Типы задач в машинном обучении и виды разметки данных

Классификация, регрессия, кластеризация
и современные подходы к разметке

Три основных парадигмы машинного обучения

Машинное обучение классифицируется по способу обучения модели на данных. Выбор типа обучения определяется наличием размеченных данных и характером решаемой задачи.

01

Обучение с учителем

Модель обучается на размеченных данных, где для каждого объекта известна правильная метка или значение целевой переменной. Применяется для задач классификации и регрессии.

02

Обучение без учителя

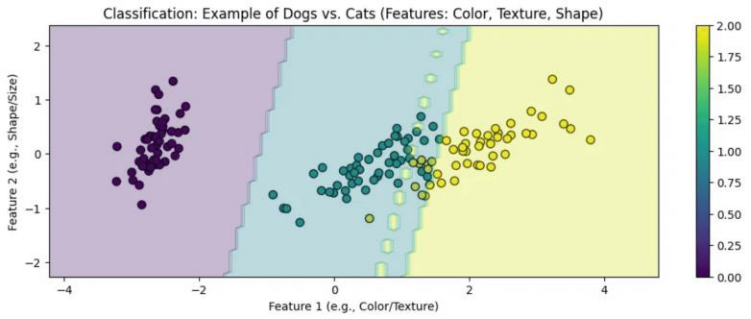
Модель самостоятельно находит закономерности в неразмеченных данных без явных целевых переменных. Используется для кластеризации, снижения размерности и выявления аномалий.

03

Обучение с подкреплением

Агент обучается принимать решения через взаимодействие со средой, получая награды или штрафы за свои действия. Применяется в робототехнике, играх и системах управления.

Классификация: предсказание категориальной переменной



Постановка задачи

По выборке $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, где $y_i \in \{1, 2, \ldots, K\}$, построить модель $f: X \to Y$ для предсказания класса.

Метрики классификации

1. Accuracy

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Доля правильных предсказаний среди всех объектов.

Матрица ошибок и метрика F1-Score

Матрица ошибок (Confusion Matrix)

Метрики на основе матрицы ошибок

Precision (Точность)

$$Precision = TP / (TP + FP)$$

Доля правильных положительных предсказаний среди всех положительных предсказаний. Отвечает на вопрос: "Какая доля объектов, которые мы предсказали как положительные, действительно положительные?"

Recall (Полнота)

$$Recall = TP / (TP + FN)$$

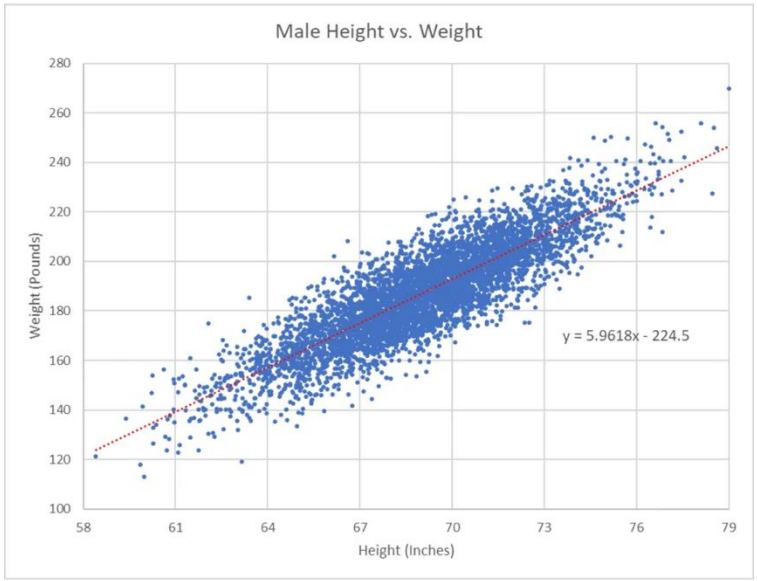
Доля найденных положительных объектов среди всех истинно положительных. Отвечает на вопрос: "Какую долю положительных объектов мы смогли найти?"

F1-Score (Гармоническое среднее)

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

Баланс между точностью и полнотой. Особенно полезна при несбалансированных классах, когда важны оба типа ошибок (FP и FN).

Регрессия: предсказание непрерывной переменной



Постановка задачи

По выборке $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, где $y_i \in \mathbb{R}$, построить модель $f: X \to \mathbb{R}$ для предсказания непрерывного значения.

Две популярные метрики

1. MSE (Mean Squared Error)

$$MSE = (1/n) \times \Sigma(y_i - \hat{y}_i)^2$$

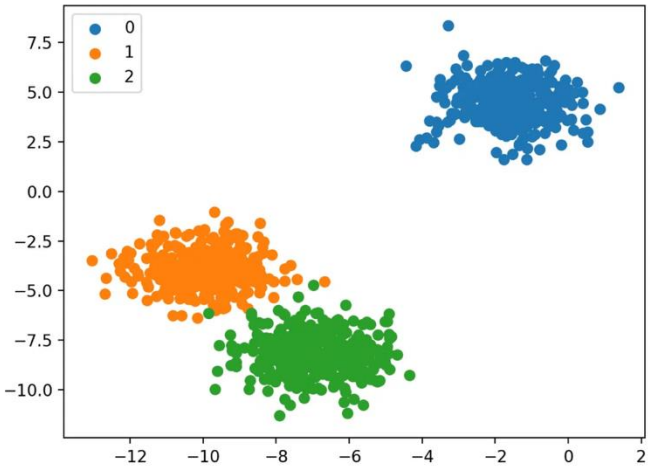
Среднее значение квадратов разностей. Сильно штрафует большие ошибки.

2. MAE (Mean Absolute Error)

$$MAE = (1/n) \times \Sigma|y_i - \hat{y}_i|$$

Среднее значение абсолютных разностей. Более устойчива к выбросам.

Кластеризация: поиск групп схожих объектов



Постановка задачи

По набору объектов $\{x_1, \ldots, x_n\}$ без меток разбить их на K кластеров так, чтобы объекты внутри кластера были схожи, а между кластерами — различны.

Две популярные метрики

1. Silhouette Score

$$s(i) = (b(i) - a(i)) / \max(a(i), b(i))$$

a(i) — среднее расстояние до своего кластера

b(i) — расстояние до ближайшего чужого кластера

Значения от -1 до 1. Высокие значения означают хорошее разделение.

2. Davies-Bouldin Index

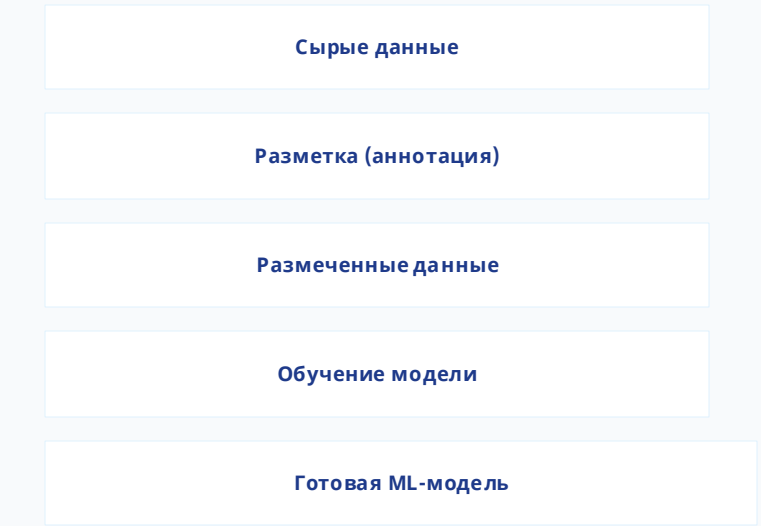
$$DB = (1/K) \times \Sigma \max_{i \neq j} [(\sigma_i + \sigma_j) / d(c_i, c_j)]$$

Разметка данных — фундамент обучения с учителем

Что такое разметка данных?

Процесс добавления меток (labels) или тэгов к сырым данным, которые показывают модели машинного обучения, что она должна предсказывать.

- **Занимает ~80% времени** ML-проекта на подготовку и обработку данных
- **Критична для supervised learning** — без разметки модель не может обучаться
- **Определяет качество модели** — garbage in = garbage out
- **Требует человеческой экспертизы** для точной и согласованной разметки



Пример

Разметка изображений кошек и собак для классификатора: каждому изображению присваивается метка "кошка" или "собака", чтобы модель научилась различать эти два класса.

"Garbage In, Garbage Out" — качество данных определяет качество модели

Принцип GIGO является фундаментальным в машинном обучении: качество выходных предсказаний модели напрямую зависит от качества входных обучающих данных.

Последствия некачественной разметки

Низкая точность предсказаний

Модель обучается на неправильных примерах и воспроизводит ошибки разметки, что приводит к систематическим ошибкам в предсказаниях на новых данных.

"Garbage In, Garbage Out" — качество данных определяет качество модели

Принцип GIGO является фундаментальным в машинном обучении: качество выходных предсказаний модели напрямую зависит от качества входных обучающих данных.

Последствия некачественной разметки

Низкая точность предсказаний

Модель обучается на неправильных примерах и воспроизводит ошибки разметки, что приводит к систематическим ошибкам в предсказаниях на новых данных.

Смещения и предвзятость (Bias)

Несогласованная или предвзятая разметка приводит к тому, что модель усваивает человеческие предрассудки и систематически ошибается на определенных группах объектов.

"Garbage In, Garbage Out" — качество данных определяет качество модели

Принцип GIGO является фундаментальным в машинном обучении: качество выходных предсказаний модели напрямую зависит от качества входных обучающих данных.

Последствия некачественной разметки

Низкая точность предсказаний

Модель обучается на неправильных примерах и воспроизводит ошибки разметки, что приводит к систематическим ошибкам в предсказаниях на новых данных.

Смещения и предвзятость (Bias)

Несогласованная или предвзятая разметка приводит к тому, что модель усваивает человеческие предрассудки и систематически ошибается на определенных группах объектов.

Неспособность обобщать

Модель переобучается на шумных или противоречивых метках и не может корректно работать на реальных данных, теряя способность к обобщению.

"Garbage In, Garbage Out" — качество данных определяет качество модели

Принцип GIGO является фундаментальным в машинном обучении: качество выходных предсказаний модели напрямую зависит от качества входных обучающих данных.

Последствия некачественной разметки

Низкая точность предсказаний

Модель обучается на неправильных примерах и воспроизводит ошибки разметки, что приводит к систематическим ошибкам в предсказаниях на новых данных.

Смещения и предвзятость (Bias)

Несогласованная или предвзятая разметка приводит к тому, что модель усваивает человеческие предрассудки и систематически ошибается на определенных группах объектов.

Неспособность обобщать

Модель переобучается на шумных или противоречивых метках и не может корректно работать на реальных данных, теряя способность к обобщению.

Потеря ресурсов

Некачественная разметка приводит к необходимости повторного сбора и разметки данных, что означает потерю времени, денег и вычислительных ресурсов.

Типы разметки по модальности данных

Изображения

Типы разметки:

- Классификация изображений
- Object Detection (bounding boxes)
- Semantic Segmentation
- Instance Segmentation
- Keypoint Detection

Примеры: распознавание объектов, медицинская диагностика, автономное вождение

Текст

Типы разметки:

- Классификация текста
- Named Entity Recognition (NER)
- Part-of-Speech Tagging
- Relation Extraction
- Question Answering

Примеры: анализ тональности, извлечение сущностей, чат-боты

Аудио

Типы разметки:

- Классификация звуков
- Speech-to-Text транскрипция
- Speaker Identification
- Emotion Recognition
- Sound Event Detection

Примеры: голосовые ассистенты, распознавание эмоций, транскрибация

Видео

Типы разметки:

- Action Recognition
- Object Tracking
- Video Segmentation
- Event Detection

Примеры: видеонаблюдение, спортивная аналитика, контент-модерация

Специализированные

Типы данных:

- LIDAR (облака точек)
- DICOM/NIfTI (медицинские изображения)
- Временные ряды
- 3D модели

Примеры: автономные автомобили, медицинская визуализация, IoT

Методы разметки данных

Ручная разметка

Человек вручную размечает каждый элемент данных

- Внутренняя (in-house)
- Краудсорсинг
- Экспертная разметка

Полуавтоматическая

Комбинация автоматической предразметки и ручной проверки

- Pre-labeling + проверка
- Active Learning
- Transfer Learning

Автоматическая

Модель автоматически генерирует метки без участия человека

- Weak Supervision
- Self-Supervised Learning
- Synthetic Data Generation

Метод	Стоимость	Скорость	Качество	Масштабируемость
Ручная разметка	Высокая	Медленно	Отличное	Низкая
Полуавтоматическая	Средняя	Средне	Хорошее	Средняя
Автоматическая	Низкая	Быстро	Среднее	Высокая

Сложности разметки данных

Высокие затраты времени и труда

Проблема:

80% времени ML-проекта на подготовку данных. Масштаб: 150,000+ изображений с 10 объектами каждое.

Решение:

Автоматизация, краудсорсинг, активное обучение

Необходимость экспертных знаний

Проблема:

Медицинская диагностика требует врачей, юридические документы — юристов. Высокая стоимость экспертов.

Решение:

Обучение аннотаторов, гибридный подход, transfer learning

Риск несогласованности

Проблема:

Разные аннотаторы применяют разные критерии. Субъективность оценок (например: положительный vs саркастический отзыв).

Решение:

Четкие guidelines, перекрестная разметка, метрики IAA

Риск ошибок

Проблема:

Человеческий фактор при работе с большими объемами. Усталость и невнимательность при монотонной работе.

Решение:

Контроль качества, золотой стандарт, регулярные перерывы

Большие языковые модели для разметки данных

Большие языковые модели (LLM) — это модели глубокого обучения с миллиардами параметров, способные решать широкий класс задач обработки естественного языка без специального обучения.

Преимущества использования LLM

Скорость и масштабируемость

LLM могут автоматически размечать тысячи примеров за минуты, что в десятки раз быстрее ручной разметки. Легко масштабируются на большие датасеты.

Согласованность и экономия

Обеспечивают единообразный подход к разметке, исключая человеческие ошибки. Значительно снижают затраты на разметку для простых и средних задач.

Применение LLM

- Классификация текстов и анализ тональности
- Выделение именованных сущностей (NER)
- Генерация синтетических данных
- Автоматическая проверка и валидация

Ограничения

Требуют валидации для специализированных задач (медицина, юриспруденция). Могут ошибаться в узкоспециализированных областях. Использование коммерческих API может быть дорогим.

Рекомендуемый подход

Гибридная стратегия: использование LLM для первичной автоматической разметки с последующей человеческой валидацией критически важных меток.

Качественная разметка — основа успешного машинного обучения

Выбор типа задачи определяется характером целевой переменной: классификация для категориальных меток, регрессия для непрерывных значений, кластеризация для поиска структуры в данных.

Метрики качества должны соответствовать специфике задачи: Accuracy и F1-Score для классификации, MSE и MAE для регрессии, Silhouette Score для кластеризации.

Принцип "Garbage In, Garbage Out" подчеркивает критическую важность качественной разметки данных. Некачественные данные приводят к некачественным моделям независимо от сложности алгоритма.

Современные подходы к разметке эволюционируют от полностью ручной к гибридным методам с использованием LLM. Большие языковые модели значительно ускоряют и удешевляют процесс разметки.

Оптимальная стратегия — комбинация автоматической разметки с помощью LLM и человеческой экспертизы для обеспечения высокого качества данных при разумных затратах ресурсов.