Предисловие

Издание, как и предыдущие книги автора — «Современные статистические методы медицинских исследований» (М.: URSS, 2008, 2013), «STATISTICA 6. Математическая статистика с элементами теории вероятностей» (М.: Бином, 2010), посвящено применению методов анализа данных в медицинских исследованиях. Но в отличие от указанных книг, в которых рассматривались традиционные многомерные методы, описано применение современных методов машинного обучения, являющихся составной частью искусственного интеллекта. В какой-то степени издание является продолжением книги «Методы машинного обучения Data Mining пакета STATISTICA» (М.: Горячая линия — Телеком, 2022), где описано решение задач классификации. В настоящем издании дополнительно рассмотрены задачи регрессии. В задачах классификации по предикторным переменным прогнозируется значение категориальной целевой переменной, например диагноз заболевания, исход лечения; в задачах регрессии прогнозируется значение непрерывной целевой переменной, например длительность операции, уровень сахара в крови. Изложение ведется на примерах общедоступных в Интернете датасетов (https://www.kaggle.com) медицинского характера, что облегчит понимание материала медиками и даст возможность читателю повторить приведенные результаты расчетов. Освещены методы машинного обучения Data Mining пакета STATISTICA: деревья решений — общие деревья классификации и регрессии, общие CHAD-модели, интерактивные деревья, бустинг деревьев классификации и регрессии, случайный лес регрессии и классификации; процедуры обучения — методы опорных векторов, k-ближайших соседей, байесовский классификатор; автоматизированные нейронные сети; кластерный анализ. Описана технология работы с мастером проектов Data Miner.

Методы машинного обучения открывают новые перспективы в создании медицинских систем поддержки принятия решений, интегрированных с искусственным интеллектом. Обработка и анализ средствами компьютерного зрения изображений, полученных рентгеновским оборудованием, томографами, ускорят

диагностику заболеваний, повысят ее точность. Прогностические модели, построенные на основе выявленных скрытых знаний в массивах медицинских данных, повысят качество идентификации заболеваний, оценки состояний больных, рисков, предсказаний развития и распространения заболеваний, эпидемий.

Машинное обучение, как и в целом анализ данных, вне зависимости от области применения — медицина, экономика, химия и т. д. не меняет своей сути, поэтому издание будет одинаково доступно восприятию всеми, кто интересуется анализом данных. Адресовано студентам и аспирантам, преподавателям вузов и научным работникам, врачам и управленцам, экономистам и социологам, представителям естественнонаучных и инженерно-технических специальностей, всем, кто в процессе обучения или профессиональной деятельности использует методы анализа данных. Простая и доступная для широкого круга читателей форма изложения, использование датасетов свободного доступа делает возможным самостоятельное изучение методов машинного обучения Data Mining. При написании книги использована русскоязычная версия пакета STATISTICA 13 (Tibco, USA).

Дополнительную информацию по методам анализа данных можно посмотреть по ссылке https://stat-lab.ru/, там же можно скачать датасет для самостоятельной работы с ранее вышедшим учебным пособием «Методы машинного обучения Data Mining пакета STATISTICA». Для контактов: E-mail: statlab@kubsu.ru.

Приступая к лечению, врач полагается не на рассуждения, а на опыт, подкрепленный разумом.

Гиппократ

ВВЕДЕНИЕ

Искусственный интеллект (ИИ) можно определить как технологию или область знаний разработки систем, воспроизводящих творческие, мыслительные функции человека в процессе сбора информации и обучения для достижения поставленных человеком целей. В настоящее время ИИ успешно применяется в различных областях человеческой деятельности — в медицине, финансовой сфере, промышленности, государственном управлении, военном деле и т. д. В медицине можно условно выделить три направления, в которых активно применяются методы ИИ: диагностика заболеваний посредством обработки изображений рентгеновских установок, компьютерных и магнитно-резонансных томографов; создание и тестирование молекул новых лекарственных препаратов; обработка медицинских записей с целью прогнозирования рисков, возможных состояний больных, исходов лечения, установления диагнозов, распространения эпидемий и т. д.

Во всех направлениях применяется машинное обучение, состоящее из методов и алгоритмов ИИ, которые в процессе решения задач обучаются на исходных данных:

- первое направление, называемое компьютерным зрением, использует алгоритмы, которые учатся «видеть», извлекая информацию из изображений для идентификации объектов на них;
- второе направление применяет алгоритмы глубокого машинного обучения на основе многослойных нейронных сетей со сложной математической структурой, обучение происходит на данных больших размеров для выявления закономерностей и последующего моделирования молекул лекарственных препаратов;
- третье направление использует алгоритмы методов классификации и регрессии машинного обучения, реализация которых в пакете STATISTICA и рассмотрена в настоящем издании.

Методы машинного обучения синтезируют методы статистики, теории вероятностей, численных методов, теории оптимизации и других дисциплин. Многие задачи машинного обучения могут быть поставлены как оптимизационные: мы строим модель, используя обучающее множество, и проверяем ее качество на тестовом наборе данных [1]. Для применения методов машинного обучения при решении задач классификации или регрессии необходимо, чтобы данные выборки имели определенную структуру, а именно были помеченными. Каждому объекту выборки наряду с признаками, его характеризующими, при помощи категориальной или непрерывной целевой переменной должна быть присвоена выходная информация, определяющая его принадлежность к некоторому классу или некоторому значению непрерывной переменной. Это положение в полной степени справедливо и для медицинских исследований. Например, предположим, что применительно к задаче классификации необходимо сделать прогноз возможности отторжения трансплантата у больного, перенесшего пересадку органа. Тогда выборка должна состоять из больных, у которых произошло отторжение трансплантата, и больных, у которых не произошло. Пометка данных состоит в том, что при помощи введенной категориальной целевой переменной первые больные определяются, например, как принадлежащие к классу «произошло», вторые — как принадлежащие к классу «не произошло». В процессе машинного обучения по клинико-лабораторным показателям больных и выходной информации создается математический образ этих классов. Если нужно сделать прогноз для некоторого больного, отсутствующего в выборке, проверяется соответствие его математического образа образу класса и больного относят к тому классу, к которому больше соответствие. Соответствие может определяться различными способами в зависимости от метода, например при помощи вероятности или расстояния до центра класса. Так как данные помечаются специалистами, то применяемый способ называют машинным обучением с учителем или контролируемым обучением. В задачах кластеризации — разделения данных на группы однородности по совокупности категориальных или непрерывных переменных — пометка данных не требуется. Поэтому соответствующий способ называют машинным обучением без учителя или неконтролируемым обучением.

Принятие решений в медицине относительно стратегии и тактики лечения больного обладает определенной спецификой.

В большинстве случаев — это и недостаточность знаний, ограниченность временных ресурсов, отсутствие возможности привлечения компетентных экспертов и т. д. Самым важным, по-видимому, является неполнота информации о состоянии больного и его рисках, перспектив лечения, так как «болезнь одна, а больные разные». По истории болезни пациента, состоянию его здоровья на данный момент времени врачу или в особо сложных случаях консилиуму врачей проблематично сделать объективный прогноз даже на неотдаленное будущее, полагаясь только на опыт, знания и интуицию. Отмеченные факторы могут быть причинами принятия неэффективных решений, приведших к длительной реабилитации или дальнейшей потере здоровья пациента. Поэтому важной является проблема создания медицинских систем поддержки принятия решений (СППР), которые являются наукоемкими и предполагают использование определенных научных направлений, например методов математики, анализа данных.

Особые перспективы в разработке медицинских СППР имеют методы анализа данных и ИИ, так как медицина по своей сути экспериментальная область знаний, накапливающая большие массивы медицинских данных. В этих данных содержатся невидимые «невооруженным глазом», скрытые знания, закономерности, присущие заболеваниям. Традиционные методы многомерного разведочного и углубленного анализа [2–5] могут быть применены для исследования медицинских записей, поиска скрытых знаний, закономерностей, их формализация в виде вероятностно-статистических моделей. Методы машинного обучения, не требующие каких-либо ограничений на данные, позволяющие строить модели классификации и регрессии по любому количеству категориальных и непрерывных предикторов (в том числе когда все предикторы категориальные), могут быть использованы для построения достоверных прогностических моделей. Методы анализа данных и ИИ в составе медицинских СППР, обеспечивая временными, информационными, познавательными ресурсами, способны, дополнив знания о состоянии больного и его заболевании, оказать врачу помощь в принятии правильных и эффективных решений.

В главе 1 рассмотрены датасеты медицинского характера, использованные при описании методов машинного обучения Data Mining пакета STATISTICA (https://www.kaggle.com). На данных ССЗ (сердечно-сосудистые заболевания), состоящих из 918

наблюдений, характеризующихся 12 показателями (переменными), рассмотрено решение задач классификации. Решение задач регрессии описано на данных «Медицинские расходы» по медицинскому страхованию, содержащих 1338 страховых случаев и 7 переменных. При решении задач классификации и регрессии временных рядов автоматизированными нейронными сетями использованы данные «COVID-19» из 270 наблюдений.

В главе 2 рассмотрены задачи машинного обучения, решение которых предусмотрено нейронными сетями в Data Mining пакета STATISTICA, — классификация и регрессия, кластеризация сетями Кохонена, классификация и регрессия временных рядов.

В главе 3 приведено решение задач классификации и регрессии процедурами обучения методом опорных векторов, k-ближайших соседей, байесовским классификатором.

В главе 4 описано решение задач классификации и регрессии методами деревьев решений — общими деревьями, CHAID-моделями, интерактивными деревьями, растущими деревьями, случайным лесом.

В главе 5 с целью сравнительного анализа по точности решения задач классификации и регрессии методов машинного обучения и традиционных методов многомерного и углубленного анализа данных описаны результаты реализации методов общих линейных моделей и общих моделей дискриминантного анализа.

В главе 6 рассмотрено решение задачи кластеризации (классификации без учителя) алгоритмами k-средних, EM и иерархической классификации.

В главе 7 приведено решение задачи регрессии Мастером проектов Data Miner.

При описании методов машинного обучения не ставилась задача разработки моделей с наилучшими прогностическими свойствами, поэтому у читателей есть возможности экспериментирования с целью улучшения прогностических свойств построенных и описанных в издании моделей.

Материал книги изложен на данных медицинской направленности, но формальная математическая сторона методов анализа данных безразлична к природе исследуемых объектов. Поэтому надеемся, что издание будет полезно и интересно самому широкому кругу читателей, всем, занимающимся анализом данных, включающим традиционные методы многомерного и углубленного анализа, а также современные методы машинного обучения — нейронные сети, деревья решений, процедуры обучения.