BÀI 2 - Kernel PCA

Hạn nộp bài: hết ngày 01/04/2024.

1 Bài tập lý thuyết

Bài 1. Một hàm đối xứng $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ được gọi là kernel hợp lệ trên $\mathbb{R}^D$ nếu với mọi $x_1, \ldots, x_n \in \mathbb{R}^D$ và $c_1, \ldots, c_n \in \mathbb{R}$ ta luôn có

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0.$$

a) Xét họ các kernel hợp lệ $(K_i)_{i\in\mathbb{N}}$, $K_i : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$. Chứng minh rằng

- Tổng $\sum_{i=1}^{n} \lambda_i K_i$ cũng là kernel hợp lệ, với $\lambda_1, \ldots, \lambda_n > 0$;
- Tích $K_1^{a_1} \cdots K_n^{a_n}$ cũng là kernel hợp lệ, với $a_1, \ldots, a_n \in \mathbb{N}$;
- Giới hạn $K = \lim_{n\to\infty} K_n$ cũng là kernel hợp lệ, nếu giới hạn này tồn tại.

b) Xét họ các kernel hợp lệ $(K_i)_i^n$, $K_i : \mathbb{R}^{D_i} \times \mathbb{R}^{D_i} \to \mathbb{R}$. Chứng minh rằng

$$K((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \prod_{i=1}^{n} K_i(x_i, y_i),$$
$$K((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sum_{i=1}^{n} K_i(x_i, y_i)$$

là các kernel hợp lệ trong không gian tích $\mathbb{R}^{D_1} \times \cdots \times \mathbb{R}^{D_n}$.

c) Xét kernel K hợp lệ trên $\mathbb{R}^D$ và tập $R \subset \mathbb{R}^D$. Chứng minh rằng hàm thu hẹp K_R của K trên R cũng là một kernel hợp lệ.

d) Xét kernel K hợp lệ trên $\mathbb{R}^D$ và $q(\cdot)$ là đa thức có hệ số không âm. Chứng minh các hàm sau là kernel hợp lệ:

$$K'(x, y) = q(K(x, y)),$$
$$K'(x, y) = \exp(K(x, y)).$$

Bài 2. Trong quá trình thiết lập kernel PCA, ta có sử dụng Kernel Trick

$$K(x, y) = \phi(x)\phi(y)^T.$$

Xác định công thức của ϕ và số chiều của không gian thuộc tính (feature space) trong các trường hợp sau:

a) $K(x, y) = \left(1 + x^T y\right)^2$ với $x, y \in \mathbb{R}^3$;

b) $K(x,y) = \exp\left(-\beta \left|x-y\right|^2\right)$ với $x, y \in \mathbb{R}^1$.

Bài 3. Tìm ánh xạ ϕ_1 cho $K_1(x,y) = 1 + x^T y$ và ánh xạ ϕ_2 cho $K_2(x,y) = x^T y + \|x\| \, \|y\|$ với $x, y \in \mathbb{R}^D$. Từ đó suy ra ánh xạ ϕ cho $K(x,y)$ theo ϕ_1, ϕ_2 trong các trường hợp sau:

a) $K(x,y) = 1 + 2x^T y + \|x\| \, \|y\|$;

b) $K(x,y) = \left(1 + x^T y\right)\left(x^T y + \|x\| \, \|y\|\right)$.

Bài 4. Trong quá trình thiết lập kernel PCA, ta có hai đẳng thức

$$K^2 \alpha_j = n\lambda_j K \alpha_j, \tag{1}$$
$$K\alpha_j = n\lambda_j \alpha_j. \tag{2}$$

Rõ ràng, mọi vector α_j thoả mãn (2) thì cũng thoả mãn (1).

a) Chứng minh rằng với bất kỳ nghiệm nào của (2) có trị riêng λ, ta đều có thể cộng thêm một bội số của vector riêng nào đó của K có trị riêng bằng 0 và thu được nghiệm của (1) cũng có trị riêng λ.

b) Chứng minh rằng việc cộng thêm ở câu b không làm ảnh hưởng để phép chiếu lên thành phần chính cho bởi

$$\phi(x)^T v_j = \sum_{i=1}^{n} \alpha_{ji} K(x, x_j).$$

Bài 5. Chứng minh rằng thuật toán PCA thông thường chính là trường hợp đặc biệt của kernel PCA khi chọn hàm kernel tuyến tính cho bởi $K(x_i, x_j) = x_i^T x_j$.

2 Bài tập thực hành

Bài 6. Dùng `sklearn.datasets.make_circles` để tạo ra 200 điểm dữ liệu dạng hình tròn (`factor=0.1`, `noise=0.1`).

a) Dùng PCA và Kernel PCA để tìm phép chiếu làm dữ liệu phân tách tuyến tính.

b) Trực quan hoá dữ liệu gốc, dữ liệu chiếu bằng PCA, dữ liệu chiếu bằng Kernel PCA và dữ liệu tái tạo từ Kernel PCA. Nhận xét.

Bài 7. Tương tự bài trên cho `sklearn.datasets.make_moons`.

Bài 8. Tập dữ liệu Wine (trong `sklearn.datasets.load_wine`) là kết quả phân tích hóa học về rượu vang được trồng trong cùng một vùng ở Ý bởi ba người nông dân.

a) Áp dụng các kernel khác nhau cho Kernel PCA, trực quan hoá và nhận xét.

b) Giảm chiều dữ liệu bằng PCA, trực quan hoá và so sánh với kết quả câu a.