# Notes: Kernel PCA & Multi-Dimensional Scaling (MDS)

## 1. Tổng quan

Hai kỹ thuật giảm chiều nâng cao trong Unsupervised Learning là Kernel PCA và Multi-Dimensional Scaling (MDS). Chúng được dùng để biểu diễn dữ liệu trong không gian có ít chiều hơn nhưng vẫn giữ nguyên mối quan hệ hoặc cấu trúc của dữ liệu.

## 2. Kernel Principal Component Analysis (Kernel PCA)

Kernel PCA là phần mở rộng của PCA truyền thống, sử dụng kỹ thuật 'kernel trick' để ánh xạ dữ liệu vào không gian có chiều cao hơn (high-dimensional space). Điều này giúp mô hình xử lý được các quan hệ phi tuyến tính giữa các biến.

Đặc điểm của Kernel PCA:  
• Có nhiều loại kernel: linear, polynomial (poly), radial basis function (RBF), sigmoid, cosine.  
• Kernel PCA không đảm bảo tái tạo hoàn hảo dữ liệu ban đầu như PCA thông thường.  
• Có thêm các siêu tham số như gamma, alpha, ảnh hưởng trực tiếp đến kết quả biến đổi.

### Cấu trúc lệnh trong Python:

from sklearn.decomposition import KernelPCA  
  
kernel\_pca = KernelPCA(kernel='rbf', gamma=10, fit\_inverse\_transform=True, alpha=0.1)  
kernel\_pca.fit(X\_train)  
X\_transformed = kernel\_pca.transform(X\_test)

Trong đó:  
- kernel: loại hàm kernel được sử dụng.  
- gamma: điều chỉnh mức độ ảnh hưởng của các điểm dữ liệu.  
- alpha: hệ số điều chuẩn để giảm overfitting.  
- fit\_inverse\_transform: cho phép khôi phục dữ liệu gần đúng về dạng ban đầu.

## 3. Multi-Dimensional Scaling (MDS)

MDS là kỹ thuật giảm chiều dữ liệu giúp bảo toàn khoảng cách (dissimilarity) giữa các điểm dữ liệu khi chiếu xuống không gian thấp hơn. Mục tiêu của MDS là biểu diễn dữ liệu trong không gian 2D hoặc 3D sao cho mối quan hệ giữa các điểm vẫn được giữ nguyên.

Có hai biến thể chính của MDS:  
• Metric MDS: giữ nguyên tỷ lệ khoảng cách thực tế giữa các điểm.  
• Non-Metric MDS: sử dụng hàm f(x) để biến đổi khoảng cách trước khi giảm thiểu sai số, giúp xử lý dữ liệu phi tuyến tính tốt hơn.

### Cấu trúc lệnh trong Python:

from sklearn.manifold import MDS  
  
embedding = MDS(dissimilarity='precomputed', random\_state=0, n\_components=2, metric=False)

Trong đó:  
- dissimilarity='precomputed': sử dụng ma trận khoảng cách tính sẵn.  
- n\_components: số chiều sau khi giảm.  
- metric=False: thực hiện Non-Metric MDS.

## 4. So sánh Kernel PCA và MDS

• Kernel PCA: tập trung vào ánh xạ dữ liệu sang không gian cao hơn để phát hiện các mẫu phi tuyến tính.  
• MDS: bảo toàn khoảng cách giữa các điểm khi biểu diễn trong không gian thấp hơn.  
• Kernel PCA thường hiệu quả hơn cho dữ liệu phi tuyến tính phức tạp.  
• MDS phù hợp hơn cho trực quan hóa dữ liệu (visualization).

## 5. Tổng kết

Kernel PCA và MDS là hai công cụ mạnh mẽ giúp giảm chiều dữ liệu mà vẫn giữ nguyên cấu trúc thông tin. Kernel PCA sử dụng hàm kernel để xử lý phi tuyến, trong khi MDS giúp trực quan hóa dữ liệu bằng cách bảo tồn mối quan hệ khoảng cách giữa các điểm.