**Искусственный интеллект (Передовые технологии) (семинары)**

**Урок 3. Создание проекта машинного обучения**

**Задание к домашней работе № 3:**

1. Подумайте о реальной проблеме, которую, по вашему мнению, можно решить с помощью машинного обучения. Чётко определите эту проблему и поставьте SMART-цели для её решения. Помните, что это должна быть проблема, в которой машинное обучение может помочь сделать прогноз, классифицировать данные или дать рекомендации.

2. Определите потенциальные источники данных, которые могут быть использованы для решения вашей проблемы. Какие данные вам понадобятся? Где вы можете их получить? Напишите краткое резюме вашей стратегии сбора данных.

3. Даже если у вас нет реальных данных для анализа, представьте, что они у вас есть. Как бы вы стали исследовать и понимать эти данные? Что бы вы искали? Напишите краткое описание вашей стратегии анализа исследовательских данных.

4. Вспомните различные модели машинного обучения, которые мы обсуждали, в частности модель "Дерево решений". Какой тип модели машинного обучения будет наиболее подходящим для вашей проблемы и почему? Объясните свой выбор.

1. Напишите краткое описание того, как вы будете обучать и оценивать выбранную вами модель машинного обучения, используя воображаемый набор данных.
2. Наконец, подумайте о том, как вы могли бы развернуть свою обученную модель. Где она будет использоваться? Кто будет ее использовать? Какую пользу она принесёт им? Напишите краткое резюме вашей стратегии развёртывания.

Помните, что целью этого задания является не создание реальной модели машинного обучения (пока!), а применение концепций, которые вы изучили на этом семинаре, к реальной проблеме.

**Выполнение:**

**1. SMART-цели для решения проблемы диагностики рака головного мозга с помощью ИИ**

Проблема диагностики рака головного мозга с помощью ИИ является сложной задачей, требующей анализа больших объёмов медицинских изображений. Машинное обучение может помочь классифицировать снимки для выявления опухолей, прогнозировать вероятность их злокачественности, давать рекомендации врачам по дальнейшим действиям и снизит нагрузку по объёмам работы на врачебный персонал.

SMART-цели:

S (Specific) — Концепция: Цель должна быть чётко и ясно сформулирована.

Разработать модель машинного обучения, которая сможет классифицировать МРТ изображения головного мозга с целью выявления наличия или отсутствие опухоли и предсказания из вероятной злокачественности.

M (Measurable) — Измеримая: Цель должна включать конкретные показатели или критерии, по которым можно оценить прогресс и успех.

Достичь точности классификации на уровне 95% на тестовом наборе данных с минимумом ложноположительных и ложноотрицательных диагнозов не более 5%.

A (Achievable) — Достижимая: Цель должна быть реальной и достижимой с учетом имеющихся ресурсов и ограничений.

Обучить модель на наборе данных из 10,000 размеченных МРТ снимков в течение 6 месяцев с использованием доступных вычислительных ресурсов и открытых библиотек для машинного обучения, TensorFlow, PyTorch, Keras, Scikit-learn.

R (Relevant) — Значимая: Цель должна быть значимой и иметь отношение к общим целям организации или проекта.

Улучшить процесс диагностики рака головного мозга в больницах, снижая время на постановку диагноза на 20% и облегчая нагрузку на врачебный персонал.

T (Time-bound) — Ограниченная по времени: Цель должна иметь чёткие сроки выполнения, которые позволяют контролировать процесс достижения.

Достичь внедрения модели в клиническую практику через 12 месяцев после начала разработки, включая этапы обучения, валидации, тестирования и интеграции с медицинскими системами.

**2. Потенциальные источники данных**

Источники данных медицинские изображения (МРТ), высококачественные МРТ снимки головного мозга пациентов, содержащие как опухоли, так и снимки без них. Для обучения модели потребуются метки, указывающие на наличие или отсутствие опухолей, а также их тип (доброкачественная или злокачественная) а также информация о пациенте, такая как возраст, пол, медицинская история и результаты биопсий, для дополнительного анализа и улучшения точности прогноза.

Получить данные можно в медицинские учреждениях и больницах при партнёрстве с крупными медицинскими учреждениями, которые могут предоставить доступ к анонимизированный данным пациентов. Из открытых баз данных типа Kaggle, сети медицинских баз данных с МРТ снимками, такие как TCIA (The Cancer Imaging Archive), BraTS (Brain Tumor Segmentation Challenge).

Краткое резюме стратегии сбора данных:

1) Начать с поиска доступных открытых баз данных (например, Kaggle) и загрузить необходимые наборы данных для предварительного обучения.

2) Установить сотрудничество с локальными медицинскими учреждениями для получения дополнительного набора анонимизированный МРТ снимков.

3) Обеспечить защиту данных пациентов и соблюдение норм конфиденциальности (например, HIPAA).

4) Построить безопасную инфраструктуру для хранения и обработки данных, чтобы поддерживать высокие стандарты безопасности.

**3. Исследование и анализ данных**

Просмотр и анализ МРТ изображений с помощью библиотек для визуализации медицинских изображений (например, Matplotlib, SimpleITK). Анализ распределения классов наличие/отсутствие опухоли, сравнение между различными возрастными группами и типами опухолей. Поиск взаимосвязей между наличием опухоли и клиническими характеристиками пациентов, чтобы выявить возможные паттерны. Различия между здоровыми снимками и снимками с опухолями. Типичные признаки злокачественных и доброкачественных опухолей. Возможные источники шума или артефактов в изображениях, которые могут повлиять на обучение модели. Возможнный дисбаланс классов, который может потребовать применения методов, таких как oversampling или undersampling.

Краткое описание стратегии анализа исследовательских данных:

1) Провести исследовательский анализ данных для понимания структуры, особенностей набора данных, размера и качества.

2) Применить методы визуализации данных для лучшего понимания их распространения, распределения, определить различия между здоровыми и больными тканями.

3) Разработать набор правил предобработки изображений, например, нормализация, аугментация.

4) Выявить ключевые переменные и признаки, которые могут влиять на точность модели, для последующей оптимизации.

**4. Выбор модели машинного обучения**

Модель "Дерево решений" — это одна из популярных моделей машинного обучения, которая работает, создавая последовательные "ветвления" на основе признаков данных для принятия решений. Эта модель интерпретируема и хорошо работает с задачами классификации и регрессии, где связи между признаками и целевыми переменными достаточно просты.

Однако в нашем случае наиболее подходящим тип это свёрточная нейронная сеть (Convolutional Neural Network «CNN») класс глубоких нейронных сетей, которые особенно эффективны для обработки данных с определённой структурой, таких как изображения, видео и временные ряды. Основное отличие CNN от других нейронных сетей заключается в использовании свёрточных слоёв, которые автоматически извлекают значимые признаки features из входных данных. Выбрана по следующим критериям:

1) Медицинские изображения головного мозга, такие как МРТ или КТ, содержат сложные паттерны и аномалии. Глубокие нейронные сети способны автоматически извлекать сложные признаки и анализировать изображения с высокой точностью, тогда как более простые модели, такие как деревья решений, могут не справиться с такой объёмной и сложной информацией.

2) Является мощным инструментом для анализа медицинских изображений, поскольку они могут находить тонкие различия в текстуре и структуре тканей мозга, которые могут свидетельствовать о наличии и границах опухоли.

3) Глубокие сети могут выполнять диагностику без ручного выделения признаков, что ускоряет процесс и снижает вероятность человеческого фактора в случае если работу выполняет не квалифицированный в медицине специалист.

4) Благодаря масштабируемости отлично справиться с анализом большого количества медицинских изображений.

**4.1 Обучение и оценивание модели**

Сначала подготовим данные для этого соберём данные в нашем случаем МРТ снимки разделённые или разделим их на метки наличие или отсутствие опухоли. Разобьём данные на тренированный, валидационный и тестовый наборы с соотношением 70% 15% 15% и нормализуем изображения.

Используем конволюционную нейронную сеть (CNN) с несколькими сверточными и полносвязанными слоями. Примерная архитектура может включать в себя сверточные слои для извлечения признаков, слои pooling для уменьшения размерности, и полносвязные слои для классификации.

Определим функцию потерь, например, кросс-энтропия для классификации и оптимизатор, например Adam. Обучаем модель на тренировочном наборе с использованием метода обратного распространения ошибки. Во время обучения модель проходит регулярную оценку на валидационном наборе, чтобы избежать переобучения.

После завершения обучения оцениваем модель на тестовом наборе. Метрики, такие как точность accuracy, полнота recall, точность precision и F1-score, помогут оценить, насколько хорошо модель справляется с задачей диагностики, использовать ROC-кривую и AUC площадь под кривой для оценки способности модели различать больных и здоровых пациентов. После успешного обучения и тестирования модели, будет развёртывание.

**4.2 Развёртывание использование модели**

Модель может быть развёрнута в медицинских учреждениях, где анализируется снимки пациентов: больницы, поликлиники, частные медицинские центры. Использовать её будут врачи-радиологи они смогут использовать модель для автоматической интерпретации изображений головного мозга, чтобы ускорить процесс диагностики и в некоторых случаях повысить точность выявления опухолей и для проведения первичной фильтрации данных перед проведением более детальных анализов.

Модель может автоматически анализировать изображения и выявлять признаки опухоли быстрее, чем ручной анализ, что важно для своевременного начала лечения. Модель может обнаружить микроскопические аномалии, которые могут быть не заметны для человеческого глаза. А так же модель может служить инструментом для поддержки принятия решений, что позволит врачам сосредоточиться на более сложных случаях.

Краткое резюме стратегии развёртывания:

1) Начальное развёртывание модели будет происходить в тестовом режиме сначала в одном медицинском учреждении. Врачи будут проверять результаты модели и сравнивать их с результатами ручной диагностики.

2) Модель будет интегрирована с системами медицинских изображений с переработанным пользовательским интерфейсом для автоматического анализа новых снимков и создания отчётов.

3) Получение отзывов от врачей и настройка модели на основе их рекомендаций и новых данных.

4) Организация тренингов для врачей и технических специалистов по правильному использованию и интерпретации результатов модели.

5) После успешного тестирования в модель будет развёрнута дополнительно в нескольких учреждениях, позже модель может быть развёрнута в более широком масштабе, чтобы обеспечить доступ к высокоточной диагностике большему числу пациентов.

6) После окончания тестированного развёртывания и решения возникших проблем в ходе развёртывания, запустить модель в массовое внедрение не только в больницы но и в профильные поликлиники, далее предложить модель частным учреждениям медицинского назначения как готовый продукт оптимизации расходов и увеличивание прибыли.