VAN ECCELPOEL
Théo

Advanced Machine Learning
Lab 7 – PCA
Report & Analysis

In this report, I will analyze the lab I conducted on Principal Component Analysis (PCA). PCA is a widely used dimensionality reduction technique that reduces the number of dimensions in a dataset while preserving as much variance as possible. By doing so, PCA helps simplify data, making it easier to analyze, visualize, or preprocess for machine learning models.
Dimensionality reduction takes various forms, such as clustering, Non-negative Matrix Factorization (NMF), PCA, or even Gaussian mixture modeling. In this lab, I explored PCA in depth, focusing on its theoretical foundations, practical applications, and limitations.

1) Theory

The core idea of PCA is to project high-dimensional data into a lower-dimensional space while maximizing the retained variance. It achieves this by finding new variables, called principal components**,** which are linear combinations of the original features. These components are orthogonal to each other and ordered by the amount of variance they capture in the dataset.

To compute the principal components, we rely on mathematical tools like Singular Value Decomposition (SVD) or eigenvalue decomposition. These methods identify the eigenvectors (directions of maximum variance) and eigenvalues (amount of variance along these directions). By sorting the eigenvectors based on their corresponding eigenvalues, we select the top kk eigenvectors that capture the most variance to form our reduced-dimensional space.

This lab involved implementing a PCA class from scratch in Python, which helped solidify my understanding of these principles. I gained hands-on experience with key steps such as performing eigen-decomposition, sorting eigenvalues, and truncating the eigenvectors to focus on the most informative ones.

2) Application in biostatistics

The biostatistics application was particularly insightful, as it demonstrated not just the mechanics of PCA but also how to make informed decisions about the number of principal components to retain.

I learned that choosing the rank r of the reduced matrix involves balancing dimensionality reduction with information retention. A common approach is to retain 80–90% of the total variance. This was straightforward to compute by examining the cumulative variance explained by the top kk principal components. Alternatively, the elbow curve technique provides a visual method for identifying the point of diminishing returns in the explained variance.

One key takeaway was the utility of PCA for data visualization**.** By reducing data to 2 or 3 dimensions, PCA makes it easier to explore patterns and relationships. For instance, in the biostatistics dataset, plotting the first two principal components revealed a clear distinction between malignant and benign cases. Additionally, the ability to trace back which original features contribute most to each principal component gives PCA interpretability, making it valuable as we can give domain-specific insights to the components.

An interesting observation during the lab was the sensitivity of PCA to features with constant variance or noise. Such features do not contribute meaningful information but can skew the results, emphasizing the need for careful preprocessing, such as removing low-variance features or normalizing the data.

3) Limitations

Despite its usefulness, PCA has notable limitations. Its linear nature means it assumes linear relationships between features, making it unsuitable for capturing complex, nonlinear structures in the data. As a result, while PCA preserves the global structure of the data (the overall distribution), it often fails to capture the local structure **(**relationships among nearby points).

In the lab, this limitation was evident when applying PCA to the MNIST dataset. PCA struggled to separate clusters of handwritten digits, whereas nonlinear methods like t-SNE performed much better, clearly grouping and separating clusters. t-SNE works by converting pairwise distances into probabilities, which helps it uncover local patterns in the data. However, t-SNE has its drawbacks: it sacrifices global structure for local structure, making it ideal for visualization but less reliable for tasks requiring interpretability.

Further exploration I made introduced me to another technique called UMAP (Uniform Manifold Approximation and Projection)**,** which aims to balance the preservation of local and global structures. This makes UMAP a promising alternative for dimensionality reduction in complex datasets.