Week 8 Learning Reflection

Summary:

This week introduced unsupervised learning techniques, with a focus on Principal Component Analysis (PCA) as a method for reducing dimensionality in datasets without using a response variable. PCA helps us find new features (principal components) that capture the most variance in the data, allowing us to visualize and interpret high-dimensional datasets more effectively. We also explored how these components are derived from the covariance matrix, and why they are useful for summarizing and simplifying complex datasets.

Concepts:

- Unsupervised Learning: Learning patterns from data without any labeled outcomes. PCA is a core example, used for structure discovery and dimension reduction.

- Principal Component Analysis (PCA):

    - A technique for finding a small number of uncorrelated linear combinations (principal components) of the original variables that retain most of the variability.

    - The first principal component is the direction (a linear combination of features) that maximizes the variance of the projected data.

    - Subsequent components are orthogonal to the previous ones and account for as much remaining variance as possible.

- Covariance Matrix and Eigenvectors:

    - PCA is performed by computing the eigenvectors and eigenvalues of the covariance matrix (or correlation matrix when variables are on different scales).

    - Eigenvectors define the directions (principal axes), and eigenvalues determine the amount of variance explained by each component.

- Interpreting PCA:

    - The proportion of variance explained (PVE) tells us how much information is retained by each component.

    - A scree plot can help visualize the PVE and decide how many components to retain.

- Centering and Scaling:

    - It is standard to center variables (subtract the mean) before PCA.

    - Scaling (dividing by standard deviation) is needed when variables are measured on different units.

- Applications of PCA:
    - Data compression
    - Visualization of high-dimensional data
    - Preprocessing step before supervised learning

Uncertainties:

While I understand the theoretical formulation of PCA, I'm still unsure how to make practical decisions like:

- When to scale variables before applying PCA.
- How to interpret loadings when many variables contribute to a component.
- How to decide on the number of components to retain in practice, beyond the scree plot.