Занятие № 6

Работа с пропусками

Содержание

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.

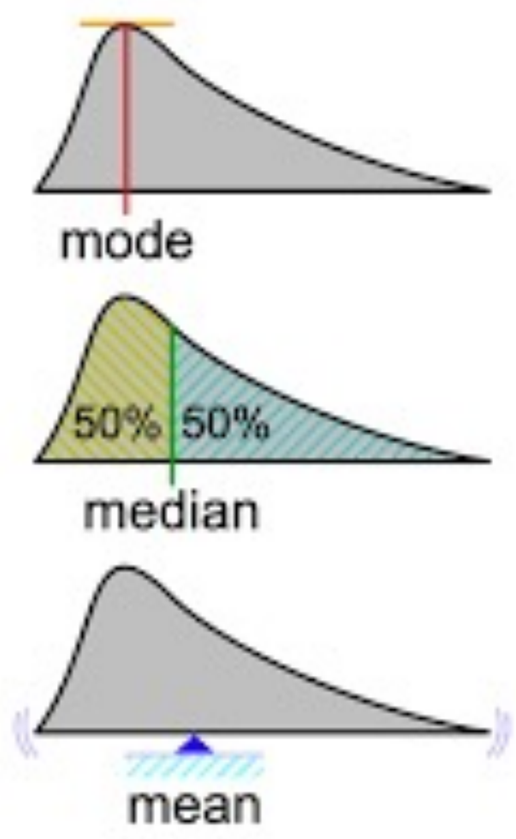
Обработка нулевых значений

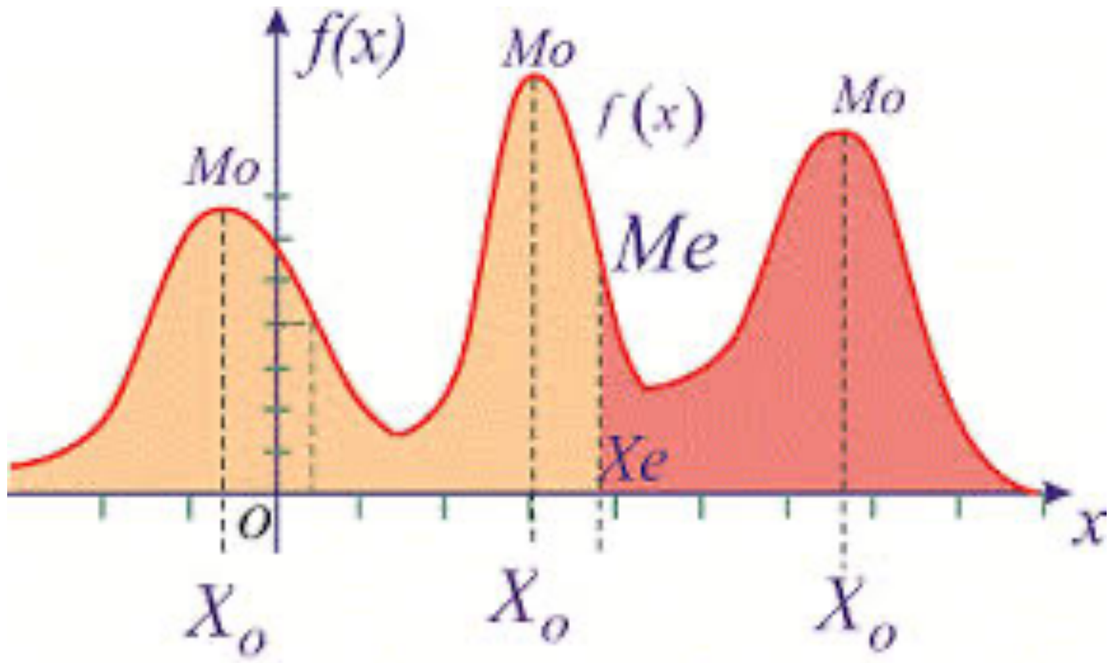
Удаление пропущенных значений:
- Удаление столбец содержащий нулевое значение (потеря информации)
- Удаление строки, в которых атрибут равен нулевому значению (потеря информации)

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

Обработка нулевых значений

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





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

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-алгорим
- Итерационное применение обучаемых моделей

ПРАКТИКА

Спасибо за внимание!