится к одному из кластеров
3. Для фиксированных кластеров вычисляются центроиды
4. Пункты 2,3 повторяются до сходимости

Методы восстановления пропущенных данных: MICE

Идея MICE (multiple imputations by chained equations)

Последовательно обучать модели для восстановления данных

Методы восстановления пропущенных данных: матричные разложения

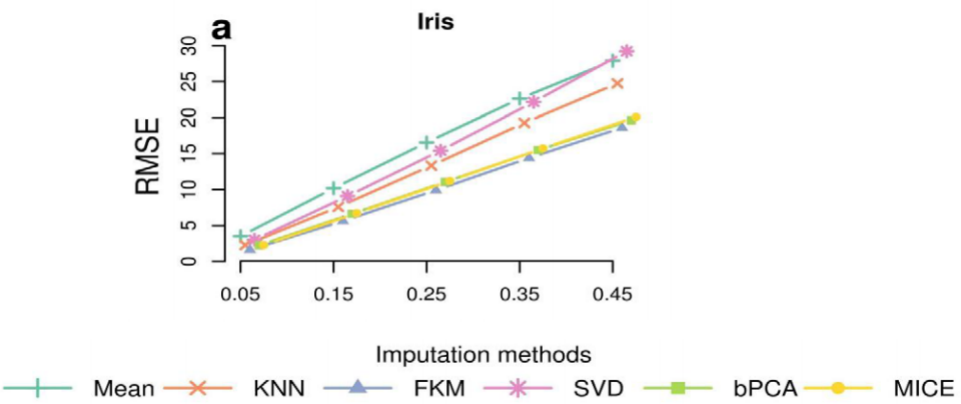
Идея

Многие матричные разложения работают и на матрицах с пропущенными данными

Примеры:

- bPCA - байесовский метод главных компонент
- SVD-разложение

Сравнение различных стратегий [1]



[1] https://lgreski.github.io/datasciencedepot/references/a-comparison-of-six-methods-for-missing-data-imputation-2155-6180-1000224.pdf

Заключение

- Большинство алгоритмов машинного обучения работают с числовыми признаками без пропущенных значений
- Стандартные методы преобразования категориальных признаков — это простое кодирование, dummy-кодирование и кодирование агрегированными значениями
- В зависимости от причин появления пропусков возникают три типа пропусков
- Метод восстановления пропущенных значений зависит от даннных
- Стандартная рекомендация — начинать с простых методов