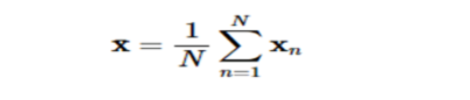
Câu 1: **Trình bày thuật toán PCA**

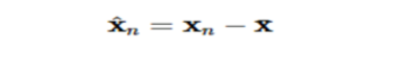
**Ý tưởng chung:** PCA là tìm một không gian mới (với số chiều nhỏ hơn không gian cũ). Các trục tọa độ trong không gian mới được xây dựng sao cho trên mỗi trục, độ biến thiên của dữ liệu trên đó là lớn nhất có thể.

**Các bước thực hiện:**

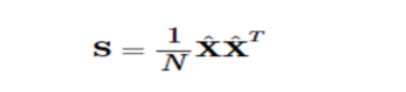
1. Tính vector kì vọng của toàn bộ dữ liệu



1. Trừ mỗi điểm dữ liệu đi vector kì vọng của toàn bộ dữ liệu



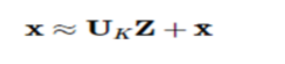
1. Tính ma trận hiệp phương sai:



1. Tính các trị riêng và vector riêng có norm bằng 1 vủa ma trận này, sắp xếp chúng theo thứ tự giảm dần của trị riêng
2. Chọn **K** vector riêng ứng với **K** trị riêng lớn nhất để xây dựng **U**K có các cột tạo thành một hệ trực giao.
3. Chiếu dữ liệu ban đầu đã chuẩn hóa **X** xuống không gian con tìm được
4. Dữ liệu mới chính là tọa độ của các điểm dữ liệu trên không gian mới



Dữ liệu ban đầu có thể tính được xấp xỉ theo dữ liệu mới như sau:



**Minh họa thuật toán PCA**

|  |  |  |
| --- | --- | --- |
| Bước 1: Thu thập dữ liệu | | |
| X | **x0** | **y0** |
| 2.5 | 2.4 |
| 0.5 | 0.7 |
| 2.2 | 2.9 |
| 1.9 | 2.2 |
| 3.1 | 3.0 |
| 2.3 | 2.7 |
| 2 | 1.6 |
| 1 | 1.1 |
| 1.5 | 1.6 |
| 1.1 | 0.9 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Bước 2: Chuẩn hóa dữ liệu** | | | | | |
| **Dữ liệu PCA Gốc** | | | **Dữ liệu PCA chuẩn hóa** | | |
| **X** | **x0** | **y0** | **Xnorm** | **x1** | **y1** |
| 2.5 | 2.4 | 0.2779 | 0.1831 |
| 0.5 | 0.7 | -0.5276 | -0.4520 |
| 2.2 | 2.9 | 0.1571 | 0.3698 |
| 1.9 | 2.2 | 0.0362 | 0.1083 |
| 3.1 | 3.0 | 0.5195 | 0.4072 |
| 2.3 | 2.7 | 0.1973 | 0.2951 |
| 2 | 1.6 | 0.0765 | -0.1158 |
| 1 | 1.1 | -0.3262 | -0.3026 |
| 1.5 | 1.6 | -0.1248 | -0.1158 |
| 1.1 | 0.9 | -0.2859 | -0.3773 |
| **mean** | 1.8100 | 1.9100 | **mean** | 0 | 0 |
| **std** | 0.7852 | 0.8465 | **std** | 0.3162 | 0.3162 |
| **var** | 0.6166 | 0.7166 | **var** | 0.1 | 0.1 |

**Bước 3: Xây dựng ma trận hiệp phương sai**

**Covar =**

|  |  |
| --- | --- |
| 0.1000 | 0.0926 |
| 0.0926 | 0.1000 |

**Bước 4 Xác định trị riêng, vector riêng**

**Eival =**

|  |  |
| --- | --- |
| 0.0074 | 0 |
| 0 | 0.1926 |

**Eivec =**

|  |  |
| --- | --- |
| -0.7071 | 0.7071 |
| 0.7071 | 0.7071 |

**Bước 5 Lựa chọn các thành phần xây dựng vector đặc trưng**

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | | --- | | **0.7071** | | **0.7071** | |
|  | |  |  | | --- | --- | | **0.7071** | **-0.7071** | | **0.7071** | **0.7071** | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Bước 6: Chuyển dữ liệu và không gian mới** | | | | |
| **Dữ liệu PCA Gốc** | | **Dữ liệu PCA chuẩn hóa** | | |
| **X** | **x2T**  **0.3259**  **-0.6927**  **0.3726**  **0.1022**  **0.6553**  **0.3482**  **-0.0278**  **-0.4446**  **-0.1702**  **-0.4690** | **Xnorm** | **x2T** | **y2T** |
| 0.3259 | -0.0671 |
| -0.6927 | -0.0534 |
| 0.3726 | 0.1505 |
| 0.1022 | 0.0510 |
| 0.6553 | -0.0794 |
| 0.3482 | 0.0691 |
| -0.0278 | -0.1360 |
| -0.4446 | 0.0167 |
| -0.1702 | 0.0064 |
| -0.4690 | -0.0646 |