Домашнее задание 1

Общие требования

- Обеспечьте воспроизводимость и должное качество оформления вашей работы, по которым ее можно однозначно оценить:
  - Все теоретические задачи должны быть размещены в ноутбуке в виде фотографии или оформлены на языке разметки markdown.
  - Все случайные сиды (‘seeds‘) зафиксированы, и повторный запуск ноутбука даёт те же результаты оценки.
  - Ноутбук должен запускаться сверху вниз без ошибок.
  - Время работы всех ячеек, связанных с построением модели, инференсом и оценкой, должно быть замерено с использованием `%%time`.
  - Все необходимые метрики должны быть отображены как результат работы ячейки ноутбука.
- Четко и аккуратно опишите ваш подход, проделенные эксперименты, полученные результаты и выводы. Подчеркните нужные выводы визуализацией.
- Пожалуйста, не списывайте и не делитесь кодом с однокурсниками.
- Срок сдачи: 2 октября 23:59

1 Задача 1 (15 баллов + 5 бонусных баллов)

Теоретическое задание

Рассмотрим задачу персонализированных top-n рекомендаций, сформулированную как выбор top-n наиболее релевантных айтемов для пользователя u:

$$\text{toprec}(u, n) = \arg\max_{i}^{n} r_{ui},$$

где r_{ui} — релевантность (не обязательно рейтинг), назначаемая моделью айтему i для пользователя u.

Покажите, что следующие модели не отвечают задаче **персонализированных** top-n рекомендаций, а именно - выбор релевантных айтемов не зависит от параметров / признаков пользователя:

1. Baseline predictors:

   $$r_{ui} \approx g_u + f_i + \mu,$$

   где векторы g_u и f_i — соответствующие предикторы пользователя и айтема.

2. Регрессионная модель, обученная в форме:

   $$r_{ui} \approx \theta x_{ui} + \epsilon,$$

   где вектор x_{ui} кодирует некоторые признаки и пользователя u, и айтема i (например, в виде конкатенации признаков пользователя и айтема), а θ — обучаемые веса регрессионной модели.

Бонус (5 баллов): Предложите другие алгоритмы или методы предобработки признаков, которые могут обеспечить более высокий уровень персонализации, чем регрессионные модели, описанные выше. Опишите алгоритм, на вход которому подается признаковое описание x_{ui} (пользователя u и айтема i), а выходом является вещественное значение – релевантность айтема для пользователя.

2 Задача 2 (20 баллов)

Датасет

Вы будете использовать данные MovieLens-1M (ML-1m) для экспериментов.

Разбиение данных и метрики

Пожалуйста, используйте метрики, приведенные в конце лекции по content-based фильтрации, а именно: близость по релевантности. Например, выберите две пары айтемов с высоким и низким рейтингами, выставленными пользователем и исключите их из тренировочной выборки. После обучения, предскажите релевантность оставшихся фильмов и оцените как часто предсказанная релевантность ближе к айтему с высоким рейтингом, чем к айтему с низким рейтингом. Усредните значение по всем пользователям.

Анализ baseline predictors

Используя baseline predictors из семинара, визуализируйте распределения смещений айтемов $\mathbf{f}$ для фильмов ужасов и драм. Опишите различия в этих распределениях. Сделайте хотя бы одну гипотезу, почему наблюдаются эти различия.

Baseline predictors с регуляризацией

Ваша задача — реализовать модель baseline predictors с регуляризацией. Необходимо решить следующую задачу оптимизации, введённую на лекции:

$$\arg\min_{\mathbf{f},\mathbf{g}} \left[\sum_{(i,j)\in\mathcal{O}} (r_{ij} - g_i - f_j - \mu)^2 + \sum_i \gamma_i g_i^2 + \sum_j \lambda_j f_j^2 \right],$$

где

- $\mathcal{O}$ — множество всех известных записей (наблюдаемые данные),
- r_{ij} — рейтинг, поставленный пользователем i айтему j,
- g_i и f_j — параметры смещения для пользователя i и айтема j, соответственно,
- μ — глобальное среднее значение рейтинга во всех взаимодействиях,
- γ_i, λ_j — гиперпараметры регуляризации.

Сравните модели в различных сценариях, когда

1. $\gamma_i = \gamma, \lambda_j = \lambda$, не отличаются для пользователей и айтемов,
2. $\lambda_j = \frac{\lambda}{p_j}, \gamma_i = \frac{\gamma}{q_i}$, где p_j — популярность айтема j, а q_i — активность пользователя i (число взаимодействий).

Попробуйте как минимум 3 различных значения для каждого типа регуляризации. Опишите наблюдаемые различия.

Baseline predictors с негативным семплированием

Трактуя отсутсвующие рейтинги как рейтинги со значением 0, реализуйте модель baseline predictors с использованием негативного семплирования (пример реализации семплирования можно найти в семинаре по content-base filtering). Задача оптимизации выглядит следующим образом:

$$\sum_{(i,j)\in\mathcal{O}} (r_{ij} - t_i - f_j)^2 + \alpha \sum_{(i,j)\notin\mathcal{O}} (0 - t_i - f_j)^2 \to \min,$$

где α – гиперпараметр.

На каждый позитивный пример $(i, j) \in \mathcal{O}$, выбирайте **10** негативных примеров. Исследуйте эффект от выбора параметра α.

Baseline predictors с использованием всей матрицы взаимодействий

Последняя часть задачи — реализовать модель baseline predictors с использованием всей матрицы взаимодействий, дополняя нулями ячейки с неизвестными взаимодействиями.

$$\arg\min_{\mathbf{f},\mathbf{g}} \|\mathbf{R} - \mathbf{g}\mathbf{e}_N^\top - \mathbf{e}_M\mathbf{f}^\top\|_F^2,$$

где

- $\mathbf{R}$ — матрица рейтингов, с дополненными нулями неизвестных взаимодействий,
- $\mathbf{e}_K$ — вектор единиц размера K,
- M и N — количество пользователей и айтемов соответственно.

Следует вывести аналитическое решение, использующее формулу Шермана-Вудбери-Моррисона. Обратите внимание, что работа с плотными матрицами запрещена.

Сравните реализованные модели между собой. Что можно сказать о качестве их рекомендаций?

Задача 3 (15 баллов)

Датасет

Для экспериментов используйте датасет RentTheRunway об аренде одежды людьми в одноименном сервисе.

Описание датасета

- **item_id**: уникальный идентификатор айтема
- **weight**: вес клиента
- **rented for**: цель аренды одежды
- **body type**: тип телосложения клиента
- **review_text**: текст отзыва клиента
- **review_summary**: краткое содержание отзыва
- **size**: размер вещи
- **rating**: рейтинг вещи
- **age**: возраст клиента
- **category**: категория одежды
- **bust size**: обхват груди клиента
- **height**: рост клиента
- **fit**: подошло или нет по размеру
- **user_id**: уникальный идентификатор клиента
- **review_date**: дата написания отзыва

Разбиение данных и метрики

Для задачи выберите топ-500 пользователей по количеству взаимодействий с айтемами. Пожалуйста, используйте метрики, приведенные в конце лекции по content-based фильтрации. Например, выберите одну пару айтемов с высоким и низким рейтингами от пользователя, постройте профиль пользователя (возьмите топ-5 айтемов с лучшим предсказанным рейтингом и усредните значения признаков), затем сравните расстояние до «положительного» айтема и до «отрицательного» (например с помощью косинусной схожести). Возьмите за

метрику количество случаев, когда построенный профиль оказывался ближе к «положительному» товару, чем к «отрицательному».

Контентные модели с персонализацией

В этой задаче вы обучите простую контентную модель *для каждого пользователя индивидуально*, чтобы достичь некоторого уровня персонализации в задаче предсказания рейтинга пользователя арендованной вещи. Следуйте общим требованиям из начала файла. Старайтесь избегать циклов в вашем коде. Каждый из ваших рекомендательных пайплайнов (предобработка признаков + обучение + предсказание) **должна выполняться менее 4 часов** для всех тестовых пользователей (засеките время!).

Ансамбль контентных моделей

Используя данные о рейтингах пользователей и информацию об айтемах, постройте контентную модель. Вы можете использовать любые методы предобработки признаков (включая те, что были использованы в семинаре). Опишите вашу модель и методы предобработки, объясните ваш выбор.

Оцените качество полученного ансамбля пользовательских моделей.

Улучшения

Улучшите модель, изменяя:

- способ обработки контентной информации (например, выбор различных признаков и методов предобработки; настройка регуляризации и других важных параметров *регрессионной* модели). Вы также можете использовать исходные тексты для обучения модели.

- способ учета истории пользователя — можно использовать случайную подвыборку айтемов фиксированного размера.

Гибридная модель (5 бонусных баллов)

Постройте рекомендательную систему на основе гибридного baseline predictor. Используйте ту же схему для оценки, что и в предыдущей задаче.

Реализация гибридного базового предиктора

$$r_{ij} = \mu + f_j + t_i + (Va_i)^\top W c_j$$

где:

- a_i — атрибуты пользователя i,
- c_j — характеристики айтема j,
- V, W — обучаемые веса.

Сравните полученный результат с регрессионным подходом и сделайте выводы.