Module: Principal Component Analysis (PCA)

Welcome! In this module, we'll explore **Principal Component Analysis (PCA)**, a fundamental and widely used technique in unsupervised learning for **dimensionality reduction**. We'll learn how PCA can help us simplify complex datasets while retaining most of the important information.

Structure of this Module

Our journey through PCA will cover these key topics:

1. **Introduction to PCA** *(Current Section)*
2. PCA Process Steps
3. Data Standardization (Requirement for PCA)
4. Finding Covariance Matrix of our Dataset
5. Eigenvectors and Eigenvalues
6. Recast Data using new PCs (Principal Components)
7. Explained Variance Ratio and Scree Plot

PCA - Definition and Purpose

**Principal Component Analysis (PCA)** is an **unsupervised dimensionality reduction algorithm**. Its primary goal is to transform a dataset with many features (high dimensions) into a new dataset with fewer features (lower dimensions) while preserving as much of the original dataset's **variance** (information) as possible.

Essentially, PCA aims to find a **more meaningful basis or coordinate system (axis)** for our data. It identifies the directions (principal components) along which the data varies the most. These principal components are linear combinations of the original features.

Why Dimensionality Reduction? The Curse of Dimensionality

PCA is particularly useful when we need to tackle the **'curse of dimensionality'**. This refers to various problems that arise when working with high-dimensional data (datasets with many features):

* **Model Complexity & Overfitting:** A higher number of dimensions means more parameters for a model to learn, increasing the risk of overfitting the training data and reducing generalization accuracy. Dimensionality reduction simplifies the model.
* **Computational Cost:** Training machine learning models on high-dimensional data can be very **slow** and require significant computational resources (memory and processing power).
* **Noise and Redundancy:** High-dimensional datasets often contain irrelevant features (noise) or highly correlated features (redundancy). PCA can help filter out noise and combine redundant information.
* **Visualization Difficulties:** Humans cannot visualize data beyond 3 dimensions. Dimensionality reduction (especially down to 2 or 3 dimensions) is essential for visualizing complex datasets to gain insights.
* **Sensitivity to Scale:** The 'curse' can be exacerbated when features have **different scales** (e.g., mixing weight, length, area, temperature, counts). The distances or variances calculated can be dominated by features with larger values.

Categories of Dimensionality Reduction

There are two main approaches to reducing dimensionality:

1. **Feature Selection:** We **select a subset** of the *original* features and discard the rest. We keep the most relevant features based on certain statistical criteria or model performance.
2. **Feature Extraction:** We derive information from the *entire* feature set to construct a **new, smaller set of features** (a feature subspace). These new features are combinations or transformations of the original features.

**PCA is a Feature Extraction method.** It creates new features (Principal Components) that are linear combinations of the original features, ordered by the amount of variance they capture.

How PCA Works (Briefly)

PCA works by analyzing the relationships between features. It finds the principal components by looking at the **covariance matrix** of the dataset (or performing Singular Value Decomposition - SVD). The first principal component captures the direction of maximum variance, the second captures the maximum remaining variance while being orthogonal (uncorrelated) to the first, and so on. By keeping only the first few principal components that capture most of the variance, we achieve dimensionality reduction.