BÀI 4 - PHÂN TÍCH MA TRẬN KHÔNG ÂM NMF

Hạn nộp bài: hết ngày 26/04/2024.

1 Bài tập lý thuyết

Bài 1. Trong bài toán NMF, cho trước ma trận $A_{u\times v}$ với các phần tử không âm, ta muốn tìm các ma trận không âm $W_{u\times k}$ và $H_{k\times v}$ với hạng k sao cho

$$A \approx WH,$$
$$k \ll rank(A).$$

Để làm điều này, ta cần tối ưu hàm mục tiêu

$$\|A - WH\|_2^2. \tag{5}$$

a) Giả sử ta đã biết W và cần tìm H. Khi đó với mỗi j ta cần cực tiểu hoá

$$\|A_{\cdot j} - WH_{\cdot j}\|_2^2. \tag{6}$$

Hãy cho biết liên hệ giữa hai biểu thức (5) và (6).

b) Chứng minh rằng nghiệm của bài toán tối ưu này là

$$H_{\cdot j} = \left(W^T W\right)^{-1} W^T A_{\cdot j}. \tag{7}$$

Bài 2. Chứng minh rằng hàm mục tiêu (5) không tăng dưới quy tắc cập nhật

$$W \leftarrow W \circ \frac{AH^T}{WHH^T} \tag{9}$$
$$H \leftarrow H \circ \frac{W^T A}{W^T WH}. \tag{10}$$

2 Bài tập thực hành

Bài 3. Tập dữ liệu 20newsgroups (trong `sklearn.datasets.fetch_20newsgroups`) được lấy từ 18846 bài thảo luận về 20 chủ đề khác nhau.

a) Hãy đọc dữ liệu và xem danh sách các chủ đề trong dữ liệu.

b) Chọn 5 trong 20 chủ đề để lấy dữ liệu. Tiền xử lý và vector hoá dữ liệu đã chọn.

c) Phân tích NMF. Giải thích ý nghĩa của các ma trận W và H.

d) In ra các 10 từ quan trọng nhất ở mỗi chủ đề sau khi phân tích. So sánh với các chủ đề được chọn ở b) và nhận xét.

Bài 4. Tập dữ liệu Olivetti faces (trong `sklearn.datasets.fetch_olivetti_faces`) chứa 400 tấm ảnh gương mặt của 40 người.

a) Hãy đọc dữ liệu và trực quan hoá 10 tấm ảnh đầu tiên (ảnh trắng đen).

b) Dùng NMF để nén ảnh gốc còn 40 chiều. Giải thích ý nghĩa của các ma trận W và H.

c) Tái tạo lại dữ liệu và trực quan hoá 10 tấm ảnh đầu tiên sau khi tái tạo. Nhận xét.