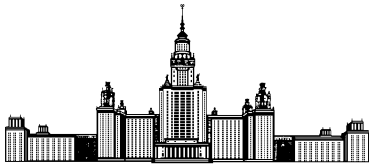
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

«Название курсовой работы»

Выполнил:

студент 3 курса 317 группы

Каюмов Эмиль Марселевич

Научный руководитель:

д.ф-м.н., профессор

Дьяконов Александр Геннадьевич

Москва, 2016

Содержание

1 Введение

В практических задачах анализа данных выборки часто содержат в себе пропущенные значения. Их причины могут быть различными, например, неответ респондента на конкретный вопрос, отказ работы датчика для измерений показателя при низких температурах, ошибки в программном обеспечении при записи данных и другие.

За редким исключением алгоритмы машинного обучения не работают с выборками, имеющими пропущенные значения. Поэтому возникает необходимость перейти к данным, не имеющим пропусков, для дальнейшней работы с ними. Существуют различные подходы к решению задачи восстановления пропусков в данных, которые различаются по своей природе, области применимости и вычислительной сложности.

...

В данной работе описаны основные методы восстановления пропусков в данных, произведено их сравнение на нескольких наборах данных, включая как данные с искусственно созданными пропусками, так и с данными, имеющими натуральные пропущенные значения.

2 Методы восстановления пропусков

Если написать классификацию MCAR, MAR, NMAR, то это, наверное, должно быть здесь. Но тогда надо объяснять, какой метод на что рассчитан (что сомнительно в какой-то мере).

Возможно, стоит делить на группы: простейшие методы (первые три), основанные на предсказаниях (логистическая регрессия, случайный лес, ...), основанные на разложениях и подобных вещах (SVD, EM, K-Means, ...).

Здесь же про постановку задачи, обозначения.

2.1 Игнорирование объектов с пропущенными значениями

Простейшим методом решения проблемы пропущенных значений в выборке является игнорироваение объектов, имеющих пропуски. Такой метод применим

только в том случае, когда малая часть объектов выборки имеет пропущенные значения. Преимуществом данного подхода является простота и невозможность испортить данные путем замены пропусков. В случае достаточно большого размера выборки метод может показывать хорошие результаты. Альтернативным вариантом в случае наличия пропусков в небольшом количестве признаков является удаление таких признаков из выборки.

2.2 Замена специальным значением

Другим простейшим методом является замена пропусков на специальное заранее определенное значение такое, как, например, 0 или -1. Данный подход позволяет не уменьшать размер выборки, однако может вносить значения, сильно отличающие от настоящих.

стоит ли писать о том, что для деревьев логично работать с -1 в категориальном случае, а для линейного такая замена может исказить результат ?

2.3 Замена самым частым или средним значением

Еще одним простым методом восстановления пропусков является замена на моду или среднее значение по конкретному признаку. В случае категориального признака все пропуски заменяются на наиболее часто встречающее значение, в случае количественного признака – на среднее значение по признаку. Данный метод, в отличие от двух предыдущих, учитывает имеющиеся данные и усредняет их. Преимуществом подхода является простота, однако на практике возникает проблема в определении, является ли конкретный признак категориальным или количественным, особенно при большом количестве признаков.

2.4 Замена с помощью SVD

...

2.5 Замена с помощью метода ближайших соседей

...

2.6 Замена с помощью случайного леса

...

2.7 Замена с помощью линейной регрессии

...

2.8 Замена с помощью EM-алгорима

...

2.9 Замена с помощью метода k средних

...

2.10 Возможные усовершенствования

Про поиск ближайшего.

3 Эксперименты с данными

Реализация всех описанных методов и эксперименты расположены в репозитории на Github [**?**].

3.1 Исходные данные и условия эксперимента

Описание трех датасетов с полностью заполненными значениями. Описание методов удаления данных.

Описание трех датасетов с натуральными пропусками в данных.

Описание того, как проводятся измерения.

3.2 Результаты эксперимента

Графики и таблицы с результатами.

3.3 Обсуждение и выводы

В каких случаях что использовать. Что лучше работает, что не стоит использовать.

4 Заключение

Список литературы