Частина I

Лабораторна робота №3, Рекомендаційні системи

1 Загальні відомості

Даний документ містить інформацію щодо лабораторної роботи №3 з алгоритмів машинного навчання. Всього передбачено 12 варіантів. Ваш варіант відповідає Вашому номеру у списку з журналу (якщо Ваш номер більше 12, то віднімiть від Вашого номера 12).

В кожному варіанті буде вказано посилання по якому можна скачати датасет для аналізу а також додаткову інформацію.

Завдання полягає у тому, щоб побудувати рекомендаційну систему.

1. Зчитайте дані, та перетворіть, за потреби у форму матриці (можливо з пропущеними даними).
2. Імплементуйте алгоритм ймовірнісної матричної факторизації в якому d, λ, σ^2-будуть параметрами.
3. Підберіть найкращі d, λ, σ^2. Найголовніше тут d. λ та σ^2 можете покласти рівними одиниці і підбирати якщо залишається час.

Якщо у Вашому датасеті більше 100 000 значень, або матриця виходить суттєво більшою ніж 1.5 млн елементів (включаючи пропущені значення), то можете урізати свій датасет. Для валідації викиньте кілька спостережень і покажіть як алгоритм їх спрогнозує.

2 Вимоги до виконання, оформлення та здачі

Роботу можна виконувати в R або Python. Вибірку слід розбити на тестову та тренувальну у пропорціх 20% та 80%.

Робота має бути оформлена у вигляді .pdf (можливо .docx, як показала практика .doc-файли не завжди відкриваються) файлу який містить всю необхідну інформацію. Роботи потрібно здавати на парах або надсилати на email. У файлі має бути вказано: призвіще ім'я, номер варіанту, назва групи. Якщо pdf/docx файл не містить коду, то файл з кодом слід прикріпити окремо. Якщо робота надсилається на email, то в темі листа має бути вказано: "[ML-LAB-3], <ПРИЗВІЩЕ ІМ'Я>, <Група>" (символи [] потрібні, символи <> - ні). Строки виконання роботи - до кінця дня 23 травня (врахуйте, що це день Києва :)).

Робота повинна містити коментарі, кожен фрагмент коду повинен мати пояснення - навіщо це робиться і які результати. Опишіть швидкість роботи алгоритму, кількість ітерацій, досягнуті значення цільової функції та

підібрані гіперпараметри. Алгорим повинен працювати притомний час, не більше години для одного набору гіперпараметрів та 10-20 ітерацій (взагалі кажучи, це повинно займати кілька хвилин).

Наприкінці роботи має бути висновок який повинен містити короткий огляд методів та отриманих результатів.

Кожна робота буде розглядатися на відповідність критеріям описаним вище, та на обгрунтованість прийнятих рішень. Кожен студент, повинен виконати свою роботу самостійно. Ідентичні, або майже ідентичні роботи прийматися до уваги не будуть. Прошу не брати чужі роботи і робити в них косметичні зміни, типу заміни на свій датасет. Також не треба присилати мені подібні задачі з інтернету, по-перше я їх вже бачив, по-друге там часто не зовсім ця задача. В разі виявлення плагіату робота буде анульована. Якщо буде виявлено, що робота містить матеріал з роботи іншого студента, анульовано буде обидві роботи.

3 Варіанти

Варіант 1
Опис даних: `http://www2.informatik.uni-freiburg.de/~cziegler/BX/`
Самі дані: `http://www2.informatik.uni-freiburg.de/~cziegler/BX/BX-CSV-Dump.zip`

Варіант 2
Опис даних: `https://www.kaggle.com/tamber/steam-video-games/`
Самі дані: `https://www.kaggle.com/tamber/steam-video-games/download`

Варіант 3
Опис даних: http://files.grouplens.org/datasets/hetrec2011/hetrec2011-movielens-readme.txt
Самі дані: http://files.grouplens.org/datasets/hetrec2011/hetrec2011-movielens-2k-v2.zip

Варіант 4
Опис даних: http://files.grouplens.org/datasets/hetrec2011/hetrec2011-lastfm-readme.txt
Самі дані: http://files.grouplens.org/datasets/hetrec2011/hetrec2011-lastfm-2k.zip
Додаткова інформація: Використовуйте кількість прослуховувань в якості метрики (рейтингу)

Варіант 5
Опис даних: `https://guoguibing.github.io/librec/datasets.html`, CiaoDVD
Самі дані: `https://guoguibing.github.io/librec/datasets/CiaoDVD.zip`

Варіант 6
Опис даних: https://www.kaggle.com/tamber/steam-video-games/
Самі дані: https://www.kaggle.com/tamber/steam-video-games/download

Варіант 7
Опис даних: http://www2.informatik.uni-freiburg.de/~cziegler/BX/
Самі дані: http://www2.informatik.uni-freiburg.de/~cziegler/BX/BX-CSV-Dump.zip

Варіант 8
Опис даних: https://www.kaggle.com/tamber/steam-video-games/
Самі дані: https://www.kaggle.com/tamber/steam-video-games/download

Варіант 9
Опис даних: http://files.grouplens.org/datasets/hetrec2011/hetrec2011-movielens-readme.txt
Самі дані: http://files.grouplens.org/datasets/hetrec2011/hetrec2011-movielens-2k-v2.zip

Варіант 10
Опис даних: http://files.grouplens.org/datasets/hetrec2011/hetrec2011-lastfm-readme.txt
Самі дані: http://files.grouplens.org/datasets/hetrec2011/hetrec2011-lastfm-2k.zip
Додаткова інформація: Використовуйте кількість прослуховувань в якості метрики (рейтингу)

Варіант 11
Опис даних: https://guoguibing.github.io/librec/datasets.html, CiaoDVD
Самі дані: https://guoguibing.github.io/librec/datasets/CiaoDVD.zip

Варіант 12
Опис даних: https://www.kaggle.com/tamber/steam-video-games/
Самі дані: https://www.kaggle.com/tamber/steam-video-games/download