1. Giới thiệu

Dimensionality Reduction (giảm chiều dữ liệu) là một kỹ thuật quan trọng trong Machine Learning.
Dữ liệu thực tế có thể có số chiều rất lớn (hàng nghìn). Việc giảm chiều giúp tiết kiệm lưu trữ, tăng tốc tính toán và có thể coi như nén dữ liệu.
Một phương pháp tuyến tính cơ bản là Principal Component Analysis (PCA).

2. Một chút toán

2.1 Norm 2 của ma trận

$$||A||_2 = \max_x ||Ax||_2 / ||x||_2 \quad (1)$$

$$||A||_2 = \max_{\{||x||_2=1\}} ||Ax||_2 \quad (2)$$

Giải bằng nhân tử Lagrange cho thấy norm 2 của ma trận chính là singular value lớn nhất của A.
Vector tương ứng là right-singular vector của A.

2.2 Biểu diễn vector trong các hệ cơ sở khác nhau

$$x = Uy, \quad y = U^{-1}x \quad (7)$$

Nếu U trực giao: $U^{-1}=U^T$, do đó $y = U^T x$.

2.3 Trace

Một số tính chất:
- trace(A) = trace(A^T)
- trace(kA) = k trace(A)
- trace(AB) = trace(BA)
- $||A||_F^2$ = trace(A^T A)
- trace(A) = tổng các trị riêng của A

2.4 Kỳ vọng và ma trận hiệp phương sai

Một chiều: $\bar{x} = (1/N) \Sigma x_n$, $\sigma^2 = (1/N) \Sigma (x_n - \bar{x})^2$
Đa chiều: $\bar{x} = (1/N) \Sigma x_n$, $S = (1/N) (X - \bar{x}1^T)(X - \bar{x}1^T)^T$

3. Principal Component Analysis (PCA)

Mục tiêu: Tìm hệ cơ sở trực chuẩn sao cho phương sai dữ liệu tập trung ở K thành phần đầu.
Dữ liệu chuẩn hoá: $\dot{X} = X - \bar{x}1^T$
Ma trận hiệp phương sai: $S = (1/N)\dot{X}\dot{X}^T$

Hàm mất mát: $J = \Sigma_{i=K+1}^{D} u_i^T S u_i$

Tối ưu tương đương chọn K vector riêng ứng với K trị riêng lớn nhất của S.

4. Các bước PCA

- Tính kỳ vọng $\bar{x}$
- Chuẩn hoá dữ liệu: $\dot{X} = X - \bar{x}1^T$
- Tính ma trận hiệp phương sai S
- Tính trị riêng & vector riêng, sắp xếp λ giảm dần
- Chọn K vector riêng lớn nhất $\rightarrow U_K$
- Tính toạ độ mới: $Z = U_K^T \dot{X}$
- Xấp xỉ khôi phục: $x \approx U_K z + \bar{x}$