Обработка пропущенных значений

Аббакумов Вадим

04.2021

Что лучше, чем заполнение средними?

Пропуски во временных рядах — там все иначе?

Что делать, если 99% наблюдений — пропуски?
Рекомендательные системы.

Наука об обработке пропущенных значений сформировалась к 1980 году.

Но первые 30 лет новые модели плохо работали на практике…

Последние 5 лет методы превратились в работающие процедуры, в технологию.

Субъективный обзор и рекомендации

Заказ был про Python, но

Ситуация в R

https://cran.r-project.org/web/views/MissingData.html

165+ пакетов только в репозитории CRAN

Нерешенная проблема:
критерий качества заполнения пропусков

Казалось бы все просто,
надо сравнить все методы

Сравнение

Взять все популярные наборы данных.

Удалить,
заполнить пропуски,
сравнить пропуски и удаленные значения.

Какой метод заполнения дает меньшие ошибки,
тот и лучше.

Увы
Заполнение пропусков не самоцель

Решаем другую задачу.

Лучше тот метод заполнения, после которого основная модель работает лучше.

Заполнение пропусков — гиперпараметр!

1 Удаление строк / Удаление столбцов

Неверно: удаление только строк

Комбинация подходов -удаляем «правильные» строки и столбцы

Возможно смещение выборки

2 Заполнение пропущенных значений (imputation)

2.1 Заполнение средними или медианами (по столбцам!)

Неверно:

заполняем только средними или только медианами.

Еще варианты

Заполнение значением, которое чаще всего встречается в столбце

Когда применяем?

Заполняем любым числом, которое понравится...

`sklearn.impute`.SimpleImputer

https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html#sklearn.impute.SimpleImputer

2.2 Заполнение средними/медианами по сегменту

«Средняя температура по больнице»

Неправильно одинаково заполнять данные
- про богатых и про бедных
- про мальчиков и про девочек
- про людей из городов миллионников и из населенных пунктов с населением до 10000 человек

Предварительная сегментация!

По какой переменной?
По всем разом?
А в других переменных тоже пропуски...

Итого:

кластеризуем,

затем заполняем средними в каждом кластере.

2.4 Заполнение методом к-го ближайшего соседа

Выбираем

максимально похожего и не имеющего пропусков,

заимствуем значение у него.

А может выбирать и соседа с пропусками...

А может 3-х максимально похожих...

sklearn.impute.KNNImputer

https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

Достоинство:
Лучше заполнения средними/медианами

Недостатки.

Надо определять число ближайших соседей.
Но можно считать это число гиперпараметром.

Только евклидово расстояние с коррекцией за присутствие пропусков.
Но можно реализовать свой способ подсчета расстояний.

Типичная ошибка при применении:
забыли стандартизовать данные.

2.5 mice: Multivariate Imputation by Chained Equations

С 2011 года, сейчас версия 3.

На C++

Авторитетный, популярный (?)

У меня плохо работал 8 лет назад,

не работает и сейчас

IterativeImputer
https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

2.6 Статистические методы

Много методов заполнения для случаев, когда верим в

Гипотезу о распределении пропусков.

Либо в

Гипотезу о многомерном нормальном распределении данных

2.6 MissForest

Появился в R, перенесен в Python

в Python активно продвигается разработчиком

Естественный

Работает не на любом наборе данных.
Готовьтесь к сюрпризам

3. Заполнение пропусков во временных рядах

4. Рекомендательные системы.

Что делать, если 99% наблюдений — пропуски?

SVD разложение по Funk’у

fancyimpute

https://pypi.org/project/fancyimpute/