


Концепция воспроизводимых исследований

A large, dark blue, diagonal shape that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

Машинное обучение (МО) нашло применение в исследованиях всех областей науки и во многом заменило традиционную статистику.

И хотя для анализа данных зачастую проще использовать именно МО, присущий этой технологии «подход чёрной коробки» вызывает серьёзные проблемы при интерпретации результатов.

Термин «кризис воспроизводимости» означает, что тревожно большое количество результатов научных экспериментов не нашли своего подтверждения при проведении тех же манипуляций другими группами учёных.



Что стоит понимать под термином «воспроизводимые исследования»?

Воспроизводимые исследования (Reproducible research) - это термин, используемый в некоторых областях исследований для обозначения определенного способа проведения анализа, который предоставляет:

- инструменты преобразующие необработанные данные и метаданные в обработанные данные;
- инструменты выполняющие анализ данных;
- инструменты агрегирующие анализы в отчет.

«Цель воспроизводимых исследований - привязать конкретные инструкции к анализу данных и экспериментальным данным, чтобы исследование можно было воссоздать, лучше понять и проверить»

Более широкое использование термина «воспроизводимость»

Ассоциации вычислительной
техники:

[Терминология](#)

На основании определений Ассоциации вычислительной техники я предлагаю принять следующие определения:

Повторяемость измерений(также сходимость результатов измерений, англ. Repeatability) (Same team, same experimental setup): Результат может быть получен с заявленной точностью при нескольких испытаниях в тех же условиях.

Повторяемость исследований (англ. Replicability) (Different team, same experimental setup): Результат может быть получен с заявленной точностью при нескольких испытаниях в тех же условиях, но другой командой.

Воспроизводимость (англ. Reproducibility): (Different team, different experimental setup): Результат может быть получен независимой группой, используя артефакты, которые они разрабатывают полностью независимо.

Машинное обучение — разновидность алхимии

Один из исследователей ИИ, Али Рахими, назвал технологии машинного обучения разновидностью алхимии. Полный текст заявления можно прочитать в [соответствующей статье](#) его блога.

Можем ли мы доверять исследованиям с использованием МО?

В феврале 2019 Джениффера Аллен сделала [тревожное заявление](#) для Американской ассоциации содействия развитию науки: учёные, полагающиеся на машинное обучение, обнаруживают определённую систематику в данных, даже если алгоритм просто заикливается на информационном шуме, который в ходе повторного эксперимента, как правило, не повторяется.

Недостаточное понимание алгоритма МО

Недостаточное понимание алгоритма — очень распространённая проблема в машинном обучении. Если вы не знаете, как алгоритм выводит результат, как вы можете быть уверены, что он не «жульничает», выводя несуществующие корреляции между переменными?

Недостаточное знакомство с исходными данными

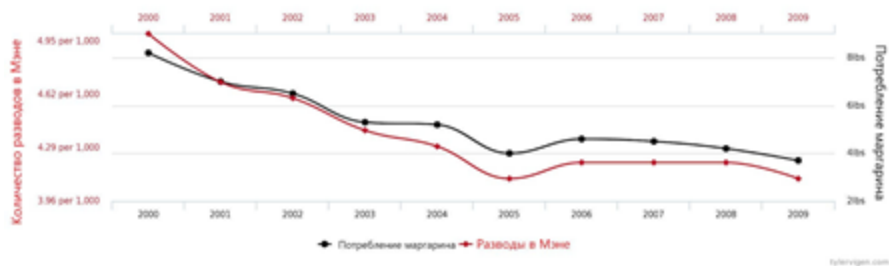
Плохое понимание исходных данных также является серьезной проблемой, но эта проблема существовала и во время работы с традиционными статистическими методами. Ошибки в сборе данных — такие как ошибки квантования, неточности считывания и использование замещающих переменных — самые распространенные затруднения.

Неверная интерпретация результатов

Количество разводов в Мэне

коррелирует с

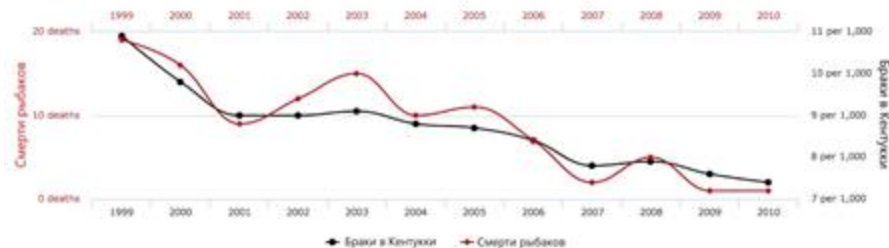
Потребление маргарина на душу населения



Люди, утонувшие, выпав из рыболовной лодки

коррелирует с

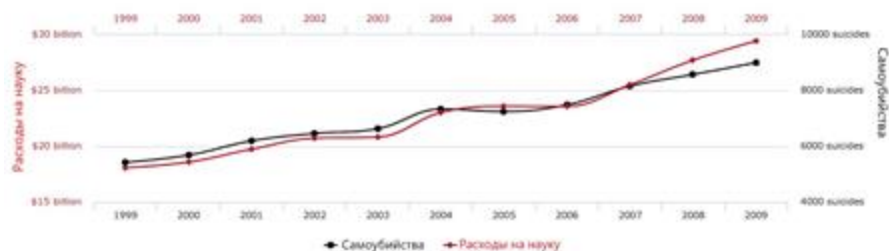
Количество браков в Кентукки



Расходы США на науку, космос и технологии

коррелирует с

Самоубийства путём повешения и удушения



Что такое p-hacking?

Суть p-hacking'a состоит в дотошном поиске в наборе данных статистически значимых корреляций и принятии их за научно обоснованные.

Чем больше у вас данных, тем вероятнее найти ложные корреляции двух переменных.

Обычно научный подход последовательно включает формулирование гипотезы, сбор данных и анализ собранных данных для подтверждения обоснованности гипотезы. В процессе же p-hacking'a сначала проводится эксперимент, и по его результатам формируются гипотезы, объясняющие полученные данные.

Форсирование корреляций

Ещё одна проблема алгоритмов машинного обучения заключается в том, что алгоритм должен делать предположения. Алгоритм не может «ничего не найти».

В настоящий момент я не знаю алгоритма машинного обучения, который мог бы прийти к заключению, что данные не подходят для того, чтобы делать обоснованные выводы. Предполагается, что это работа учёного.

Зачем использовать машинное обучение?

Хороший вопрос.

Машинное обучение упрощают анализ данных и алгоритмы МО делают за пользователя громадную работу.

В тех областях, где учёные имеют дело с действительно большими объемами данных, традиционные методы статистического анализа оказываются неэффективными и применение МО — единственный разумный способ обработки информации.

Что можно сделать?

Конечно, не всё так трагично. Та же проблема всегда присутствовала при использовании традиционных статистических методов анализа. Она лишь усугубилась с появлением больших наборов данных и алгоритмов, которые находят корреляции автоматически и не настолько прозрачны, как стандартные методы. И это усиление выявило недостатки научного процесса, которые ещё предстоит преодолеть.

10 правил для проведения воспроизводимых исследований

Правило №1— Для
каждого полученного
результата сохраните
алгоритм его
получения.

Важно знать каким образом вы получили те или иные результаты. Знание того, как вы перешли от необработанных данных к заключению, позволяет вам:

- защищать результаты
- обновлять результаты, если обнаружены ошибки
- воспроизводить результаты при обновлении данных
- представить свои результаты на обсуждение

Правило №2 – избегайте этапов ручного управления данными или процессом.

Может возникнуть соблазн открыть файлы данных в редакторе и вручную исправить пару ошибок форматирования или удалить выбросы. Кроме того, современные редакторы позволяют легко форматировать файлы огромных размеров. Однако соблазну сократить ваш алгоритм следует сопротивляться. Ручная обработка данных - это скрытая манипуляция.

Правило №3 – сохраните точные версии всех использованных внешних инструментов.

В идеале вы должны настроить виртуальную машину или контейнер со всем программным обеспечением, используемым для запуска ваших скриптов. Это позволяет сделать снимок вашей аналитической экосистемы, что упрощает воспроизведение ваших результатов.

По крайней мере, вам необходимо задокументировать выпуск и версию всего используемого программного обеспечения, включая операционную систему. Незначительные изменения в программном обеспечении могут повлиять на результаты.

Правило №4 – используйте контроль версий.

Для отслеживания версий ваших скриптов следует использовать систему контроля версий, такую как Git. Вы должны пометить (сделать снимок) текущее состояние скриптов и ссылаться на этот тег во всех получаемых вами результатах. Если вы затем решите изменить свои алгоритмы, что вы обязательно сделаете, можно будет вернуться во времени и получить точные сценарии, которые использовались для получения заданного результата.

Правило №5 – храните все промежуточные результаты в стандартизированном виде.

Если вы соблюдаете Правило № 1, в теории уже возможно воссоздать любые результаты на основе необработанных данных. Однако, хотя это может быть теоретически возможно, на практике могут быть ограничивающие факторы.

В этих случаях может быть полезно начать исследование с набора производных данных, которые уже могут представлять больше пользы или быть более удобными, чем необработанных данных. Хранение этих промежуточных наборов данных (например, в формате CSV) предоставляет больше возможностей для дальнейшего анализа и может упростить определение проблемных мест поскольку нет необходимости все переделывать.

Правило №6 – для алгоритмов использующих случайность записывайте их случайное зерно.

Одна вещь, которую специалисты по данным часто не могут сделать - это установить исходные значения для своего анализа. Это делает невозможным точное воссоздание исследований машинного обучения. Многие алгоритмы машинного обучения включают стохастический элемент, и, хотя надежные результаты могут быть статистически воспроизводимыми, нет ничего, что можно было бы сравнить с теплым сиянием в глазах проверяющего при точном совпадении результатов.

Правило №7 – всегда храните вместе с графиками данные.

Если вы используете скриптовый язык программирования, ваши графики скорее всего генерируются автоматически. Однако, если вы используете такой инструмент, как Excel, убедитесь, что вы сохранили начальные данные. Это позволяет не только воспроизвести график, но также более детально просмотреть лежащие в основе данные.

Также стоит всегда сохранять алгоритмы, которые вы использовали для получения график на основе которых вы потом приводите какие-либо утверждения.

Правило №8 – иерархический подход при генерировании результатов анализа.

Наша задача как специалистов по обработке данных - обобщить данные в той или иной форме. Вот что включает в себя извлечение информации из данных.

Однако резюмирование также является простым способом неправильного использования данных, поэтому важно, чтобы заинтересованные стороны могли разбить сводку на отдельные точки данных. Для каждого итогового результата укажите ссылку на данные, использованные для расчета итогового значения.

Правило №9 – всегда указывайте вместе текстовые утверждения и результаты исследования.

В конце работы результаты анализа данных оформляются в текстовом виде. И слова неточны. Иногда бывает трудно определить связь между выводами и анализом. Поскольку отчет часто является самой важной частью исследования, важно, чтобы его можно было связать с результатами и, в соответствии с правилом № 1, с исходными данными.

Правило №10 – обеспечивайте доступность ваших результатов, данных и исследований.

В коммерческих условиях может быть нецелесообразно предоставлять открытый доступ ко всем данным. Однако имеет смысл предоставить доступ другим пользователям в вашей организации. Облачные системы управления исходным кодом, такие как Bitbucket и GitHub, позволяют создавать частные репозитории, к которым могут получить доступ любые авторизованные коллеги.

Заключение

Машинное обучение в науке представляет проблему из-за того, что результаты недостаточно воспроизводимы. Однако учёные в курсе этой проблемы и работают над моделями МО, дающими более воспроизводимый и прозрачный результат. Настоящий прорыв произойдет, когда эта задача будет решена для нейросети.

Как сказал физик Ричард Фейнман в своей речи перед выпускниками Калифорнийского технологического института в 1974 году:

“Первый принцип науки заключается в том, чтобы не одурачить самого себя. И как раз себя-то одурачить проще всего.”

Список источников

“Кризис машинного обучения в научных исследованиях” <https://tproger.ru/translations/machine-learning-crisis>

“Reproducibility vs. Replicability: A Brief History of a Confused Terminology”
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5778115>

“Repeatability, Reproducibility, and Replicability: Tackling the 3R challenge in biointerface science and engineering”
<https://avs.scitation.org/doi/full/10.1116/1.5093621>

“10 RULES FOR CREATING REPRODUCIBLE RESULTS IN DATA SCIENCE” <https://dataconomy.com/2017/07/10-rules-results-data-science/>