Министерство науки и высшего образования Российской Федерации
Калужский филиал федерального государственного автономного образовательного учреждения высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(КФ МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУК Информатика и управление

КАФЕДРА ИУК4 Программное обеспечение ЭВМ, информационные технологии

КОНТРОЛЬНАЯ РАБОТА №1

по дисциплине: «Методы машинного обучения»

Выполнил: студент группы ИУК4-72Б _____ Губин Е.В.
(Подпись) (И.О. Фамилия)

Проверил: _____ Семененко М.Г.
(Подпись) (И.О. Фамилия)

Дата сдачи (защиты):

Результаты сдачи (защиты):

- Балльная оценка:
- Оценка:

Калуга, 2025

Вариант №4

Вопрос №1. Дайте определение сингулярному разложению.

Сингулярное разложение (SVD) — это метод линейной алгебры, который позволяет представить любую прямоугольную матрицу как произведение трёх специальных матриц: ортогональных и диагональной.

Формально:

$$A = UEV^T$$

где:

- A — исходная матрица размера m x n;
- U — ортогональная матрица размера m x m;
- E — прямоугольная диагональная матрица размера m x n;
- V^T — транспонированная ортогональная матрица V размера n x n.

Матрица U:

Она состоит из ортонормированных собственных векторов матрицы AA^T. Эти векторы называются левыми сингулярными векторами. Свойство:

$$U^T U = I_m$$

Объяснение свойства: строки и столбцы взаимно ортогональны и имеют длину 1.

Матрица V:

Она состоит из ортонормированных собственных векторов матрицы A^TA. Эти векторы называются правыми сингулярными векторами. Свойство:

$$V^T V = I_n$$

Матрица E (сигма):

Диагональная (точнее, прямоугольная диагональная) матрица, в которой на диагонали стоят сингулярные значения — числа $\sigma_1, \sigma_2, \ldots, \sigma_r$. Эти значения всегда неотрицательные и упорядочены по убыванию. Они являются квадратными корнями собственных значений матриц A^TA и AA^T:

$$\sigma_i = \sqrt{lambda_i}$$

где lambda_i – собственные значения.

Геометрический смысл:

Сингулярное разложение показывает, как линейное преобразование, описываемое матрицей AAA, воздействует на пространство:

1. V^T — поворачивает/отражает координатную систему;
2. E — растягивает или сжимает оси на величиныσ_i;
3. U — снова поворачивает систему координат.

То есть A можно рассматривать как комбинацию поворота → масштабирования → поворота.

Значение в машинном обучении:

Сингулярное разложение — это основа многих методов машинного обучения и анализа данных, например:

Область	**Применение**
Снижение размерности	Алгоритм PCA (Principal Component Analysis) — реализуется через SVD.
Решение переопределённых систем	Используется в методе наименьших квадратов, особенно при наличии линейно зависимых признаков.
Рекомендательные системы	Алгоритмы, как в Netflix или YouTube, используют SVD для факторизации матрицы пользователь–объект.
Сжатие данных	Например, сжатие изображений с помощью сохранения только нескольких наибольших сингулярных значений.

Свойства SVD:

- Всегда существует для любой матрицы (в отличие от спектрального разложения, которое требует квадратности и симметрии).
- Позволяет определить ранг матрицы — это количество ненулевых сингулярных значений.
- Число обусловленности матрицы:

$$k(A) = \frac{\sigma_1}{\sigma_r}$$

  показывает, насколько "устойчива" матрица к численным ошибкам.
- Через SVD можно вычислять псевдообратную матрицу:

$$A^+ = VE^+U^T$$

где E^+ — диагональная матрица, в которой каждое ненулевое σ_i заменено на $\frac{1}{\sigma_i}$.

Вопрос №2. Охарактеризуйте решающие деревья.

Решающее дерево (Decision Tree) — это один из самых интуитивно понятных и популярных методов машинного обучения, который используется для классификации и регрессии.

Модель представляет собой дерево, где:

- Внутренние узлы (nodes) — это условия (вопросы) о признаках данных,
- Ветви (branches) — это результаты проверки условия,
- Листья (leaves) — это итоговые решения: класс (для классификации) или численное значение (для регрессии).

Структура решающего дерева:

- Корень - начальная точка, где выбирается первый признак для разделения.
- Внутренний узел – условие (например, возраст > 30)
- Ветвь (ребро) - путь, соответствующий результату проверки (да/нет).
- Лист (leaf) - конечный результат (предсказание класса или значения).

Как строиться решающее дерево:

Построение дерева — это рекурсивный процесс разбиения данных на подмножества так, чтобы каждое разбиение делало выборку максимально “чистой” (т.е. чтобы объекты в одной ветви были как можно более похожи по целевому признаку).
Основная идея:

1. Выбрать признак и порог, по которому данные делятся на 2 (или больше) группы.
2. Оценить качество этого разбиения с помощью метрики (например, энтропии или индекса Джини).
3. Повторять процесс для каждой ветви, пока не достигнем:
    - чистых классов,
    - или заданной глубины,
    - или минимального числа объектов в узле.

Метрики чистоты (критерии выбора признака):ш

1. Энтропия (Entropy)

   Показывает степень неопределённости:

$$H(S) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

   где p_i — доля объектов класса i в множестве S.

   - Если все объекты одного класса → энтропия = 0 (полная определённость).
   - Если классы смешаны поровну → энтропия максимальна.

2. Индекс Джини (Gini index)

   Часто используется в алгоритме CART:

$$Gini(S) = 1 - \sum_{i=1}^{k} p_i^2$$

   Чем меньше Gini, тем “чище” узел.

Алгоритмы построения деревьев:

- ID3 - использует энтропию и информационный прирост. Работает только с категориальными признаками.
- C4.5 - улучшение ID3: поддерживает числовые признаки, нормализует прирост.
- CART - основан на индексе Джини; может использоваться для классификации и регрессии.

Преимущества:

- Понятность и интерпретируемость. Можно визуализировать и объяснить решение (“если..., то...” — как правила).
- Не требует масштабирования данных. Неважно, в каких единицах измеряются признаки.
- Работает с категориальными и числовыми данными.
- Автоматическое выделение важнейших признаков.

- Быстрое обучение и предсказание.

Недостатки:

- Переобучение (overfitting) — если дерево слишком глубокое, оно “запоминает” данные.
- Нестабильность — небольшие изменения в данных могут привести к совершенно другому дереву.
- Жадный алгоритм — выбор лучшего признака на каждом шаге не гарантирует глобально оптимального дерева.
- Не подходит для линейно разделимых задач (иногда слишком сложная структура для простой зависимости).

Значение в машинном обучении:

Область	**Значение**
Классификация	Определение типа клиента, болезни, спама и т.д.
Регрессия	Прогнозирование цен, спроса, температуры и т.д.
Обработка признаков	Определение важности признаков, построение интерпретируемых правил.