**So sánh sự khác nhau giữa K-means, K-medoids và Kernel K-means**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **K-means** | **K-medoids** | **Kernel K-means** |
| **Nguyên lí hoạt động** | Phân chia bộ dữ liệu thành k cụm sao cho tổng khoảng cách Euclidean nhỏ nhất | Tương đối giống K-means, chọn một điểm dữ liệu làm đại diện cho một cụm sao cho khoảng cách trung bình đến các điểm khác nhỏ nhất. | Ánh xạ dữ liệu từ không gian gốc sang không gian đặc trưng và áp dụng phân cụm. |
| **Khoảng cách sử dụng** | Euclidean | Euclidean, Manhattan | Gaussian, Polynomial |
| **Ưu điểm** | * Tốc độ nhanh, hiệu quả, dễ cài đặt. * Phù hợp với dữ liệu phân bố tuyến tính, không nhiễu và dữ liệu số lượng lớn. | * Có thể áp dụng vào dữ liệu có nhiễu, outliers. * Hoạt động tốt với nhiều loại dữ liệu (số, phân loại). | Xử lý tốt dữ liệu có ranh giới phi tuyến phức tạp. |
| **Nhược điểm** | * Nhạy cảm với dữ liệu nhiễu do tính khoảng cách trung bình. * Chỉ dùng được với dữ liệu số và hoạt động kém với dữ liệu dạng phi tuyến. * Chỉ định k cụm trước. | * Tốn nhiều thời gian do cần tính toàn bộ khoảng cách. * Không hiệu quả với dữ liệu lớn. | * Tốn tài nguyên hơn do ánh xạ kernel. * Cần chọn kernel phù hợp. * Khó trực quan hóa dữ liệu. |
| **Ứng dụng** | * Phân tích thị trường. * Xử lý ảnh. | * Phân tích văn bản. * Phân tích chứng khoán. | * Phân tích gen sinh học. * Nhận dạng khuôn mặt. |