Занятие № 6

Работа с пропусками

Содержание

1 Основные способы заполнения пропусков

2 Практика.

Проблема пропущенных значений

- Ошибочные пропуски в данных и во время использования с новыми данными пропусков не будет.

- Валидные пропуски в данных и во время использования будут приходить данные с пропусками.

Проблема пропущенных значений

- Пропущенные значения ведут к снижению статистической мощности (то есть снижают вероятность нахождения реальных закономерностей в данных), а также могут быть причиной систематических ошибок.

- За редким исключением алгоритмы машинного обучения не работают с выборками, имеющими пропущенные значения.

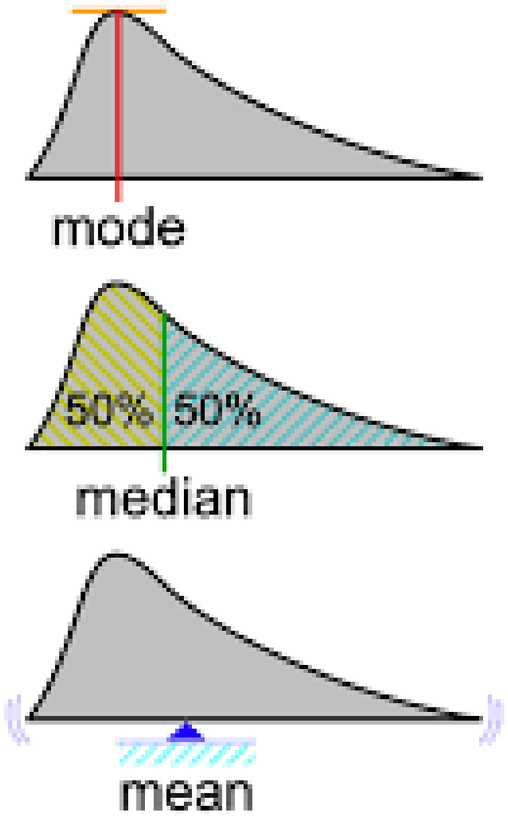
Обработка нулевых значений

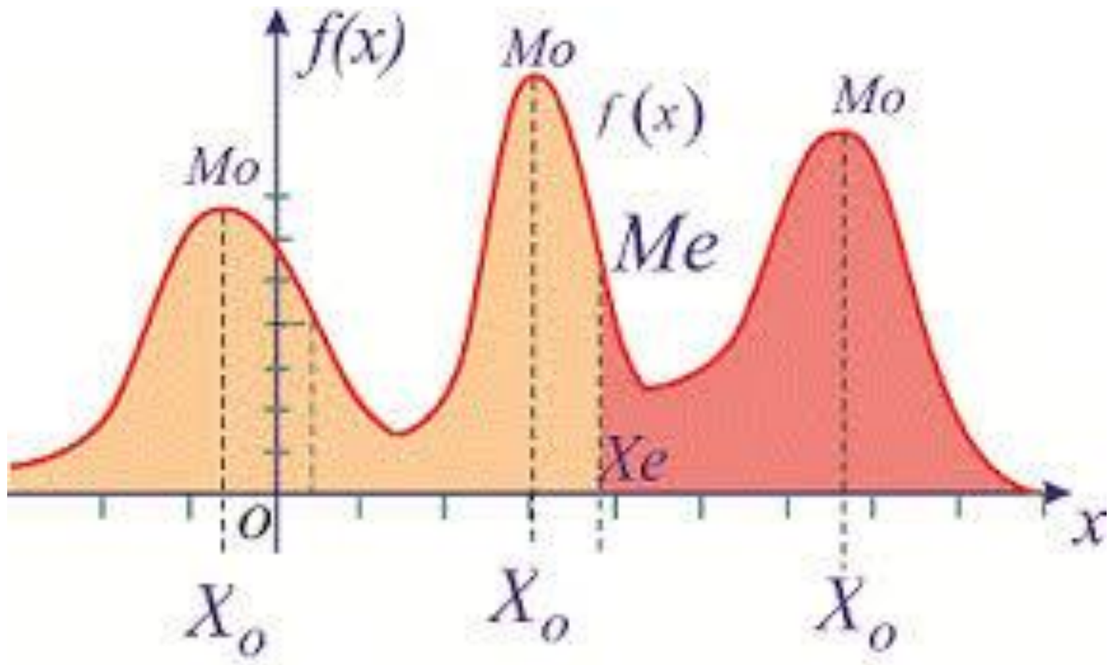
Удаление пропущенных значений:
- Удаление столбец содержащий нулевое значение (потеря информации)
- Удаление строки, в которых атрибут равен нулевому значению (потеря информации)

Подстановка значений:
- Статистический подход.
- Indicator Method
- Восстановление пропусков на основе моделей
- Интерполяция/моделирование (в случае последовательностей)

Обработка нулевых значений

Статистический подход. Заменять на среднее значение, медиану, моду и др.





Обработка нулевых значений

Статистический подход.

Плюсы:
- быстро
- просто
- не зависит от наличия пропусков в других признаках

Минусы:

При обучении
- не учитывает зависимости между данными
- не учитывает разделение на страты

При использовании
- дрейф в данных

Обработка нулевых значений

Indicator Method - замена пропущенных значений нулями и создание новой переменной индикатора (где она принимает значение 1 при наличие пропуска и 0 в остальных случаях)

...	**5**	...
...	None	...
...	7	...

...	**5**	**0**	...
...	0	1	...
...	7	0	...

Обработка нулевых значений

Indicator Method

Плюсы:
- быстро
- просто
- не зависит от наличия пропусков в других признаках
- модель сама подбирает значение

Минусы:

При обучении
- не учитывает зависимости между данными
- не учитывает разделение на страты
- дополнительные параметры в модель

При использовании
- дрейф в данных

Обработка нулевых значений

Восстановление пропусков на основе моделей

Обучаемые модели

- Линейная регрессия
- Логистическая регрессия
- Деревья решений (Случайный лес и тд)
- kNN – метода ближайшего соседа
- ...

Итерационные алгоритмы

- SVD
- EM-алгоритм
- Итерационное применение обучаемых моделей

...	**5**	...
...	None	...
...	7	...

...	**5**	...
...	5	...
...	7	...

Обработка нулевых значений

Восстановление пропусков на основе моделей

Обучаемые модели

Плюсы:
- учитываются зависимости между данными и страты
- достаточно быстро если признаков с пропущенными значениями очень мало

Минусы:

При обучении
- для заполнения одних признаков нужно, что бы были заполнены пропущенные значения в признаках, на основе которых строится модель
- при большом количестве признаков с пропущенными значениями нужно найти правильную последовательность заполнения
- увеличивается время обучения

При использовании
- дрейф в данных

Обработка нулевых значений

Восстановление пропусков на основе моделей

Итерационные алгоритмы

Плюсы:
- учитываются зависимости между данными и страты
- можно получить хорошие значения
- можно заполнять пропущенные значения одновременно несколько признаков

Минусы:
При обучении
- сильно увеличивается время обучения
- должно сходиться

При использовании
- дрейф в данных

Обработка нулевых значений

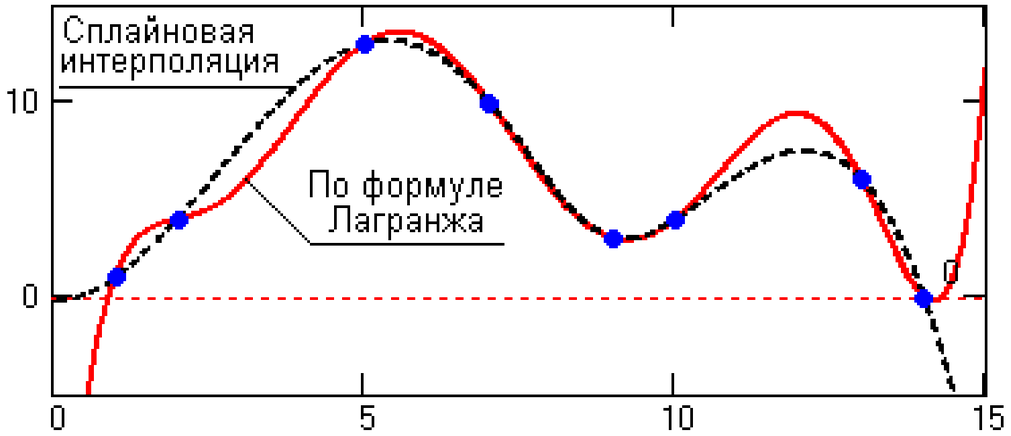
Восстановление пропусков временных рядов

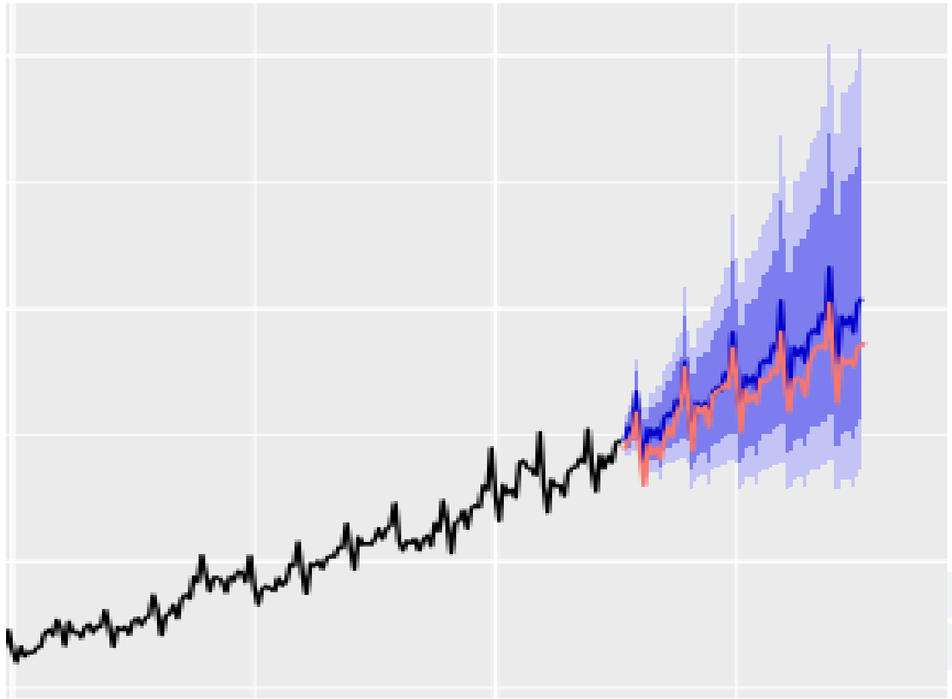
Аналитический

a) Аппроксимация ряда в окрестности пропуска аналитической функцией
b) интерполяция пропущенного значения

Модельный

a) Описание ряда с помощью модели
b) Предсказание пропущенного ряда полученной моделью





ПРАКТИКА

Спасибо за внимание!