ЗАНЯТИЕ 1. ОСНОВНЫЕ ПОНЯТИЯ МАШИННОГО ОБУЧЕНИЯ

ОБЪЕКТ - абстрактная сущность, с которой мы работаем (абстрактная, т.е. не число, а набор признаков)

ЦЕЛЕВАЯ ПЕРЕМЕННАЯ - величина, которую мы хотим прогнозировать

ПРИЗНАК - характеристика объекта

МОДЕЛЬ МАШИННОГО ОБУЧЕНИЯ - математическая функция, которая по признакам объекта каким-то образом предсказывает ответ (функция, которая сопоставляет объекту значение целевой переменной)

Признаки объекта x можно записать в виде вектора $(f_1(x), \dots, f_n(x))$

Матрица "объекты - признаки":

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_n) & \dots & f_n(x_n) \end{pmatrix}$$

В строчках - объекты

В столбцах - признаки

Схема получения предсказания

В задачах обучения с известными классами (обучение по прецедент) всегда есть **2 этапа:**

1. **Этап обучения** (training): по выборке $X = \{(x_i, y_i)\}$ строим алгоритм a
2. **Этап применения** (testing): алгоритм a для новых объектов X выдает ответы $a(x)$

ОБУЧАЮЩАЯ ВЫБОРКА – конечный набор объектов, для которых известны значения целевой переменной.

Виды признаков:
- числовые
- бинарные (0/1)
- категориальные (имеющие бесконечное множество вариантов)
- признаки со сложной внутренней структурой

Признаки, которые выглядят как числа, но не ведут себя как числа, называются категориальными.

Виды данных:
- таблицы (-xls, -csv и т.д.)
- текстовые данные
- изображения
- звуки
- логи

Типы задач в зависимости от целевой переменной

КЛАССИФИКАЦИЯ:

Целевая переменная – некоторый класс из конечного числа

Пример: вернёт ли человек кредит;
болен ли человек по анализам

- $Y = \{0, 1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$ – классификация на M непересекающихся классов
- $Y = \{0, 1\}^M$ – классификация на M классов, которые могут пересекаться

Мультиклассовая классификация:

- определение типа объекта на изображении
- определение наиболее подходящей профессии для данного кандидата

РЕГРЕССИЯ:

Целевая переменная может принимать бесконечно много значений, т.е. может быть нецелым или отрицательным числом.

Пример: предсказание прибыли ресторана; стоимость квартиры

КЛАСТЕРИЗАЦИЯ:

Задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.

Также есть задачи:

- генерации (сгенерировать изображение/текст)
- поиск аномалий (выбросов)
- ранжирования (сортировка в порядке релевантности)
- понижения размерности (сделать у объектов меньше признаков)

Все задачи можно разбить на 2 класса:

1) Используем целевую переменную - **ОБУЧЕНИЕ С УЧИТЕЛЕМ** (классификация, регрессия, ранжирование)

2) Не используем целевую переменную - **ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ** (кластеризация, понижение размерности и т.д.)

Основные типы в экзамене: классификация, регрессия и немного кластеризация

Обучение алгоритма, оценка предсказательной способности.

Допустим, мы хотим предсказать стоимость дома у по его площади (x_1) и кол-ву комнат (x_2).

Используем линейную модель для предсказания: $a(x) = w_0 + w_1x_1 + w_2x_2$, где w_0, w_1, w_2 - параметры модели (веса), но мы их не знаем. Наша задача при обучении модели найти эти веса

Чтобы найти параметры мы можем перебирать все возможные прямые ($a(x)$ - линейная функция) таким образом, чтобы минимизировать ошибку, т.е. нам нужно среди всех возможных прямых, среди всех возможных весов, найти ту прямую, у которой будет наименьшая ошибка.

ОБУЧЕНИЕ МОДЕЛИ - процесс подбора весов модели; процесс подбора оптимальной прямой такой, что на ней минимизируется ошибка

ФУНКЦИОНАЛ ОШИБКИ - функционал, измеряющий качество работы алгоритма

Пример: среднеквадратичная ошибка, MSE:

$$Q(a, X) = \frac{1}{L} \sum_{i=1}^{L} (a(x_i) - y_i)^2$$

X - объекты
L - кол-во объектов
a - алгоритм
$a(x_i)$ - ответ алгоритма на объекте x_i
y_i - истинные ответы

Метрики качества

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используют **метрики качества**.

Примеры:
- среднеквадратичная ошибка - для регрессии
- доля правильных ответов - для классификации

Доля правильных ответов:

$$accuracy(a, x) = \frac{1}{L} \sum_{i=1}^{L} [a(x_i) = y_i]$$

Алгоритм решения задачи анализа данных

1) Постановка задачи
2) Выделение признаков
3) Формирование выборки
4) Выбор функции потерь и метрики качества
5) Предобработка данных