**Инструкции для преподавания по теме «Проблема конфиденциальности данных при использовании ML»**

1. **Титульный слайд (название модуля, автор)**

Приветствие. Занятие «Проблема конфиденциальности данных при использовании ML» в модуле «Искусственный интеллект в информационной безопасности».

1. **План лекции**

Рассмотреть план лекции со слайда.

1. **Постановка проблемы**

Как вы уже знаете, глубокое обучение, являясь подразделом машинного обучения, основано на изучении данных. Но часто изучаемые данные являются глубоко личными. Многие модели анализируют частную информацию, рассказывающую о жизни людей то, что иным способом трудно было бы узнать.

Основным ресурсом в глубоком обучении являются обучающие данные (естественные или синтетические). Без этих данных глубокое обучение невозможно; а поскольку самые ценные модели часто используют наборы личных данных, глубокое обучение нередко становится причиной, почему компании стремятся собрать такие данные. Они нужны им для использования в конкретных сферах.

Кроме отсутствия желания пользователей делиться своими данными, существует еще ряд законодательных ограничений. Их мы рассмотрим на следующем слайде.

Все эти пункты можно разобрать на примере формирования модели, позволяющей диагностировать раковые опухоли на ранних этапах.

1. **Классификация информации по видам доступа**

Классификация информации по видам доступа приведена на схеме ниже (в соответствии с российским законодательством). Кратко привести примеры по каждому из видов информации.

Отсюда важный вывод, что конфиденциальные данные и персональные данные это не одно и тоже. Персональные данные — это подмножество конфиденциальных данных. Детальное определение будет дальше. В лекции мы сосредоточимся на персональных данных, так как часто именно они представляют интерес для обучения моделей. Однако сделанные выводы могут быть перенесены и на другие виду конфиденциальной информации.

На самом деле зачастую одни и те же сведения могут относится к различным категориям тайн. Например, сведения, составляющие врачебную тайну, наверняка являются чьими-то персональными данными. Поэтому это деление весьма условно.

Подробнее и с примерами: <https://www.securitylab.ru/blog/personal/aguryanov/29908.php>

1. **Нормативные правовые акты, регламентирующие вопросы обработки персональных данных**

Видим довольно большой перечень документов разного уровня, направленных на защиту персональных данных, поэтому вопрос является актуальным. Детально на этом вопросе останавливаться не будем, а лишь поверхностно посмотрим на положения 152-ФЗ.

1. **ФЗ-152**

Персональные данные - любая информация, относящаяся к определенному физическому лицу. Разобрать примеры персональных данных со слайда.

1. **ФЗ-152. Определения**

На этом слайде разобраны определения из ФЗ-152, на которые нужно будет опираться в рамках лекции. Кратко разобрать все определения.

1. **ФЗ-152. Категории ПД**

Разобрать категории персональных данных и их состав. Привести примеры.

1. **ФЗ-152. Требования законодательства при обработке ПД**

Основные требования:

* обработка персональных данных должна ограничиваться достижением конкретных, заранее определенных и законных целей;
* содержание и объем обрабатываемых персональных данных должны соответствовать заявленным целям обработки;
* обрабатываемые персональные данные не должны быть избыточными по отношению к заявленным целям их обработки;
* должно быть получено согласие субъекта на размещение его персональных данных в форме, установленной ч. 4 ст. 9 Федерального закона № 152-ФЗ.
* при обработке персональных данных должны быть обеспечены точность персональных данных, их достаточность, а в необходимых случаях и актуальность по отношению к целям обработки персональных данных. Оператор должен принимать необходимые меры либо обеспечивать их принятие по удалению или уточнению неполных или неточных данных;
* хранение персональных данных должно осуществляться в форме, позволяющей определить субъекта персональных данных, не дольше, чем этого требуют цели обработки персональных данных;
* обрабатываемые персональные данные подлежат уничтожению либо обезличиванию по достижении целей обработки или в случае утраты необходимости в достижении этих целей.

1. **Ответственность за нарушение законодательства в области ПД**

За нарушение законодательства в области персональных данных для физических и юридических лиц предусмотрена как административная, так и уголовная ответственность. С некоторыми наиболее «популярными» статьями можно ознакомить на слайде.

1. **Потенциальное решение – обезличивание ПД**

Ученые предложили множество методов обеспечения защиты конфиденциальности при анализе данных. Популярным методом является удаление личных данных или замена случайными значениями перед анализом данных. Как правило, такие детали, как номера телефонов и почтовые индексы, анонимны. Однако анонимных данных не всегда достаточно для удовлетворения требований. Когда злоумышленник получает вспомогательную информацию о лицах, представленных в наборе данных, конфиденциальность, обеспечиваемая этой анонимной операцией, будет значительно снижена. Таким образом, сложно определить и защитить конфиденциальность, а также оценить объем информации, которую может получить злоумышленник.

Разобрать определение обезличивания ПД со слайда. В РФ процедура по обезличиванию персональных данных регламентирована Приказом Роскомнадзора от 5 сентября 2013 г. № 996 "Об утверждении требований и методов по обезличиванию персональных данных". Так, в соответствии с Приказом, к наиболее перспективным и удобным для практического применения относятся следующие методы обезличивания:

* введение идентификаторов – замена части сведений идентификаторами с созданием таблицы соответствия идентификаторов исходным данным;
* изменение состава или семантики – изменение состава или семантики персональных данных путем замены результатами статистической обработки, обобщения или удаления части сведений;
* декомпозиция – разбиение множества персональных данных на несколько частей с последующим раздельным хранением подмножеств;
* перемешивание – перестановка отдельных записей, а так же групп записей в массиве персональных данных.

Важно обратить внимание, что есть различие между обезличенными и анонимизированными данными. Полностью анонимизированные данные представляют собой статистику, которая доступна в свободном доступе и относится к открытым данным (например, статистика Росстата и соцопросы). Такие данные не несут той же ценности, как обезличенные, на основании которых можно определить некоторые особенности разных видов социальных групп. Анализ поведенческих особенностей малых социальных групп, прогнозирование возрастных трендов, измерение настроения людей и определение их отношения к тем или иным явлениям – все эти функции на основании анонимизированных данных невозможны. Другими словами, полностью анонимизированные данные не представляют ценности для бизнеса, а для некоторых областей искусственного интеллекта даже обезличенные данные не представляют ценности – для его обучения требуется опыт, а если такой опыт с пробелами, его обучение будет соответственным.

Подробнее - <https://www.garant.ru/news/1464529/>

1. **Потенциальное решение – Федеративное обучение**

В 2017 году компания Google опубликовала очень интересную статью и пост в блоге, которые внесли значительный вклад в обсуждение этой темы. В Google предположили, что для обучения моделей не нужен централизованный набор данных. Компанией было предложено рассмотреть вопрос: что если вместо сбора данных в одном месте, попытаться перенести модель в данные? Этот новый и увлекательный раздел машинного обучения получил название федеративное обучение. Рассмотреть определение.

Эта простая перестановка имеет чрезвычайно большое значение. Во-первых, это означает, что для участия в цепочке глубокого обучения людям не нужно отправлять свои данные кому бы то ни было. Ценные модели в здравоохранении, управлении личными активами и других чувствительных областях можно обучать без необходимости раскрывать личную информацию.

Основная идея приведена на изображении из блога Google. Кратко: на девайс загружается модель, дообучается (обновляет коэффициенты модели) на данных конкретного пользователя. Эти изменения модели собираются в некий апдейт небольшого размера, которая отправляется в облако с использование зашифрованного соединения. Затем эти изменения вливаются в основную модель. При этом все данные, на которых обучается модель не покидают пределы пользовательского устройства.

Подробнее - <https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>

1. **Пример Федеративного обучения – Gboard**

Первое крупномасштабное применение федеративного обучения нашлось в Gboard. При старом подходе машинного обучения разработка более совершенных предсказаний клавиатуры была бы чрезвычайно инвазивной - все, что мы печатали, все наши личные сообщения и странные поиски в Google должны были быть отправлены на центральный сервер для анализа.

Федеративный подход работает по-другому. При выводе предлагаемых слов для ввода смартфон локально хранит информацию о текущем контексте и выбранных предложениях. С помощью федеративного обучения эти накопленные данные обрабатываются локально на устройстве с целью улучшения работы модели на данном устройстве.

На устройстве для дообучения используется миниатюрная версия TensorFlow. Дообучения осуществляется только когда телефон не используется, подключен к зарядке. Для передачи используются только не тарифицируемые сети. Это сделано для сведения к минимуму неудобств, доставляемых пользователю.

Подробнее - <https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>

1. **Федеративное обучение для сферы здравоохранения**

В 2018 году Intel установила партнерские отношения с Центром биомедицинских вычислений и аналитики изображений в Пенсильванском университете, чтобы продемонстрировать, как федеративное обучение может быть применено к медицинским данным в качестве доказательства концепции.

Сотрудничество показало, что в рамках федеративного подхода к обучению их конкретная модель глубокого обучения может быть подготовлена ​​с точностью 99% по сравнению с той же моделью, обученной традиционными методами.

1. **Пример практического использования федеративного обучения. NVIDIA**

Исследователи из NVIDIA и Королевского колледжа Лондона использовали федеративную архитектуру клиент-сервер с центральным сервером для поддержки глобальной глубокой нейронной сети. При таком подходе участвующим больницам будет предоставлена ​​копия их нейронной сети для обучения по их собственному набору данных.

Подробнее - <https://servernews.ru/995564>

1. **Безопасное агрегирование. Google Secure Aggregation**

Подход федеративного обучения предполагает агрегирование данных моделей от большого количества пользователей. В общем случае параметры модели от конкретного пользователя могут отображать специфичную информацию о пользователе. Поэтому встает вопрос безопасного агрегирования получаемых данных. Основная идея в том, чтобы усреднять веса до того, как кто-то сможет их увидеть (в том числе и сервер).

Есть два основных пути, используемых на практике. В действительности их больше, но рассмотрим основные.

Протокол безопасной агрегации использует многоступенчатые вычисления для определения среднего значения группы сводок пользовательских данных, не раскрывая сводки данных какого-либо отдельного лица на сервере или любой другой стороне.

В этой системе каждая из пользовательских сводок шифруется перед тем, как покинуть пользовательское устройство, и они не могут быть расшифрованы сервером до тех пор, пока они не будут добавлены вместе и усреднены с заданным числом других пользовательских сводок. Это позволяет серверу обучать свою модель в среднем по пользователю, не раскрывая отдельных сводок, которые могут быть использованы для раскрытия личных данных отдельных лиц.

Secure Aggregation не только предотвращает доступ сервера к пользовательским сводкам, но также человек посередине атаки гораздо сложнее.

При разработке были учтены особенности мобильных устройств: узкий канал передачи данных и возможные частые сбои (пользователь выключил телефон и т.д.).

Теоретическая основа - схема разделения секрета Шамира. Схема Шамира позволяет реализовать (k,n) — пороговое разделение секретного сообщения (секрета) между n сторонами так, чтобы только любые k и более сторон k ≤ n) могли восстановить секрет. При этом любые k-1 и менее сторон не смогут восстановить секрет.

Кратко рассмотреть алгоритм, изображенный на картинке. Подробнее про Google Secure Aggregation - <https://eprint.iacr.org/2017/281.pdf>

1. **Дифференциальная приватность. Определение**

Другой способ – использование модели дифференциальной приватности.

Дифференциальная приватность — это математическое определение понятия «наличия приватности». Это не какой-то конкретный процесс, а, скорее, свойство, которым может обладать процесс. Например, можно рассчитать (доказать), что данный конкретный процесс удовлетворяет принципам дифференциальной приватности.

Этот термин был введён Синтией Дворк в 2006 году. Рассмотреть иллюстрацию на слайде, вывод по ней приведен в рамке. Кто бы ни посмотрел на результаты, он не сможет сказать, в каком случае использовались данные Ивана, а в каком не использовались.

Подробнее - <https://habr.com/ru/company/domclick/blog/526724/>

1. **Пример потребности в дифференциальной приватности**

Предположим, что у нас есть база данных медицинских записей D1, где каждая запись представляет собой пару (Имя, X), где X является нулём или единицей, обозначающим, имеет ли человек гастрит или нет. Теперь предположим, что злоумышленник хочет найти, имеет ли Михаил гастрит или нет. Также предположим, что он знает, в какой строке находится информация о Михаиле в базе данных. Теперь предположим, что злоумышленнику разрешено использовать только конкретную форму запроса Qi, который возвращает частичную сумму первых i строк столбца X в базе данных. Чтобы узнать, есть ли гастрит у Михаила, злоумышленник выполняет запросы: Q4(D1) и Q3(D1), затем вычисляет их разницу. В данном примере, Q4(D1)=3, а Q3(D1)=2, поэтому их разность равна 1. Это значит, что поле «Наличие гастрита» в строке Михаила должно быть равно 1. Этот пример показывает, как индивидуальная информация может быть скомпрометирована даже без явного запроса данных конкретного человека.

Если бы злоумышленник получал значения Qi через ε-дифференциально приватный алгоритм, для достаточно малого ε, то он не смог бы отличить два набора данных по определению дифференциальной конфиденциальности, так как Q4(D1) и Q3(D1) были бы неотличимы.

Посмотрим, как на практике может быть реализована дифференциальная конфиденциальность.

1. **Формальное определение дифференциальной приватности**

Мы контролируем требуемый уровень приватности через изменение параметра приватности ε, который также называют потерей приватности (privacy loss) или бюджетом приватности (privacy budget). Чем меньше значение ε, тем менее различимы результаты и тем больше защищены данные отдельных людей. Рассмотреть определение.

В соответствии с этим определением дифференциальная приватность является условием механизма публикации данных (то есть определяется доверенной стороной, выпускающей информацию о наборе данных), а не самим набором. Интуитивно это означает, что для любых двух схожих наборов данных, дифференциально-приватный алгоритм будет вести себя примерно одинаково на обоих наборах. Определение также даёт сильную гарантию того, что присутствие или отсутствие индивидуума не повлияет на окончательный вывод алгоритма.

Случай, когда ε = 0, является идеальным для сохранения конфиденциальности, поскольку наличие или отсутствие любой информации о любом человеке в базе данных никак не влияет на результат алгоритма, однако такой алгоритм является бессмысленным с точки зрения полезной информации, так как даже при нулевом количестве людей он будет давать такой же или подобный результат.

Подробнее - <https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%84%D1%84%D0%B5%D1%80%D0%B5%D0%BD%D1%86%D0%B8%D0%B0%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D0%BF%D1%80%D0%B8%D0%B2%D0%B0%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C>

1. **Реализация дифференциальной приватности – метод Лапласа**

В связи с тем, что дифференциальная приватность является вероятностной концепцией, любой её метод обязательно имеет случайную составляющую. Некоторые из них, как и метод Лапласа, используют добавление контролируемого шума к функции, которую нужно вычислить.

Разобрать вывод формулы на слайде.

Если мы попытаемся использовать эту концепцию в вышеприведённом примере про наличие гастрита, то ∆𝑓=1. То есть 𝛼 явно определяет параметр приватности.

Кроме шума Лапласа также можно использовать другие виды шума (например, гауссовский), но они могут потребовать небольшого ослабления определения дифференциальной приватности.

1. **Реализация дифференциальной приватности – пример**

Дифференциальная приватность основана на введении случайности в данные.

Простой пример заключается в том, чтобы попросить человека ответить на вопрос «Есть ли у вас атрибут А?» в соответствии со следующей процедурой:

* Человек подбрасывает монету
* Если выпал орел, отвечает честно на вопрос.
* Иначе подбрасывает ещё раз, если выпадет орел, ответь «Да», если решка — «Нет»

Конфиденциальность возникает, так как невозможно по ответу точно узнать, обладает ли человек данным атрибутом. Но тем не менее эти данные значительны, так как положительные ответы дают четверть от тех людей, у которых нет этого атрибута, и три четверти от тех, кто на самом деле им обладают. Таким образом, если p — истинная доля людей с A, то мы ожидаем получить (1/4) (1- p) + (3/4) p = (1/4) + p / 2 положительных ответов. Следовательно, можно оценить р.

То есть общая закономерность сохраняется, ценность выборки не теряется. Однако достоверно определить наличие атрибута А у конкретного человека не представляется возможным.

1. **PATE – алгоритм, реализующий дифференциальную приватность**

Локальные модели тренируются независимо друг от друга (без какого-либо взаимодействия). Каждая модель обучается на своем наборе данных.

Приватный набор данных разделяется на подмножество данных (разделов). На каждом из этих данных учится своя модель. Метод обучения модели учителя неограничен, что также является одним из основных преимуществ PATE.

Подробнее - <https://arxiv.org/pdf/2004.06567.pdf>

1. **PATE. Предсказание**

Как использовать дальше эти отдельно обученные модели? Каждая модель выполняет предсказание, для того чтобы внести конфиденциальность используется подход дифференциальной приватности, а именно добавляется случайный шум Лапласа или Гаусса (для нарушения статистики).

Однако в таком подходе есть явные проблемы. Во-первых, каждый прогноз, сделанный механизмом агрегирования, увеличивает набор статистики, в далекой перспективе снижая конфиденциальность. Во-вторых, набор моделей учителей не может быть опубликован с открытым исходным кодом, в противном случае злоумышленник может проверить опубликованные параметры модели, чтобы узнать об обучении.

1. **PATE. Модель студента**

Для решения описанных выше проблем создается дополнительная модель студента, которая обучается на за счет наличия моделей учителей. Этот процесс отображен на слайде, рассмотрим его подробнее. Модель студента обучается с соблюдением конфиденциальности путем интеграции знаний, полученных учителями. Модель студента выбирает входные данные из набора немаркированных общедоступных данных и отправляет эти входные данные в модели учителей для получения меток, а затем модель ученика использует помеченные данные для обучения. Защита конфиденциальности и правильность тегов, предсказываемых механизмом агрегирования, являются результатом консенсуса, достигнутого среди учителей.

1. **Проблемы и ограничения федеративного обучения**

Федеративное обучение страдает двумя большими проблемами, особенно трудноразрешимыми, когда у каждого человека имеется лишь маленькая горстка обучающих примеров, — скорость и конфиденциальность.

Как оказывается, если у кого-то имеется лишь несколько обучающих примеров (или модель, присланная вам, была обучена лишь на нескольких примерах: обучающем пакете), вы все еще можете довольно много узнать об исходных данных. Если представить, что у вас есть 10 000 человек (и у каждого имеется очень небольшой объем данных), большую часть времени вы потратите на пересылку модели туда и обратно и не так много — на обучение (особенно если модель очень большая).

Перехват при передаче данных уточненной модели – простое решение: шифрование.

«Отравление» модели: злоумышленник может испортить модель через свое собственное устройство или путем захвата устройств других сторон, участвующих в обучении алгоритмической модели. Это актуальная проблема, особенно для протокола безопасной агрегации Google, так как там мы не сможем вычислить злоумышленника, чтобы перестать принимать от него данные.

Компании должны защищать свою интеллектуальную собственность, и похоже, что отправка модели напрямую на устройства пользователей может легко привести к тому, что эти модели будут выставлены любому желающему. Однако есть решения, которые компании могут использовать для защиты своих алгоритмических моделей.

Одним из них является использовать секретный обмен многопартийных вычислений. Это позволяет организациям скрывать взвешивание модели, распределяя ее фрагменты по устройствам. В рамках этой системы ни одна из секретных сторон не может знать всю модель.

Это позволяет организациям передавать свои алгоритмические модели обучения на устройства, не беспокоясь о краже их интеллектуальной собственности.

Обсудим ограничения, которые накладывает федеративное обучение. Во-первых, конечные устройства должны обладать достаточной вычислительной мощностью и памятью. Именно поэтому 10-15 лет назад идея федеративного обучения была просто нереализуемой. Современные носимые устройства уже обладают достаточным потенциалом.

Другое техническое ограничение связано с пропускной способностью. Федеративное обучение проводится через Wi-Fi или 4G, в то время как традиционное машинное обучение происходит в центрах обработки данных. Пропускная способность Wi-Fi или 4G на порядок ниже, чем между рабочими узлами и серверами в этих центрах. Это узкое место, которое увеличивает задержку и замедляет процесс обучения по сравнению с традиционным подходом.

Несмотря на попытки оптимизировать время запуска обучения это может вызвать замедление работы устройства (например, когда пользователь неожиданно для системы снимает телефон с зарядки и начинает им активно пользоваться).

1. **Достоинства федеративного обучения**

Федеративное обучение может также помочь организациям улучшить свои алгоритмические модели, не раскрывая данные клиента и не нарушая законодательные нормы. Законы, такие как Европейское общее положение о защите данных (GDPR) и Закон о переносимости медицинского страхования США от 1996 года, ФЗ-152 содержат строгие правила в отношении данных отдельных лиц и способов их использования.

Другая ключевая причина принятия федеративного подхода к обучению заключается в том, что он потенциально может снизить задержку. В вероятном будущем сценарии, когда на наших дорогах будет большое количество автомобилей с автоматическим управлением, они должны быть в состоянии быстро реагировать друг на друга во время происшествий, связанных с безопасностью.

Традиционное облачное обучение включает в себя передачу больших объемов данных и более медленный темп обучения, поэтому существует вероятность того, что федеративное обучение может позволить автономным транспортным средствам действовать быстрее и точнее, уменьшая количество аварий и повышая безопасность.

В завершении лекции еще раз напомнить основные моменты и прокомментировать картинку, демонстрирующую отличия традиционного обучения, федеративного обучения и федеративного обучения с безопасной агрегацией.

1. **Финальный слайд**

Спасибо за внимание!

**Перечень рекомендованной литературы.**

1. Dominik, Jens Sjlund, Tobias J. Oechtering. Decentralized Differentially Private Segmentation with PATE. <https://arxiv.org/pdf/2004.06567.pdf>
2. Э. Траск. Грокаем машинное обучение (глава 16).
3. Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical Secure Aggregation for Privacy-Preserving Machine Learning. <https://eprint.iacr.org/2017/281.pdf>
4. Блог SecurityLab. Виды информации ограниченного доступа. <https://www.securitylab.ru/blog/personal/aguryanov/29908.php>
5. Информационно-правовой портал «Гарант.ру». Обезличивание данных: сохранение баланса между правами граждан и развитием инноваций. <https://www.garant.ru/news/1464529/>
6. Google AI blog. Federated Learning: Collaborative Machine Learning without Centralized Training Data. <https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>
7. Новостной портал servernews.NVIDIA использовала федеративное машинное обучение при создании ИИ для здравоохранения <https://servernews.ru/995564>
8. Портал «Хабр». Дифференциальная приватность — анализ данных с сохранением конфиденциальности. <https://habr.com/ru/company/domclick/blog/526724/>
9. [Википедия. Дифференциальная приватность. https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%84%D1%84%D0%B5%D1%80%D0%B5%D0%BD%D1%86%D0%B8%D0%B0%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F\_%D0%BF%D1%80%D0%B8%D0%B2%D0%B0%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C](Википедия.%20Дифференциальная%20приватность.%20https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%84%D1%84%D0%B5%D1%80%D0%B5%D0%BD%D1%86%D0%B8%D0%B0%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D0%BF%D1%80%D0%B8%D0%B2%D0%B0%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C)