So sánh các phương pháp giảm chiều dữ liệu (Data Dimensional Reduction)

|  |  |  |  |
| --- | --- | --- | --- |
|  | PCA | SVD | LSA |
| Mục đích | Giảm chiều dữ liệu bằng cách tìm các thành phần chính với mục tiêu nắm bắt được phương sai lớn nhất trong dữ liệu. Chuyển đổi dữ liệu sang hệ tọa độ mới. | Giảm chiều dữ liệu bằng cách phân rã ma trận chính thành ba ma trận con. | Giảm chiều dữ liệu bằng cách tìm mối quan hệ ngữ nghĩa giữa các từ và áp dụng phương pháp SVD. |
| Kỹ thuật | Sử dụng ma trận hiệp phương sai để tính toán vecto riêng và trị riêng của ma trận để xác định các thành phần chính. | Phân tích giá trị kỳ dị  A = U∑VT  Giữ lại k giá trị lớn nhất trong ∑. | Áp dụng SVD lên ma trận văn bản để giảm nhiễu và tìm các chủ đề của văn bản. |
| Dữ liệu đầu vào | Dữ liệu dạng số. | Ma trận. | Ma trận từ ngữ. |
| Dữ liệu đầu ra | Ma trận đặc trưng mới. | Ba ma trận (U, ∑, VT) | Không gian ngữ nghĩa mới (dạng vector giảm chiều). |
| Ưu điểm | Giảm chiều dữ liệu hiệu quả, giúp mô hình chạy nhanh hơn. | Hoạt động tốt trong dữ liệu tuyến tính lẫn phi tuyến. Không giới hạn ma trận đầu vào. | Xử lý tốt dữ liệu văn bản, cải thiện tìm kiếm. |
| Nhược điểm | Cần chuẩn hóa dữ liệu trước, hoạt động kém với dữ liệu phi tuyến, không giữ được mối quan hệ của dữ liệu. | Tốn tài nguyên tính toán, nhạy cảm với dữ liệu nhiễu. | Không thể xử lý tốt dữ liệu lớn, nhạy cảm với dữ liệu nhiễu, tốn nhiều tài nguyên tính toán. |
| Ứng dụng | Nhận dạng mẫu, thị giác máy tính, phân cụm, giảm chiều dữ liệu số và phân tích phương sai. | Giải quyết hệ phương trình, nén và khôi phục ảnh, tìm kiếm thông tin. | Xử lý ngôn ngữ tự nhiên (NLP), phân tích ngữ nghĩa văn bản, tìm kiếm tài liệu liên quan. |