# «Современные методы машинного обучения»

Вопросы для экзамена по курсу

## Тема 1. Задачи машинного обучения

1. Основные понятия машинного обучения (Модель, Алгоритм, признак, обучающая и тестовая выборка). Постановки задач машинного обучения.
2. Классификация и кластеризация. Примеры моделей: Наивный байесовский классификатор, метод k-средних
3. Проблемы доверенного ИИ. Основные классы дефектов машинного обучения (информационная безопасность, функциональная надёжность, социогуманитарные аспекты). Примеры угроз и их важность для доверенного ИИ

## Тема 2. Введение в нейронные сети

1. Основные понятия нейронных сетей: искусственный нейрон, функция активации, полносвязный слой, архитектура многослойной нейронной сети.
2. Постановка задачи обучения с учителем. Основные функции потерь и их назначение. Градиентный спуск: идея метода, влияние шага обучения на сходимость, проблема выбора оптимального шага. Стохастический и пакетный градиентный спуск: особенности, преимущества и ограничения. Граф вычислений.
3. Обработка изображений с помощью нейросетей. Проблема применения полносвязных сетей к данным высокой размерности. Архитектура свёрточных нейронных сетей. Операция двумерной свёртки, шаг и дополнение нулями, формирование карт признаков, уменьшение размерности через свёртку или пулинг.

## Тема 3. Математика полносвязной сети: геометрия, интерпретируемость, SLAP атака

1. Геометрическая интерпретация многослойного персептрона: разбиение входного пространства на ячейки первым и последующими слоями, роль знаков линейных неравенств нейронов, линейность модели внутри каждой ячейки, связь глубины сети с числом ячеек.
2. Функция регрессии. Классический байесовский классификатор. Проблема поведения вне носителя. Модифицированный Байесовский классификатор. Отказ от принятия решения. Нейросетевая аппроксимация байесовского решения. Понятие состоятельности.
3. Построение объясняющего дерева eXBTree. Интерпретация решений модели через анализ ячеек. SLAP атака: постановка задачи, однослойный и многослойный случаи, ограничения и уязвимости.

## Тема 4. Атаки на модели машинного обучения и методы защиты

1. Градиентные атаки: FGSM и PGD. Укажите преимущества и ограничения методов. Опишите различия между целевыми и нецелевыми атаками. Как

необходимо модифицировать алгоритм FGSM, чтобы он работал в режиме целевой атаки?
2. Атаки чёрного ящика и переносимость. Опишите, как строятся атаки чёрного ящика в условиях ограниченного доступа к модели. Score-based и decision-based подходы. Атака черного ящика на основе запросов OPT.
3. Методы защиты от состязательных атак. Опишите основные классы методов защиты: детектирование атакованных примеров, преобразования входных данных и adversarial training. Объясните сильные и слабые стороны каждого класса и почему ни один из методов не дает полной гарантии безопасности.

## Тема 5. Задачи обработки изображений. Атаки на методы обработки изображений

1. Атака с помощью дистилляции/аппроксимации целевой нейросети. Опишите основную идею и выпишите лосс-функцию.
2. Атаки на метрики качества видео. В чем сложность атак на видео и как можно ее обойти. Методы атак на метрики Zhang et al., Korhonen et al., SSAH, UAP. Основная идея метода IOI.
3. Методы атак без ограничений на состязательный шум: основные идеи (cAdv, диффузионная). Физическая атака TI-Patch.

## Тема 6. Диффузионные модели: атаки и защита

1. Прямой и обратный диффузионные процессы в DDPM. Общая схема обучения и инференса DDPM (качественно). Основные отличия DDPM и DDIM.
2. Latent Diffusion Models (LDM). Ключевые особенности устройства LDM и общая схема их работы.
3. Принцип работы метода SDEdit для редактирования изображений. Метод редактирования изображений Prompt-to-prompt, его основная идея.
4. *Что такое h-space в диффузионных моделях и как он используется для интерпретируемости? Как определяется h-space?

## Тема 7. Доказательства в машинном обучении. Новые результаты и открытые проблемы

1. Определение переносимой состязательной атаки. Определение и описание процесса дистилляции знаний. Определение липшицевой нейронной сети и связь липшицевости с устойчивостью нейронной сети к аддитивным возмущениям.
2. Применение дистилляции знаний для построения переносимого состязательного возмущения в случае липшицевых сетей, основная идея доказательства.
3. Предсказания нейронной сети как случайные величины (интерпретация, когда возникает, выборочные оценки математического ожидания). Примеры процедур оценки вероятности больших отклонений случайной величины (качественно). Связь вероятности больших отклонений и устойчивости нейронной сети в среднем.

## Тема 8. Медицинские приложения, ЭКГ. Устойчивость.

1. Задача автоматического анализа ЭКГ: постановка проблемы, методы сегментации сигнала и основные подходы к задаче классификации

2. Методы самообучения и контрастного обучения для временных рядов на примере ЭКГ
3. Федеративное обучение в медицине: алгоритмы, угрозы безопасности, византийские атаки

## Тема 9. GAN. Создание и детекция  дипфеков

1. Основные компоненты StyleGAN. Цикл обучения: обучение генератора, обучение дискриминатора.
2. Диффузия: прямой проход, обратный проход. Classifier guidance.
3. Признаки, по которым можно обнаружить дипфейки (пространственные, темпоральные).

## Тема 10. Маркировка данных. Борьба с дипфейками

1. Основные цели и задачи маркировки. Основные цели и задачи водяных знаков. Сходства и отличия маркировки и цифровых водяных знаков. Основные свойства ЦВЗ. Основные типы ЦВЗ, для каких задач какие типы ЦВЗ применяются.
2. Основные метрики оценки незаметности ЦВЗ для изображений. Алгоритм идентификации пользователей, выполняющих незаконную ретрансляцию онлайн-трансляции.
3. Алгоритмы обеспечения контроля целостности изображений и видео для обнаружения дипфейков, создаваемых на их основе.

## Тема 11. Дрейф данных. OOD. Инструменты MLOps.

1. Многообразие задач, относящихся к непрерывному обучению, Разница в постановках задач AD/ND/OOD/OSR/OD, метрики для всех задач.
2. Основные подходы к Inference-time детекции примеров вне распределения, смысл и вывод Energe-score, его преимущества.
3. Основные подходы к Train-time детекции примеров вне распределения, подходы к получению и генерации OOD-выборок.

## Тема 12. Обучение LLM. RuAdapt

1. Что такое задача языкового моделирования. Чем отличается трансформер-декодер от трансформера-энкодера с точки зрения архитектуры. Какие основные этапы обучения больших языковых моделей.
2. Что такое zero-shot, few-shot. Идея FLAN. Chain-of-thought
3. В чем оценивают “стоимость” обучения языковых моделей. Что важнее, количество данных или размер модели. Каким образом текст подается в языковые модели. Основная идея и шаги методологии Ruadapt.