Задание

«Сравнение рекомендательных систем на основе SVD, матричной факторизации и нейронных сетей»

Цель работы

Исследовать и сравнить эффективность трёх различных подходов к построению рекомендательных систем: сингулярного разложения (SVD), матричной факторизации (MF) с градиентным спуском и полносвязной нейронной сети (FNN).

Задачи

1. Изучить теоретические основы методов SVD, матричной факторизации и нейронных сетей для рекомендательных систем.
2. Освоить подготовку и предобработку данных для задачи предсказания рейтингов.
3. Реализовать или адаптировать модели SVD, матричной факторизации и полносвязной нейронной сети.
4. Изучить метрики оценки качества рекомендательных систем (Precision, Recall).
5. Провести сравнительный анализ моделей по точности и времени выполнения.

Исходные данные (выбрать один из датасетов)

- **Вариант 1: MovieLens 100K Dataset**
  https://www.kaggle.com/datasets/prajitdatta/movielens-100k-dataset
  100,000 оценок от 943 пользователей для 1,682 фильмов. Классический учебный набор.
- **Вариант 2: Jester Joke Ratings (первые 100K записей)**
  https://www.kaggle.com/datasets/vikashrajluhaniwal/jester-17m-jokes-ratings-dataset
  Оценки шуток. Плотная матрица взаимодействий, проще для алгоритмов факторизации.
- **Вариант 3: Anime Recommendations (подмножество 100K)**
  https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database
  Оценки аниме от -1 до 10. Интересная тематика с отрицательными оценками.

Примечание по подготовке данных:

- Для Jester и Anime датасетов необходимо самостоятельно создать подмножество из первых 100,000 записей.
- Все датасеты содержат как минимум три колонки: UserID, ItemID, Rating.

Описание выбранного датасета

- Название датасета: [указать выбранный вариант]
- Описание: [краткое описание структуры данных: пользователи, предметы, оценки]
- Разреженность матрицы взаимодействий: [рассчитать как (1 - (кол-во оценок) / (пользователи * предметы)) * 100%]
- Диапазон оценок: [например, от 1 до 5]
- Распределение оценок: [описать или визуализировать]

Методика выполнения

Часть 1: Подготовка данных и предварительный анализ

1. Загрузка данных и формирование матрицы взаимодействий $\mathbf{R}$ (пользователи $\times$ предметы).
2. Разделение данных на обучающую (80%), валидационную (10%) и тестовую (10%) выборки с соблюдением временного порядка или случайным разбиением.
3. Статистический анализ: распределение оценок, активность пользователей, популярность предметов.
4. Исследование уровня разреженности данных.
5. Нормализация оценок (при необходимости, например, центрирование).

Часть 2: Реализация и обучение моделей

Задача: Предсказание отсутствующей оценки r_{ui}.

1. **Модель 1: SVD (сингулярное разложение).**
   - Реализовать или использовать функцию SVD для матрицы $\mathbf{R}$.
   - Восстановить матрицу оценок: $\hat{R} = U_k \Sigma_k V_k^T$.
   - Подобрать оптимальную размерность латентного пространства k на валидационной выборке.
2. **Модель 2: Матричная факторизация (MF) с регуляризацией.**
   - Реализовать оптимизацию функции потерь: $\min_{P,Q} \sum_{(u,i)\in \text{train}} (r_{ui} - p_u^T q_i)^2 + \lambda(\|p_u\|^2 + \|q_i\|^2)$.
   - Использовать стохастический градиентный спуск (SGD) или метод Alternating Least Squares (ALS).
   - Подобрать гиперпараметры: количество факторов k, скорость обучения η, коэффициент регуляризации λ, количество эпох.
3. **Модель 3: Полносвязная нейронная сеть (FNN).**

- Спроектировать архитектуру сети. Вход: векторное представление пользователя и предмета. Выход: предсказанная оценка.
- Подобрать гиперпараметры: количество и размер скрытых слоев, функции активации.

Часть 3: Оценка качества рекомендаций

1. Рассчитать **ранжирующие метрики** (для этого преобразовать задачу в бинарную: оценка $\geq$ порога = релевантный предмет):
   - Precision@K (точность среди топ-K рекомендаций).
   - Recall@K (полнота среди топ-K рекомендаций).
2. Оценить время обучения и время предсказания для каждой модели.

Часть 4: Анализ влияния разреженности данных

1. Создать подвыборки обучающих данных с разной степенью разреженности (например, взять случайные 30%, 50%, 70% данных).
2. Повторить обучение и оценку моделей на каждой подвыборке.
3. Проанализировать, как качество моделей зависит от количества доступных данных.

Метрики оценки качества

Для сравнения моделей использовать:

- **Precision@10** – точность среди 10 наиболее рекомендованных пользователю предметов.
- **Recall@10** – полнота среди 10 наиболее рекомендованных пользователю предметов.
- **Время обучения** – общее время подбора гиперпараметров и финального обучения.
- **Время предсказания** – время предсказания оценок для всех пар пользователь-предмет в тестовой выборке.

Требования к отчету

Содержание отчета

1. Титульный лист.
2. Цель, задачи и описание датасета.
3. Теоретическое описание моделей (SVD, MF, FNN) с формулами.
4. Методика исследования (предобработка, валидация, метрики).
5. Результаты:
   - Подробное описание подобранных гиперпараметров для каждой модели.

- Таблицы и графики с результатами экспериментов.
- Визуализация процесса обучения (графики функции потерь).

6. Сравнительный анализ всех моделей по всем метрикам.
7. Выводы и практические рекомендации по выбору метода в зависимости от условий.