1. Введение в большие данные

Большие данные — это разнообразные данные, поступающие с огромной скоростью, объем которых постоянно растет.

Почему мир практически свихнулся на термине "большие данные"? С точки зрения чистой математики - чем больше и разнообразнее у вас данные, тем более сложные регрессионные модели вы можете на них натренеровать. Возможно вы помните простое правило из алгербы 6/7 класса - для однозначного решения системы уравнений необходимо, чтобы число переменных равнялось числу линейно независимых уравнений в системе.

С большими данными все то же самое - огромное разнообразие переменных и большое накопление, позволяет нам зафитить самые запутанные или даже может быть самые незначительные казалось бы зависимости.

И естественно, что область медицины одина из фаворитов на использование больших данных для тренировки многопараметрических моделей.

▎2. 5 V больших данных

Как мы уже сказали, великость этих самых больших данных характеризуются в первую очередь разнообразием, скоростью поступления и объемом

• Объем (Volume): количество данных, которые были сгенерированы в здравоохранении. Официальные цифры ТАСС на 22 год - создается порядка 2 млн документов в сутки (рост по сравнению с 19 годом в 40 раз). Большие госпитали создают порядка 10 ТБ в год. На национальной конференции Big Data Health Science 2024 года, быза озвучена очень интересная цифра, что около 30% данных, генерирующихся на планете земля, это медицинские данные

• Скорость (Velocity): скорость, с которой данные создаются и обрабатываются.

• Разнообразие (Variety): огромное количество типов данных создается в медицинской области - это и табличные данные, и нереляционные данные, тексты, изображения, видео, сейчас уже появляется и аудио данные. И структурированные и неструктурированные. И все это различные срезы одного и того же процесса.

• Достоверность (Veracity): надежность и точность данных. У любых зафиксированных данных есть точность их фиксации. Связано это может быть связано с человеческим фактором, с шумом при регистрации показаний, либо с помехами и ошибками самого прибора. С помехами приборов и шумом при записи показаний борятся сами производители приборов путем введения жестких норм использования прибора, процедур контроля истиности показаний прибора и т.д. Но огромное количество информации генерируется вручную врачами. В последнее время МИСы берут на себя частичную ответственность за правильность ввода данных - это сппр, внедрение клинреков непосредственно в программы в цифровом виде, чеклисты, утвержденные цифровые формы документов, прописанные технические процессы (нельзя закрыть документ, пока не заполнено определнное поле) и т.д. То есть достоверность растет - вводятся шаблоны работы и т.д.

• Ценность (Value): ценность данных измеряется уровнем возможности их анализа и понимания. То есть, можем ли мы использовать данные для улучшения глобалььного здоровья населения или же оптимизации медицинских техпроцессов. Решение подобных амбициозных задач требует датасета высокого качества - он должен быть достаточно полный, структурированный, с высокой достоверностью и разнообразным. В данный момент в медицине существует проблема отсутствия единого формата медицинского документа и единого формата хранения мед документации. Плюс к этому, специфика медицинской информации не позволяет ею свободно делиться. Там очень много персональных данных, у каждой клиник своя МИС, каждый врач пишет диагноз по своему и так далее. из-за этого эти данные очень трудно друг с другом пожениить. И как такового медицинского корпуса нет. Первая задача, которая сейчас стоит перед датасаентистами и перед ML-инженерами это сделать данные анализируемыми и понятными, поженить разные форматы данных. То есть сделать так, чтобы модель могла опираться не только на персональные данные Иван Иваныча из москвы, но на Иван Иваныча из владивостока, а то и Иван Иваныча из австралии. Мы приходим к ситуации, когда данных очень много, но ценность у них пока еще не так велика. Требуется огромная аналитическая работа.

Источники:

▎3. Источники больших данных в медицине

• Электронные медицинские записи (ЭМЗ): данные о пациентах, истории болезни, диагнозах и лечении.

• Геномные данные: информация о генетическом коде (500гб)

• Данные с носимых устройств: информация о физической активности, сердечном ритме и других показателях здоровья.

• Данные из клинических испытаний: результаты исследований и испытаний новых методов лечения.

Это все сырые данные

Отдельно заслуживает внимание уже переработанный невероятно качественный массив информации - медицинская литература, научные статьи. Когда мы работаем с моделями машинного обучения, то нам приходится выбирать - использовать ли многопараметрические модели глубокого обучения на сырых диких данных, или же испольщовать модели поменьше и попроще, но на данных высокого качества.

• Данные о страховании: информация о страховых полисах, выплатах и обращениях за медицинской помощью.

Информация, популяризирующая медицину в широких массах

• Социальные сети и мобильные приложения: отзывы пациентов, обсуждения заболеваний и их лечения.

Проблемы больших данных в медицине

1. Самая проблема, которая приходит на ум - это проблема этическая, проблема конфиденциальность и деликатность медицинской информации. Развитие новых способов защиты информации и новых способов дешифровки и взлома - классическое соревнование оружия и брони. Плюс к этому, чем лучше мы защищаем наши данные, тем, как правило, они менее пригодны для анализа. Это связано со спецификой методов защиты информации. В данный момент возникает еще и проблема с точки зрения закона - понятие персональных данных определено очень широко - это любые данные, с помощью которых можно идентифицировать индивида. Тут можно привести историю про нетфликс и imdb, результаты анонимного опроса пользователей нетфликса были деанонимизированны благодаря открытым базам сервиса imdb.

2. Вторая проблема, которая приходит на ум, это проблема техническая - как уже было сказано, данные не стандартизированы. Разным программным продуктам очень сложно друг с другом общаться.

И третья тоже техническая проблема связана с защитой медицинских данных - их очень тяжело получать и передавать.

Где хранятся?

как хранятся?

диски для БД