Giảm chiều dữ liệu (tiếng Anh: dimensionality reduction, hay dimension reduction), là sự biến đổi dữ liệu từ không gian chiều-cao thành không gian chiều-thấp để biểu diễn ở dạng chiều-thấp đồng thời giữ lại một số thuộc tính có ý nghĩa của dữ liệu gốc, có ý tưởng là gần với chiều nội tại (intrinsic dimension).

Phân tích dữ liệu trong không gian chiều-cao có thể khó khăn vì nhiều lý do; dữ liệu thô thường có tính thưa thớt (sparse matrix) là một hậu quả của lời nguyền chiều, và do đó việc phân tích thường khó tính toán; hơn nữa các thuật toán có thể mất rất nhiều thời gian để xử lý dữ liệu. Giảm chiều dữ liệu là phổ biến trong các lĩnh vực có số lượng quan sát lớn và/hoặc số lượng biến lớn, chẳng hạn như xử lý tín hiệu, nhận dạng tiếng nói, thông tin học thần kinh (tin học thần kinh, neuroinformatics), và tin sinh học.

Các phương pháp giảm chiều dữ liệu thông thường được chia thành cách tiếp cận tuyến tính và phi tuyến tính. Các cách tiếp cận cũng được chia thành chọn đặc tính (feature selection) và trích chọn đặc trưng (feature extraction). Giảm chiều dữ liệu có thể được sử dụng cho giảm nhiễu (noise reduction), trực quan hóa dữ liệu (data visualization), phân tích cụm, hoặc là một bước trung gian để tạo điều kiện thuận lợi cho các phân tích khác.

số lượng các điểm dữ liệu cũng thường rất lớn. Nếu thực hiện lưu trữ và tính toán trực tiếp trên dữ liệu có số chiều cao này thì sẽ gặp khó khăn cả về việc lưu trữ và tốc độ tính toán. Vì vậy, giảm số chiều dữ liệu là một bước quan trọng trong nhiều bài toán. Đây cũng được coi là một phương pháp nén dữ liệu.

Dimensionality Reduction, nói một cách đơn giản, là việc đi tìm một hàm số, hàm số này lấy đầu vào là một điểm dữ liệu ban đầu [Math Processing Error] với [Math Processing Error] rất lớn, và tạo ra một điểm dữ liệu mới [Math Processing Error] có số chiều [Math Processing Error].

Ví dụ:

Dữ liệu có nhiều biến, nhiều đặc trưng

▰ Văn bản

▻ 1-gram: số chiều bằng số từ

▻ 2-gram: số chiều bằng bình phương số từ

▰ Bài toán gợi ý sản phẩm

▻ Bảng dữ liệu: người dùng (dòng) x sản phẩm (cột)

▻ Netflix: 480189 người dùng x 17770 bộ phim

Dữ liệu có nhiều biến, nhiều đặc trưng :

▰ Hành vi người sử dụng

▻ Một hoặc nhiều chuỗi thời gian

▻ Ví dụ: lưu lượng sử dụng 3G theo giờ

▰ Ảnh cộng hưởng từ (MRI)

▻ Vài chục nghìn voxels (điểm trong 3 chiều) trở lên