АНАЛИЗ МНОГОМЕРНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

С. В. ДРОНОВ

«Инфра-Инженерия»

С. В. ДРОНОВ

АНАЛИЗ МНОГОМЕРНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

Монография

Москва Вологда
«Инфра-Инженерия»
2025

УДК 519.23
ББК 22.172
Д75

Рецензенты:
д. ф.-м. н., профессор, ведущий научный сотрудник
(Алтайский государственный технический университет)
Пышнограй Григорий Владимирович;
д. ф.-м. н., профессор
(Алтайский государственный университет)
Родионов Евгений Дмитриевич

Дронов, С. В.

Д75 Анализ многомерных статистических данных : монография / С. В. Дронов. – Москва ; Вологда : Инфра-Инженерия, 2025. – 308 с. : ил., табл.
ISBN 978-5-9729-2524-7

Подробно рассмотрен ряд методов обработки многомерных данных, способов сокращения размерности и некоторых продвинутых методик визуализации этих данных. При этом приводится математическое обоснование и неформальное объяснение механизма действия методов. В частности, изучаются проблемы исключения грубых ошибок наблюдения, кластерный анализ, методы главных компонент, экстремальной группировки признаков и корреляционных плеяд, а также многомерного шкалирования и анфолдинга. Обсуждаются подходы к переводу качественных данных в числовую форму. Уделено значительное внимание выявлению и оцениванию силы связи как между числовыми, так и качественными показателями. Кроме этого, в книгу включены материалы о post-hoc задачах кластерного анализа и метрической структуре семейства кластерных разбиений конечного множества.

Для тех исследователей, кто, имея математическую подготовку в объеме университетских курсов высшей математики, хочет не просто использовать многомерные статистические методы, а разобраться в механизме их действия. Может быть использовано как основа для чтения специальных курсов магистрантам и аспирантам направлений, использующих в своих предметных областях статистические методы.

УДК 519.23
ББК 22.172

ISBN 978-5-9729-2524-7

Оглавление

Предисловие

Эта книга написана для тех, кто хочет не просто применить методы статистической обработки многомерных данных, но и разобраться в том, почему и как эти методы работают. Я много лет преподаю в университете вероятностные и статистические курсы студентам разных специальностей и постоянно отмечаю, что стремление разобраться в сути предмета у большинства студентов с годами все быстрее уменьшается. Это, видимо, отчасти связано с тем, что сегодня имеется достаточно большое количество разнообразных библиотек алгоритмов и стандартных программ работы с данными, легко доступных через интернет. Более того, очень популярный нынче предмет «Машинное обучение» лозунг «Вы только намекните, что вам нужно, а соответствующий алгоритм мы подберем автоматически» сделал своим главным девизом. Поэтому многие склонны рассматривать любой курс, посвященный методам обработки данных как большой справочник с приложенным к нему автоматическим поиском. А, поскольку пытаться изучать справочник – заведомо бессмысленная затея, то и отношение к моим предметам возникает соответственное.

Тем не менее, сколько бы самых замечательных методов обработки данных не было бы создано, практик неизбежно столкнется с ситуацией, которая адекватно не обрабатывается ни одним из этих методов. Я неоднократно убеждался в этом, помогая специалистам различных конкретных предметных областей в анализе их данных. Чаще всего это были специалисты-медики, которые оказались плотно привязаны к структурированию и обработке своих данных в силу всеобщего принятия концепции доказательной медицины (нужно не просто предложить, например, новую методику лечения, но и убедительно показать, что она работает лучше существующих). Единственный шанс не спасовать в ситуации, когда классические методики отказываются работать, это предложить собственную модификацию метода, которая уже сработает на конкретном

практическом примере. Конечно же, такую модификацию вы сможете предложить только тогда, когда сумеете полностью разобраться в сути проблемы и возможных альтернативных подходах к ее решению.

Целью, которую я видел перед собой, занимаясь написанием настоящей книги, является именно объяснение причин и механизмов действия основных статистических методов в надежде, что это поможет внимательному читателю сделать что-то подобное в своей конкретной задаче.

При этом, поскольку предмет наш, разумеется, математический, для адекватных объяснений приходится привлекать некоторые математические конструкции и факты. Поэтому лучше, если у читателя имеется соответствующая математическая база. В самой большой степени нам придется использовать университетские курсы теории вероятностей и математической статистики. Коротко важнейшие для нашей тематики сведения из них я попытался напомнить в первой главе книги.

Не ждите здесь описания каких-то новейших продвинутых методов. Теория анализа данных слишком быстро развивается, и попытка успеть за ней в печатном тексте не может вызвать ничего, кроме саркастической усмешки. Посмотрите, например, как выглядят в старых книгах и справочниках по статистике обзоры программного обеспечения! В предлагаемой вам книге рассмотрен лишь набор проверенных временем методов обработки данных и предпринята попытка сделать их анализ в заявленном ключе.

Тем не менее, я счел возможным включить в текст некоторые собственные результаты, в частности, посвященные обработке нечисловых данных, ступенчатых зависимостей, а также главу, посвященную геометрии семейств кластерных разбиений, которые ранее не публиковались нигде, кроме специализированных журналов.

Можно, видимо, эту книгу использовать и как справочник, снабженный примерами, но ее ценность от этого значительно снизится. Впрочем, о том, насколько удачной и ценной получилась книга, судить не мне, а вам – ее читателям.