Постановка задачи классификации и требования к обучающей выборке
Слайд 1: Заголовок и ваше имя
Начало выступления (0-1 минута):
Классификация – это одна из ключевых задач машинного обучения, и понимание ее основ и требований к данным критически важно для создания эффективных моделей. В моем докладе мы разберем, что такое задача классификации, какие типы классификации существуют, как оценивается качество моделей классификации и, самое главное, какие требования предъявляются к данным, на которых эти модели обучаются."
Слайд 2: Что такое задача классификации (1-3 минуты):
(На экране: Заголовок "Что такое задача классификации?", основные тезисы и примеры)
"Итак, что же такое задача классификации? В самом простом понимании, это задача разделения объектов на заранее определенные классы. Представьте, что у нас есть набор объектов, например, фотографии, и нам нужно отнести каждую фотографию к одной из категорий: "кошка", "собака", "птица" и так далее. Или, как в примере из текста, отделить спам от не-спама, или легальную транзакцию от мошеннической.
В терминах машинного обучения, у нас есть множество объектов X и множество классов Y. Мы предполагаем, что существует неизвестная функциональная зависимость f, которая связывает объекты и их классы. Наша цель – построить алгоритм, который сможет приблизить эту функцию f на основе обучающей выборки S.
Обучающая выборка S – это набор пар "объект – ответ", где для каждого объекта xi из X известен его класс yxi из Y. Задача классификации заключается в том, чтобы для нового, неизвестного объекта, алгоритм смог предсказать его класс.
Важно отметить, что множество классов Y всегда конечно и дискретно. Каждый класс соответствует определенной категории объектов."
Слайд 3: Типы классификации (3-4 минуты):
(На экране: Заголовок "Типы классификации", описание и примеры для каждого типа)
"Классификация бывает разных видов, в зависимости от количества классов и типа предсказания. Основные типы, как указано в материале, это:
Бинарная классификация: Здесь всего два класса, например, "да/нет", "0/1", "спам/не спам", "легальная/мошенническая транзакция". Это самый простой вид классификации, но очень распространенный.
Многоклассовая классификация: В этом случае классов три и более. Примеры: распознавание цифр от 0 до 9, классификация видов цветов, например, ирисы, розы, тюльпаны.
Детерминированная и Вероятностная классификация: Это различие касается типа предсказания. Детерминированная модель выдает четкий ответ, например: "это кошка". Вероятностная модель оценивает вероятность принадлежности к каждому классу, например: "кошка – 85%, собака – 15%". Вероятностные модели дают больше информации и позволяют лучше оценить уверенность модели в своем предсказании."
Слайд 4: Матрица ошибок (4-6 минуты):
(На экране: Заголовок "Матрица ошибок", таблица матрицы ошибок, описание TP, FP, TN, FN)
"Для оценки качества работы алгоритмов классификации используется матрица ошибок, особенно часто для бинарной классификации. Она показывает, насколько хорошо модель классифицирует объекты по классам.

Представьте себе задачу бинарной классификации, где "положительный класс" – это, например, "болен", а "отрицательный класс" – "здоров". В матрице ошибок мы выделяем 4 типа исходов:
True Positive (TP) – Истинно положительный: Модель правильно предсказала положительный класс. Например, модель сказала "болен" и пациент действительно болен.

False Positive (FP) – Ложноположительный (ошибка первого рода): Модель ошибочно предсказала положительный класс. Например, модель сказала "болен", а пациент здоров (ложная тревога).
True Negative (TN) – Истинно отрицательный: Модель правильно предсказала отрицательный класс. Например, модель сказала "здоров" и пациент действительно здоров.
False Negative (FN) – Ложноотрицательный (ошибка второго рода): Модель ошибочно предсказала отрицательный класс. Например, модель сказала "здоров", а пациент болен (опасная ошибка, пропуск болезни).
Важно понимать, что в разных задачах ошибки FP и FN могут иметь разную "цену". Например, в медицинской диагностике ошибка FN (пропуск болезни) может быть гораздо опаснее, чем ошибка FP (ложная тревога)."
Слайд 5: Метрики качества классификации (6-8 минуты):
(На экране: Заголовок "Метрики качества классификации", формулы и описания Accuracy, Precision, Recall, F1-score)
"На основе матрицы ошибок рассчитываются различные метрики качества классификации. Вот основные из них:
Accuracy (точность): Это доля правильных ответов от общего числа объектов. Формула: (TP + TN) / (TP + TN + FP + FN). Accuracy проста и понятна, но не всегда информативна, особенно если классы несбалансированы. Например, если 99% объектов относятся к одному классу, модель, которая всегда предсказывает этот класс, будет иметь accuracy 99%, но при этом будет абсолютно бесполезна.
Precision (точность, положительная предсказательная ценность): Показывает, какую долю объектов, отнесенных моделью к положительному классу, действительно являются положительными. Формула: TP / (TP + FP). Precision важна, когда мы хотим минимизировать ложноположительные ошибки (FP).
Recall (полнота, чувствительность): Показывает, какую долю из всех объектов положительного класса модель смогла правильно идентифицировать. Формула: TP / (TP + FN). Recall важна, когда мы хотим минимизировать ложноотрицательные ошибки (FN).
F1-score (F-мера): Это гармоническое среднее между precision и recall. Формула: 2 * (precision * recall) / (precision + recall). F1-score дает сбалансированную оценку качества модели, особенно полезен при несбалансированных классах. Существует также Fβ-мера, которая позволяет задать вес для precision и recall, где β контролирует важность recall относительно precision.
Выбор метрики зависит от конкретной задачи и приоритетов. Важно понимать, что нет универсальной "лучшей" метрики, и нужно выбирать ту, которая наиболее соответствует целям и особенностям задачи."
Слайд 6: Требования к обучающей выборке - Разметка данных (8-9 минуты):
(На экране: Заголовок "Требования к обучающей выборке", подзаголовок "Разметка данных", основные тезисы)

"Теперь перейдем к самому важному – требованиям к обучающей выборке. Качество обучающих данных – это ключевой фактор успеха в машинном обучении. Даже самый сложный алгоритм не сможет хорошо работать на плохих данных.
Первое и самое важное требование – это качественная разметка данных (labeling).
Достоверность меток: Каждый объект выборки должен иметь достоверную метку класса. Метки должны соответствовать реальности. Если мы обучаем модель распознавать кошек и собак, метка "кошка" должна соответствовать фотографии кошки, а не собаки.
Консистентность меток: Метки должны быть последовательными и непротиворечивыми. Один и тот же тип объекта должен всегда размечаться одним и тем же классом. Разметчик не должен "путать" классы или менять свое мнение в процессе разметки.
Проблемы с разметкой могут возникать из-за субъективности разметки (например, при оценке эмоций в тексте) или ошибок разметчиков (опечатки, невнимательность). Для решения этих проблем применяют несколько разметчиков и автоматическую проверку данных."
Слайд 7: Требования к обучающей выборке - Репрезентативность, Объем, Качество, Дополнительные требования (9-10 минуты):

(На экране: Заголовок "Требования к обучающей выборке", подзаголовки "Репрезентативность", "Объем", "Качество", "Дополнительные требования", основные тезисы для каждого подзаголовка)
"Помимо разметки, к обучающей выборке предъявляются и другие важные требования:
Репрезентативность выборки: Выборка должна отражать реальное распределение данных в целевой задаче. Если мы хотим, чтобы модель хорошо работала в реальных условиях, обучающие данные должны быть похожи на реальные данные. Важные аспекты репрезентативности:
Баланс классов: Избегать сильного дисбаланса классов. Если один класс встречается в выборке значительно чаще, модель может сместиться в сторону этого класса. Для решения проблемы дисбаланса применяют стратегии ребалансировки: oversampling, undersampling, взвешивание классов.
Разнообразие данных: Выборка должна включать все возможные варианты объектов, разные условия, разные подтипы внутри классов. Например, для распознавания лиц нужны изображения людей разного возраста, пола, расы и в разных условиях освещения.
Объем данных: Достаточный объем данных для обучения модели. Недостаток данных может привести к переобучению и плохой обобщающей способности. Необходимый объем данных зависит от сложности модели. Линейным моделям нужно меньше данных, нейронным сетям – значительно больше. При нехватке данных применяют аугментацию данных (искусственное увеличение выборки) и transfer learning (использование предобученных моделей).
Качество данных: Данные должны быть чистыми и без шума. Основные проблемы с качеством данных:
Шум в данных: Ошибки в признаках, некорректные метки. Решения: автоматическая очистка, визуальный аудит.
Пропущенные значения: Решения: удаление, импутация.
Избыточность признаков: Коррелирующие признаки. Решения: feature selection.
Дополнительные требования:

Временная согласованность: Для данных, меняющихся со временем (например, временные ряды), выборка должна включать актуальные примеры.
Юридические ограничения: Соблюдение GDPR и других юридических ограничений при работе с персональными и медицинскими данными.
Заключение (10 минута):
"В заключение, задача классификации – это важная задача машинного обучения, требующая внимательного подхода к подготовке данных. Для построения эффективной модели классификации необходимо обеспечить высокое качество разметки данных, репрезентативность выборки, достаточный объем данных и чистоту данных. Соблюдение этих требований – залог успеха в решении задач классификации.