ДИСЦИПЛИНА **Интеллектуальные системы и технологии**

(полное наименование дисциплины без сокращений)

ИНСТИТУТ **информационных технологий**

КАФЕДРА **корпоративных информационных систем**

полное наименование кафедры

ВИД УЧЕБНОГО МАТЕРИАЛА **Дополнительные материалы**

(в соответствии с пп.1-11)

ПРЕПОДАВАТЕЛЬ **Демидова Лилия Анатольевна**

(фамилия, имя, отчество)

СЕМЕСТР **1 семестр (осенний), 2024 – 2025 учебный год**

(семестр обучения, учебный год)

Обработка пропусков в данных

Часто в данных, с которыми необходимо работать, присутствуют пропуски, в результате чего аналитик оказывается перед выбором: игнорировать, отбросить или же заполнить пропущенные значения.

Заполнение пропусков зачастую, и вполне обоснованно, кажется более предпочтительным решением.

Однако это не всегда так. Неудачный выбор метода заполнения пропусков может не только не улучшить, но и сильно ухудшить результаты.

Исключение и игнорирование строк с пропущенными значениями стало решением по умолчанию в некоторых популярных прикладных пакетах, в результате чего у начинающих аналитиков может возникнуть представление, что данное решение – правильное.

Существуют довольно простые в реализации и использовании методы обработки пропусков, получившие название **«ad-hoc методы»**: заполнение пропусков нулями, медианой, средним арифметическим значением, введение индикаторных переменных и тому подобное, простота которых может послужить причиной для выбора именно этих методов.

«Ad-hoc» – латинская фраза, означающая «специально для этого», «по особому случаю». Как правило, фраза обозначает способ решения специфической проблемы или задачи, который невозможно приспособить для решения других задач и который не вписывается в общую стратегию решений, составляет некоторое исключение.

Вероятно, именно из-за своей простоты ad-hoc методы широко использовались на заре развития современной теории обработки пропусков.

И хотя по состоянию на сегодняшний день известно, что применение этих методов может приводить к искажению статистических свойств выборки и, как следствие, к ухудшению результатов, получаемых после такой обработки пропусков, их по-прежнему часто используют.

Так, известны статьи, посвященные сбору и оценке статистики использования методов заполнения пропусков в научных работах медицинской тематики, из результатов которых можно сделать вывод, что даже ученые часто отдают предпочтение интуитивно-понятным ad-hoc методам и игнорированию/удалению строк, несмотря на то, что применение этих методов в контексте решаемой задачи иногда неуместно.

Механизмы формирования пропусков

Для того чтобы понять, как правильно обработать пропуски, необходимо определить механизмы их формирования.

Различают 3 механизма формирования пропусков: MCAR, MAR, MNAR.

MCAR (Missing Completely At Random, совершенно случайный пропуск) – механизм формирования пропусков, при котором вероятность пропуска для каждой записи набора одинакова.

Например, если проводился социологический опрос, в котором каждому десятому респонденту один случайно выбранный вопрос не задавался, причем на все остальные заданные вопросы респонденты отвечали, то имеет место механизм MCAR. В таком случае игнорирование/исключение записей, содержащих пропущенные данные не ведет к искажению результатов.

MAR (Missing At Random, случайный пропуск) – на практике данные обычно пропущены не случайно, а ввиду некоторых закономерностей.

Пропуски относят к MAR, если вероятность пропуска может быть определена на основе другой имеющейся в наборе данных информации (пол, возраст, занимаемая должность, образование…), не содержащей пропуски. В таком случае удаление или замена пропусков на значение «Пропуск», как и в случае MCAR, не приведет к существенному искажению результатов.

MNAR (Missing Not At Random, неслучайный пропуск) – механизм формирования пропусков, при котором данные отсутствуют в зависимости от неизвестных факторов. MNAR предполагает, что вероятность пропуска могла бы быть описана на основе других атрибутов, но информация по этим атрибутам в наборе данных отсутствует. Как следствие, вероятность пропуска невозможно выразить на основе информации, содержащейся в наборе данных.

Рассмотрим различия между механизмами MAR и MNAR на примере.

Люди, занимающие руководящие должности и/или получившие образование в престижном вузе чаще, чем другие респонденты, не отвечают на вопрос о своих доходах. Поскольку занимаемая должность и образование сильно коррелируют с доходами, то в таком случае пропуски в поле «Доходы» уже нельзя считать совершенно случайными, то есть говорить о случае MCAR не представляется возможным.

Если в наборе данных есть информация об образовании и должности респондентов, то зависимость между повышенной вероятностью пропуска в графе доходов и этой информацией может быть выражена математически, следовательно, выполняется гипотеза MAR. В случае MAR исключение пропусков вполне приемлемо.

Однако если информация о занимаемой должности и образовании у нас отсутствует, то тогда имеет место случай MNAR.

При MNAR просто игнорировать или исключить пропуски уже нельзя, так как это приведет к значительному искажению распределения статистических свойств выборки.

Удаление/игнорирование пропусков

Complete-case Analysis (полный анализ случая) или **Listwise Deletion Method** (метод удаления по списку) – метод обработки пропусков, применяемый в прикладных пакетах как метод по умолчанию.

Заключается в исключении из набора данных записей/строк или атрибутов/колонок, содержащих пропуски.

В случае первого механизма пропусков (MCAR) применение данного метода не приведет к существенному искажению параметров модели. Однако удаление строк приводит к тому, что при дальнейших вычислениях используется не вся доступная информация, стандартные отклонения возрастают, полученные результаты становятся менее репрезентативными.

В случаях, когда пропусков в данных много, это становится ощутимой проблемой.

В случае второго механизма пропусков (MAR) и, особенно, третьего механизма пропусков (MNAR) смещение статистических свойств выборки, значений параметров построенных моделей и увеличение стандартных отклонений становятся еще сильнее.

Таким образом, несмотря на широкое распространение, применение данного метода для решения практических задач ограничено.

Available-case analysis (анализ доступных случаев) или **Pairwise Deletion** (попарное удаление) – методы обработки, основанные на игнорировании пропусков в расчетах. Эти методы, как и **Complete-case Analysis**, тоже часто применяются по умолчанию.

Статистические характеристики, такие как средние значения, стандартные отклонения, можно рассчитать, используя все непропущенные значения для каждого из атрибутов/столбцов.

Как и в случае **Complete-case Analysis**, при условии выполнения гипотезы MCAR, применение данного метода не приведет к существенному искажению параметров модели.

Преимущество данного подхода в том, что при построении модели используется вся доступная информация.

Главный же недостаток данных методов заключается в том, что они применимы для расчета далеко не всех показателей и, как правило, сопряжены с алгоритмическими и вычислительными сложностями, приводящие к некорректным результатам.

Например, рассчитанные значения **коэффициентов корреляции** могут оказаться вне диапазона [–1; 1]. Кроме того, не всегда удается однозначно ответить на вопрос об оптимальном выборе числа отсчетов, используемого при расчете стандартных отклонений.

Пример, демонстрирующий проблемы методов **Available-case analysis**

Рассмотрим следующую задачу: необходимо рассчитать линейный коэффициент корреляции (коэффициент корреляции Пирсона) между двумя факторами/переменными X и Y, истинные значения которых приведены в таблице 1.

Таблица 1 – Данные без пропусков

X	Y	X	Y	X	Y	X	Y
1	−0,1983	9	−7,9492	11	−12,1990	18	−14,9021
1	0,3767	6	−7,3456	10	−11,0234	17	−18,9796
5	−2,2939	7	−9,8047	11	−14,2284	19	−19,5486
4	−3,0886	7	−10,6003	14	−13,1966	20	−20,5562
5	−3,1942	14	−9,9677	16	−15,8271	21	−21,0093

На основе таблицы 1 определим истинные значения статистических параметров.

Среднее значение		**Ковариация**		
X	***Y***	**σ11**	**σ12**	**σ22**
10,8000	−10,7768	37,7600	-38,1691	42,6608

Оценка (ко)вариации:

$\sigma 11=\sum X_i^2/n-(\sum X_i/n)^2=37{,}7600.$

$\sigma 12=\sum(X_i\cdot Y_i)/n-(\sum X_i\cdot\sum Y_i)/n^2=-38{,}1691.$

$\sigma 22=\sum Y_i^2/n-(\sum Y_i/n)^2=42{,}6608.$

n – число наблюдений (n=20).

Значение коэффициента корреляции:

$r=\sigma 12/(\sigma 11\cdot\sigma 22)^{0.5}=-0{,}9510.$

Рассмотрим результаты аналогичных расчетов при наличии пропусков в данных.

Таблица 2 – Данные с пропусками

X	Y	X	Y	X	Y	X	Y
?	−0,1983	9	−7,9492	11	−12,1990	18	−14,9021
?	0,3767	6	−7,3456	10	−11,0234	17	−18,9796
5	−2,2939	7	−9,8047	11	−14,2284	19	−19,5486
4	−3,0886	7	−10,6003	14	−13,1966	20	−20,5562
5	−3,1942	14	−9,9677	16	−15,8271	21	−21,0093

Работаем с тем же набором данных (что и в таблице 1), с тем лишь отличием, что в данном случае нам неизвестны два первых значения переменной **X**.

В рамках **Available-case analysis** подхода мы считаем среднее значение, используя всю доступную информацию, то есть для переменной **X** на основе 18 известных значений, а для переменной **Y** на основе всех 20 значений.

Таким образом, на основе таблицы 2 получим следующие результаты:

- среднее значение X: 11,8889 (214/18);
- среднее значение Y: −10,7768;
- оценка (ко)вариации:

  $\sigma 11$=30,0988;

  $\sigma 12$=−43,6174;

  $\sigma 22$=60,2952.

Примечание. В расчетах использовалось n наблюдений, для которых известны как значения **X**, так и **Y** (n=18).

Значение коэффициента корреляции:

$r = -1{,}0239$.

Таким образом, расчет среднего значения на основе подхода **Available-case Analysis** привел к смещению данного значения, что в свою очередь, проявилось в рассчитанном значении коэффициента корреляции меньшим −1. Таким образом, рассчитанное значение вышло за пределы теоретически возможного диапазона [-1; 1]), что противоречит физическому смыслу.

Примечание. Если же рассчитать значение коэффициента корреляции в рамках подхода **Complete-case Analysis**, то получим значение коэффициента корреляции:

− 0,9311.

Когда гипотеза **MCAR** не выполняется, методы **Available-case analysis** так же, как и методы **Complete-case Analysis** приводят к существенным искажениям статистических свойств выборки (среднего значения, медианы, вариации, корреляции...).

К недостаткам первых двух методов обработки пропусков (**Complete-case Analysis** и **Available-case analysis**) относится и то, что, далеко не всегда исключение строк в принципе приемлемо.

Нередко процедуры последующей обработки данных предполагают, что все строки и колонки участвуют в расчетах (например, когда пропусков в каждой колонке не очень много, но при этом строк, в которых нет ни одного пропущенного поля мало).

Заполнение пропусков на основе имеющейся информации

Эти методы часто объединяют в одну группу, называемую **Single-imputation methods** (методы с одним вменением).

Заполнение пропуска средним значением

Заполнение пропуска средним значением (Mean Substitution) (другие варианты: заполнение нулем, медианой и т.п.).

Примечание. Медиана – это такое число, что половина из элементов выборки больше него, а другая половина меньше.

Всем вариантам данного метода свойственны одни и те же недостатки.

Рассмотрим эти недостатки на примере одного из наиболее простых способов заполнить пропуски **непрерывной характеристики**: заполнению пропусков средним арифметическим значением и модой.

Пример 1. На рисунке 1 показано распределение значений непрерывной характеристики до заполнения пропусков средним значением и после него.

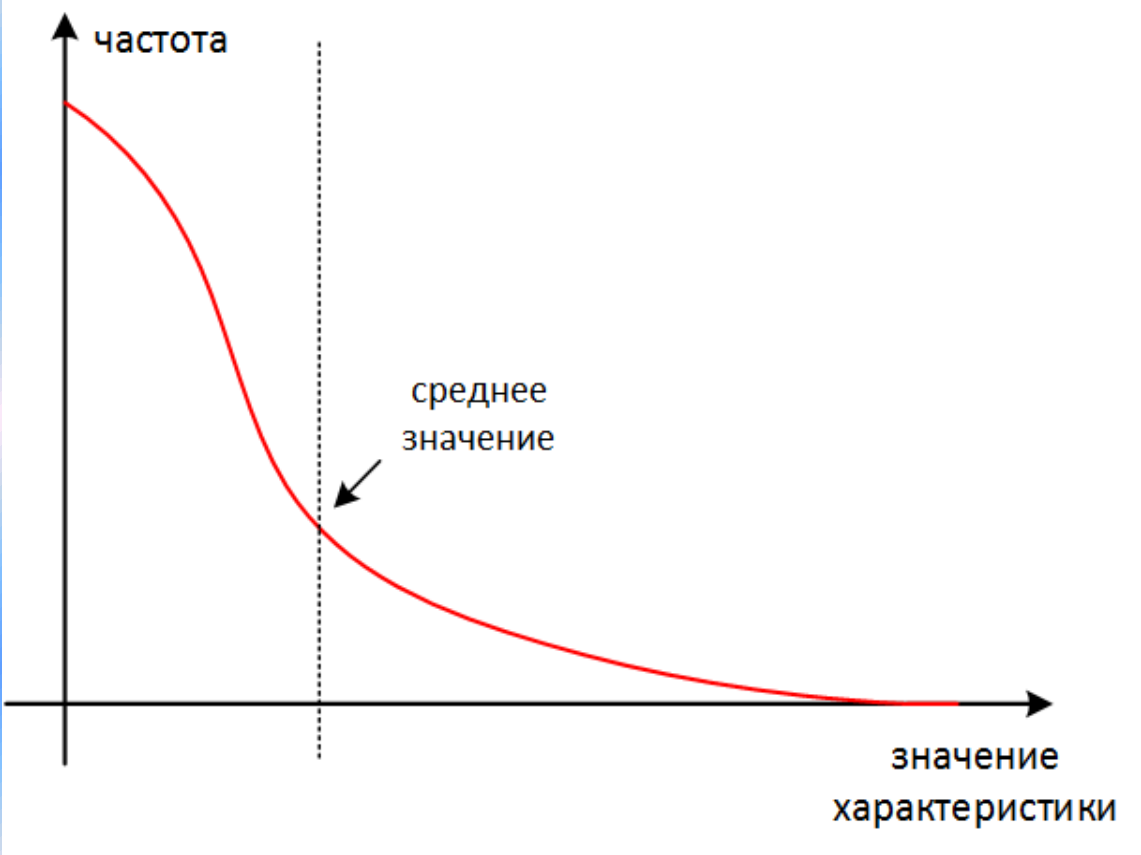

Рисунок 1а – Распределение значений непрерывной характеристики до заполнения пропусков

Рисунок 1б – Распределение значений непрерывной характеристики после заполнения пропусков

На рисунке 1 хорошо видно, что распределение после заполнения пропусков выглядит крайне неестественно. Это в итоге проявляется в искажении всех показателей, характеризующих свойства распределения (кроме среднего значения), заниженной корреляции и завышенной оценке стандартных отклонений.

Таким образом, данный метод приводит к существенному искажению распределения характеристики даже в случае MCAR.

Пример 2. В случае **категориальной/дискретной** характеристики наиболее часто используется заполнение модой.

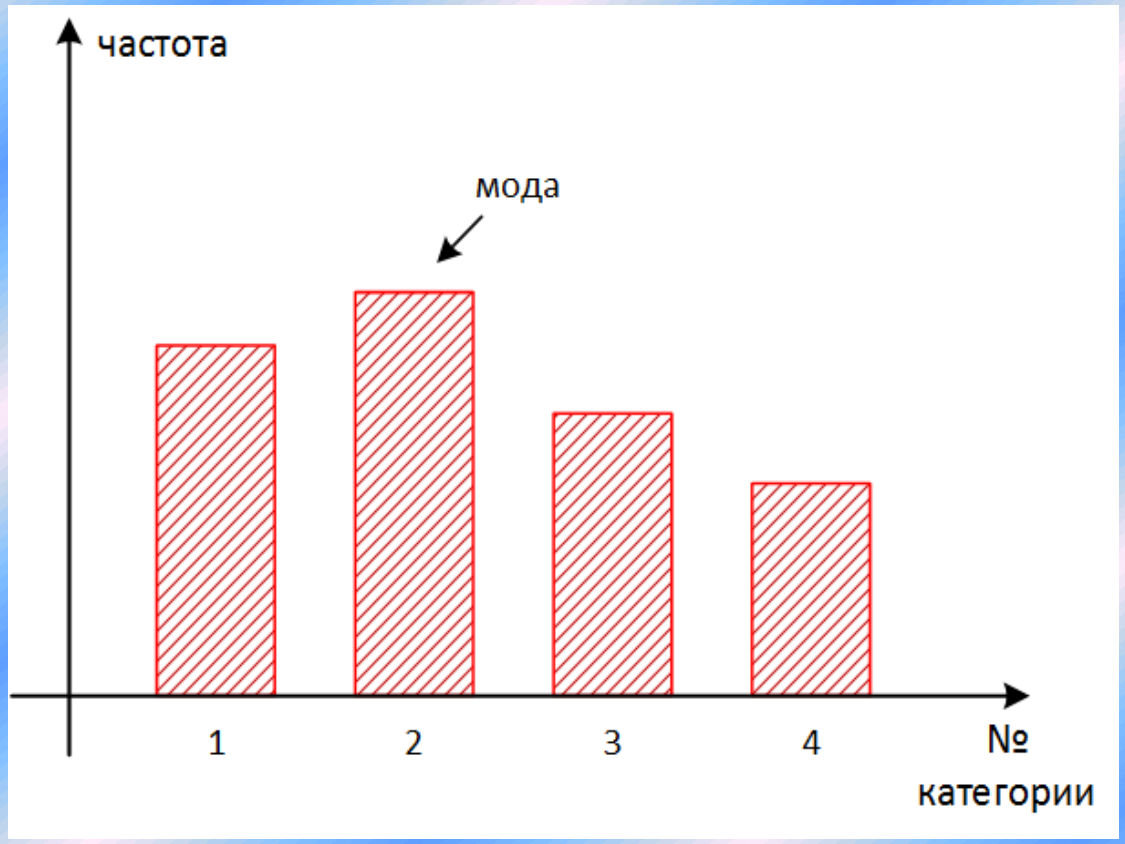

Рисунок 2а – Распределение дискретной характеристики до заполнения пропусков модой

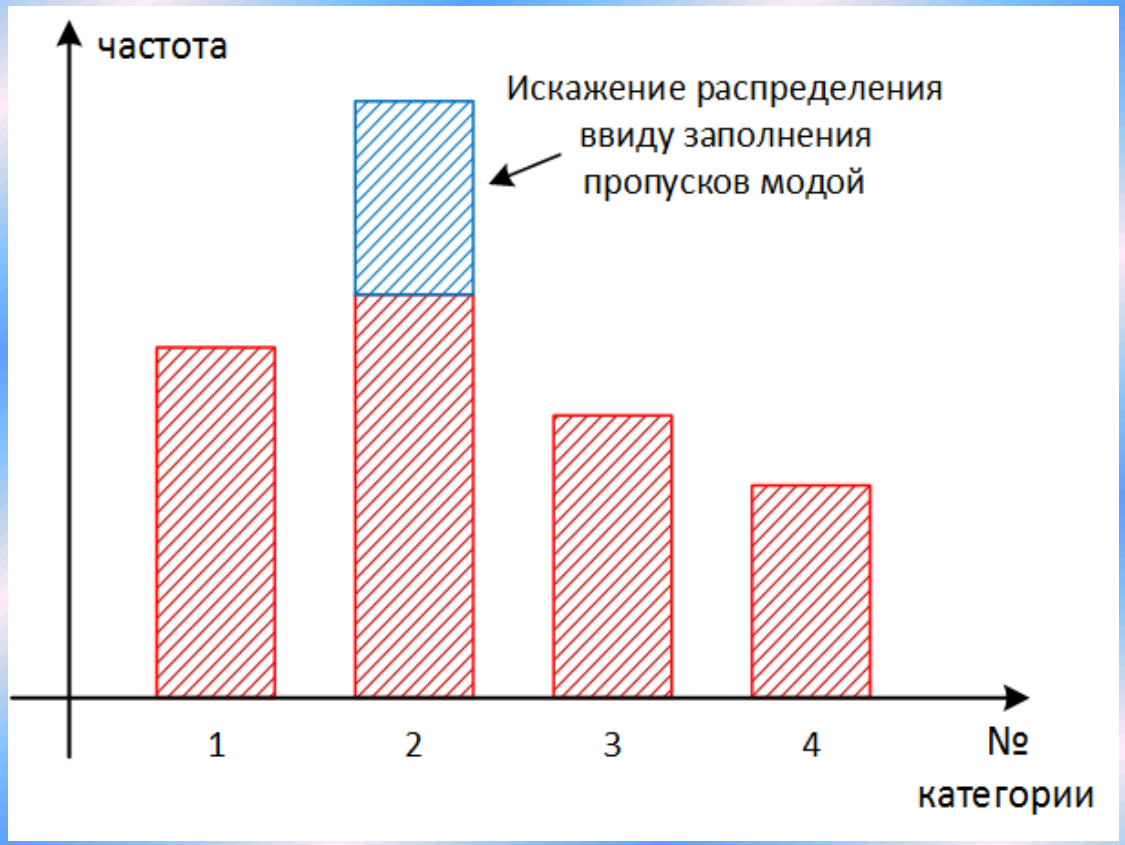

Рисунок 2б – Распределение дискретной характеристики после заполнения пропусков модой

На рисунке 2 показано распределение категориальной характеристики до и после заполнения пропусков.

Таким образом, при заполнении пропусков категориальной характеристики модой проявляются те же недостатки, что и при заполнении пропусков непрерывной характеристики средним арифметическим (нулем, медианой и тому подобным).

Мода – значение во множестве наблюдений, которое встречается наиболее часто.

Повторение результата последнего наблюдения

LOCF (Last observation carried forward) – повторение результата последнего наблюдения.

Данный метод применяется, как правило, при заполнении пропусков во временных рядах, когда последующие значения априори сильно взаимосвязаны с предыдущими.

Рассмотрим 2 случая, когда применение LOCF обосновано.

Случай 1. Если мы измеряем температуру воздуха в некоторой географической точке на открытом пространстве, причем измерения проводятся каждую минуту, то при нормальных условиях – если исключить природные катаклизмы – измеряемая величина априори не может резко (на 10–20 °C) измениться за столь короткий интервал времени между последующими измерениями. Следовательно, заполнение пропусков предшествующим известным значением в такой ситуации обоснованно.

Случай 2. Если данные представляют собой результаты измерения (например, температуры воздуха) в один и тот же момент времени в близких географических точках таким образом, что гипотеза о малых изменениях значений от одной точки набора данных до другой остается справедливой, то опять же использование LOCF логично.

Ситуации, когда использование LOCF обосновано, не ограничиваются только этими двумя случаями.

Хотя в описанных выше ситуациях метод логичен и обоснован, он тоже может привести к существенным искажениям статистических свойств даже в случае MCAR.

Так, возможна ситуация, когда применение LOCF приведет к дублированию выброса (заполнению пропусков аномальным значением).

Кроме того, если в данных много последовательно пропущенных значений, то гипотеза о небольших изменениях уже не выполняется и, как следствие, использование LOCF приводит к неправильным результатам.

Indicator Method

Indicator Method – метод, предполагающий замену пропущенных значений нулями и добавление специального атрибута-индикатора, принимающего нулевые значения для записей, где данные изначально не содержали пропусков и ненулевые значения там, где ранее были пропуски.

Пример. В таблице 3 приведены данные до заполнения пропусков.

Таблица 3 – Данные до заполнения пропусков

Параметр	12,7	7,5	?	3,1	?	?	5	5,8	3,7

Знаком ? обозначены пропуски в наборе данных.

В таблице 4 приведены данные после заполнения пропусков.

Таблица 4 – Таблица после заполнения пропусков

Параметр	12,7	7,5	?	3,1	?	?	5	5,8	3,7
Флаг	0	0	1	0	1	1	0	0	0

На практике применяются и модификации этого метода, предполагающие заполнение пропусков ненулевыми значениями. Стоит отметить, что при таком заполнении (например, средним) допустимо использование инверсных значений поля флагов (то есть 0 – для случая, когда в исходных данных значения изначально были пропущены, и ненулевое значение для случаев, когда значение поля исходных данных было известно).

Также при заполнении пропусков ненулевыми значениями часто добавляется взаимодействие поля-флага и исходного поля.

К преимуществам данного метода относятся:

- использование всего набора данных (репрезентативность выборки не страдает),
- явное использование информации о пропущенных значениях.

Несмотря на эти преимущества, даже при выполнении гипотезы MCAR и небольшом числе пропущенных значений, данный метод может привести к существенному искажению результатов.

Восстановление пропусков на основе регрессионных моделей

Этот метод заключается в том, что пропущенные значения заполняются с помощью модели линейной регрессии, построенной на известных значениях набора данных.

На рисунке 3 показан пример результатов заполнения пропущенных значений характеристики 1 на основе известных значений характеристики 2.

Рисунок 3 – Заполнение пропусков на основе линейной регрессии

Метод линейной регрессии позволяет получить правдоподобно заполненные данные. Однако реальным данным свойственен некоторый разброс значений, который при заполнении пропусков на основе линейной регрессии отсутствует.

Как следствие, вариация значений характеристики становится меньше, а корреляция между характеристикой 2 и характеристикой 1 искусственно усиливается.

В результате данный метод заполнения пропусков становится тем хуже, чем выше вариация значений характеристики, пропуски в которой мы заполняем, и чем выше процент пропущенных строк.

Существует метод, решающий эту проблему: метод стохастической линейной регрессии, проиллюстрированный на рисунке 4 (аналогично рисунку 3).

Рисунок 4 – Заполнение пропусков на основе стохастической линейной регрессии

Модель стохастической линейной регрессии отражает не только линейную зависимость между характеристиками, но и отклонения от этой линейной зависимости. Этот метод обладает положительными свойствами заполнения пропусков на основе линейной регрессии и, кроме того, не так сильно искажает значения коэффициентов корреляции.

Заполнение пропусков с помощью стохастической линейной регрессии в общем случае приводит к наименьшим искажениям статистических свойств выборки.

В случае, когда между характеристиками прослеживаются явно выраженные линейные зависимости, метод стохастической линейной регрессии нередко превосходит даже более сложные методы.

Более сложные методы – Multiple Imputation, методы функции максимального правдоподобия.

Применение рассмотренных методов может приводить к существенному искажению статистических свойств набора данных (среднее значение, медиана, вариация, корреляция…) даже в случае MCAR.

Однако, они остаются часто используемыми не только среди обычных пользователей, но и в научной среде (как минимум в областях, связанных с медициной).

Пример

В 82 работах из 100, посвященных проблеме раковых заболеваний, которые были опубликованы в 2002 году, то есть в 82% случаев, авторы указали, что столкнулись с необходимостью заполнения пропусков в данных.

При этом в 32 случаях был явно указан метод заполнения пропусков. В 12 из этих 32 работ использовался Complete Case Analysis, еще в 12 – Available Case Analysis, в 4 – Indicator Method, в 3 — ad-hoc методы и только в 1 случае использовался более сложный метод.

Спустя десятилетие ситуация не сильно изменилась.

В 2012 году в 54% случаев (в 21 статье) использовался Complete Case Analysis, в 7 случаях – LOCF, в 3 случаях – заполнение средним значением, в 1 случае – Indicator Method.

По состоянию на 2014 год рекомендуемые к использованию методы заполнения пропусков (Multiple Imputation, методы функции максимального правдоподобия) в научных статьях медицинской тематики по-прежнему применялись редко.