Лабораторна робота №4
з алгоритмів машинного навчання.

Горбунов Даніел

6 червня 2020 р.

1 Вступ.

У даній роботі наведені результати використання моделі ймовірнісної факторизації матриць з метою передбачити значення метрик окремих користувачів. Те, що розуміється під терміном метрики, буде вказано в наступному розділі. Усі обчислення та реалізація моделі були проведені за допомогою Python.

2 Дані.

Маємо таблицю музичних вподобань для користувачів з сайту http://www.last.fm. Серед усіх запропонованих таблиць, обрали одну що зберігалася у файлі "user_artists.dat". Вона складається з трьох колонок:

1. "userID" - коди користувачів, зареєстрованих на last.fm;
2. "artistID" - коди музичних виконавців, чиї пісні доступні для прослуховування;
3. "weight" - кількість прослуховувань пісень різних виконавців. Це наша значення метрики, що потрібно спрогнозувати.

Це типова задача розділу "навчання без учителя". Доречно було перетворити вхідні дані у вигляді матриці та застосувати ймовірнісну факторизацію. Про те, що вдалося отримати, буде вказано в наступному розділі.

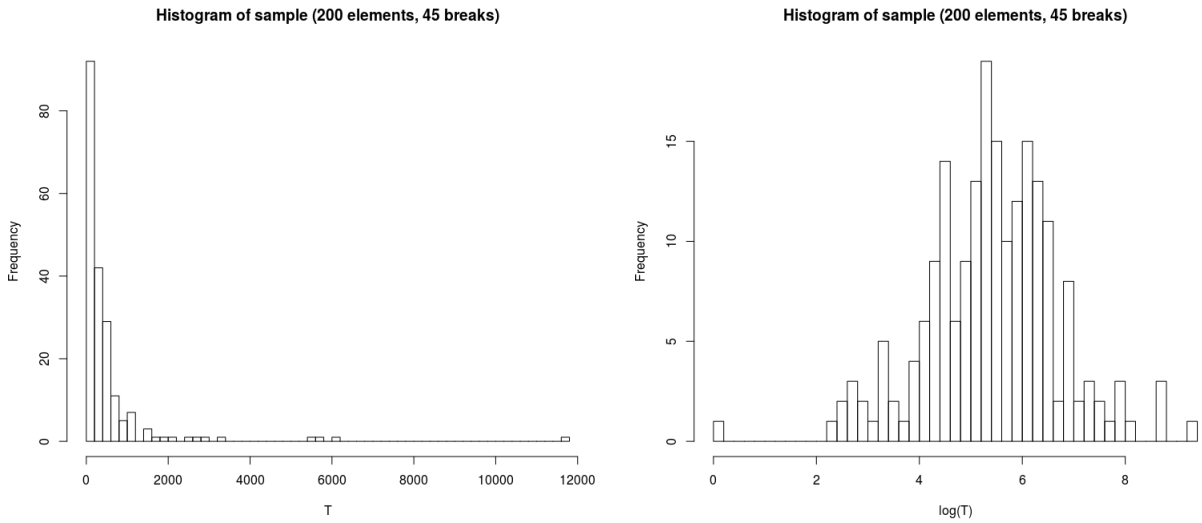

Рис. 1: Гістограми для випадкової вибірки з 200 елементів. Зліва - без модифікацій, справа - після логарифмування.

Маємо досить цікаве спостереження. Спочатку була гіпотеза про те, що розподіл кількості прослуховувань є експоненційним. Однак виявилось, якщо взяти логарифм від елементів вибірки, то форми розподілу та квантилі будуть дещо схожі з тими, що можливі для гаусового розподілу. Тому було висунуте припущення про нормальність логарифму вибірки.

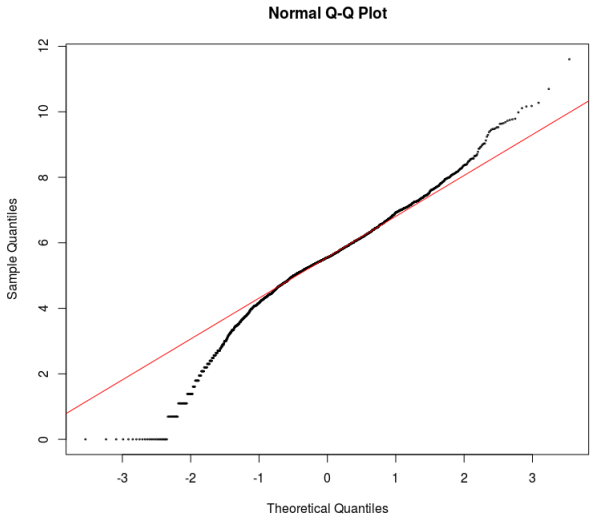

Рис. 2: QQ-діаграма логарифмованих даних, порівняння з квантилями нормального розподілу.

Застосуємо критерій нормальності Шапіро-Вілка для заданої вибірки. Значення статистики та p-значення показані нижче:

```
Shapiro-Wilk normality test
W = 0.98243, p-value = 0.01333
```

Звідси маємо, що для достатньо малого рівня значущості гіпотезу про нормальність відхиляємо. Однак з діаграм можна вважати, що розподіл даних підпорядковується логнормальному закону. Тому надалі припускаємо, що висунуте припущення про нормальність, взагалі кажучи, має місце.

3 Результати PMF.

Підбір оптимального значення параметра d проводився простим ітераційним методом: на кожному кроці тренували модель та визначали її якість середньоквадратичною похибкою. Зрозуміло, що цього не зовсім достатньо, але у рамках роботи цим можна обмежитися.

Алгоритм припиняє роботу, якщо $\exists i_0 : \forall i, j \geqslant i_0$:

$$\left| \mathcal{L}_i - \mathcal{L}_j \right| < \varepsilon := 0.01,$$

де $\mathcal{L}_k$ - значення функції витрат на k-ій ітерації.

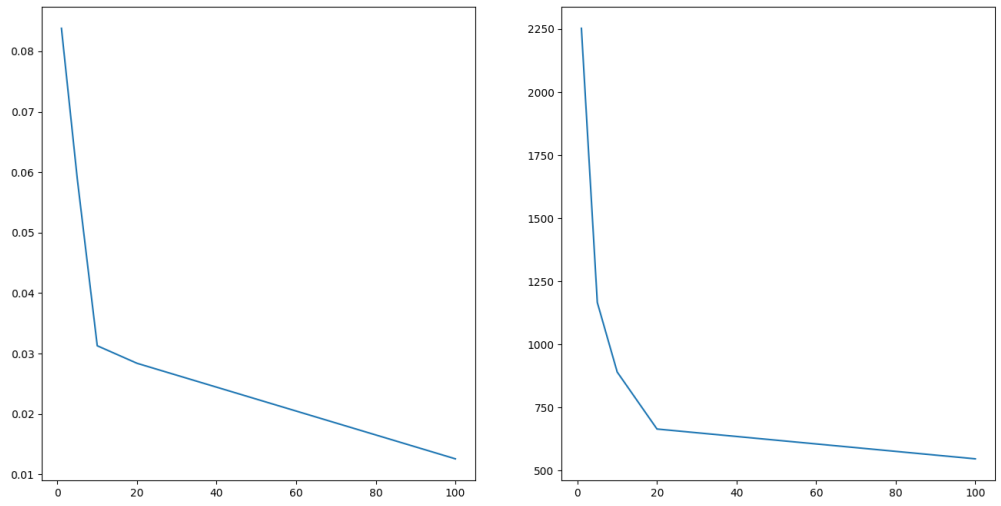

Рис. 3: Графік залежності значення функції витрат (зліва) та RMSE (справа) від числа розмірності d. Показники зафіксовані на третьому циклі.

Нижче наведені результати моделі, яка була натренована на матриці розмірності 2000×4000:

	$\mathcal{L}_i$	$\lvert\mathcal{L}_i - \mathcal{L}_{i-1}\rvert$
d = 1, iter 1	20.7003	20.7003
d = 1, iter 2	20.3061	0.3942
d = 1, iter 3	20.2223	0.0838
Estimated RMSE:		2252.9433
d = 5, iter 1	20.5096	20.5096
d = 5, iter 2	20.1045	0.405
d = 5, iter 3	20.0457	0.0588
Estimated RMSE:		1166.59
d = 10, iter 1	20.2946	20.2946
d = 10, iter 2	19.9898	0.3048
d = 10, iter 3	19.9585	0.0313
Estimated RMSE:		891.0165
d = 20, iter 1	20.2392	20.2392
d = 20, iter 2	19.9727	0.2665
d = 20, iter 3	19.9442	0.0284
Estimated RMSE:		665.1911
d = 100, iter 1	20.1903	20.1903
d = 100, iter 2	19.9032	0.2871
d = 100, iter 3	19.8905	0.0126
Estimated RMSE:		547.1051

4 Висновки.

Модель була коректно імплементована. Похибка висока, але з більшою розмірністю тренувальних матриць та значення параметра d, це виправити можливо. Видно, що для матриці зафікованої розмірності, більше число для d повертало краще передбачення кількості прослуховувань.