Qingyu Zhang / Andy

N-number: N19903322

Prof. Pascal Wallisch

Fundamentals of Machine Learning

HW5 Report: Unsupervised Learning

Question 1:

1） What was done:

We first standardized the original 13-dimensional wine dataset using StandardScaler, then applied Principal Component Analysis (PCA) to project the data into two dimensions. After fitting PCA, we examined the eigenvalues to determine how many principal components have eigenvalues greater than 1. We also calculated how much of the total variance is explained by the top two principal components and visualized the 2D projection.

2） Why this was done:

Standardization is crucial before applying PCA because the features in the wine dataset (e.g., alcohol percentage, magnesium concentration) have different units and scales. Without standardization, features with larger numerical ranges would dominate the principal components, distorting the true structure of the data. PCA was then used to reduce dimensionality, which helps simplify the data visualization and analysis while retaining as much information as possible. Examining eigenvalues (>1 rule) helps identify the number of meaningful dimensions according to the Kaiser criterion.

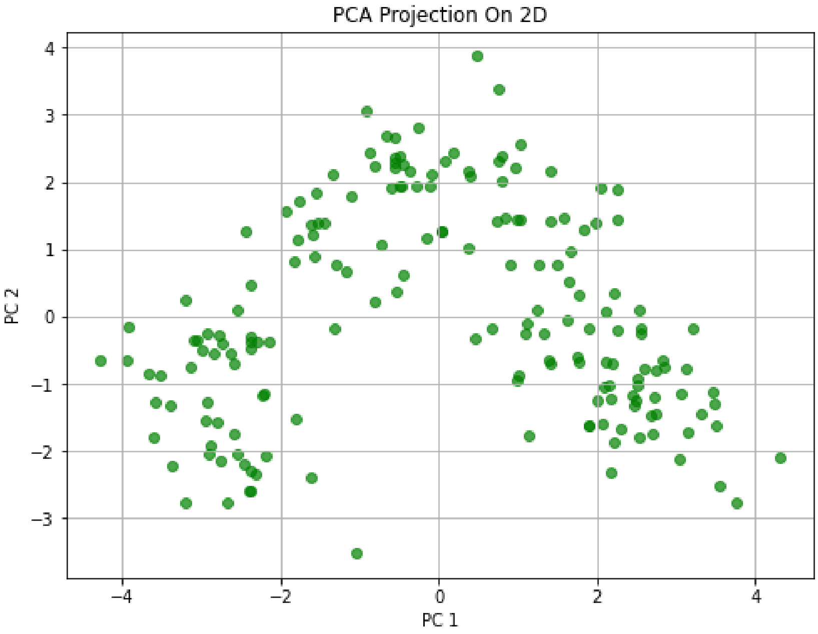
3) Findings:

- **Eigenvalues:** [4.732, 2.511, 1.454, 0.924, 0.858, 0.645, 0.554, 0.351, 0.291, 0.252, 0.227, 0.170, 0.104]

- **Number of eigenvalues > 1: 3**

- Variance explained by **top 2** principal components: **0.5541 (55.41%)**

- PCA 2D Projection:



4) Interpretation:

The PCA results show that only **3 components** have eigenvalues above 1, suggesting that the intrinsic dimensionality of the dataset is relatively low. The first two principal components alone explain **55.41%** of the total variance, meaning that over half the information is captured in a 2D representation. The scatter plot reveals broad groupings, indicating that wines cluster based on their chemical properties even without supervision. However, around 44% of the variance remains unexplained in 2D, suggesting more subtle structures exist in higher dimensions.

Question 2:

1) What was done:

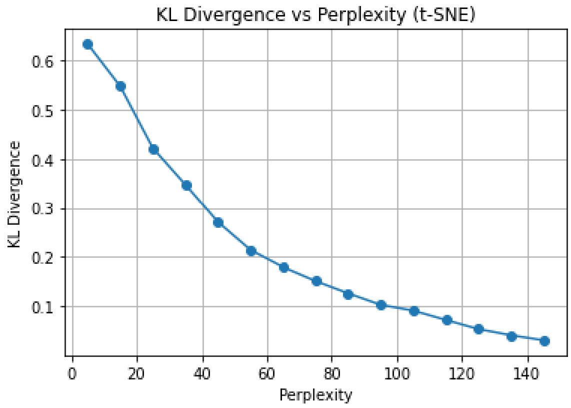
   After standardizing the wine dataset, we applied t-distributed Stochastic Neighbor Embedding (t-SNE) to project the data into two dimensions. We varied the perplexity parameter from 5 to 150 in steps of 10 and recorded the KL-divergence at each setting. We plotted KL-divergence versus perplexity to assess how perplexity affects the embedding, and generated a 2D scatter plot for perplexity = 20.

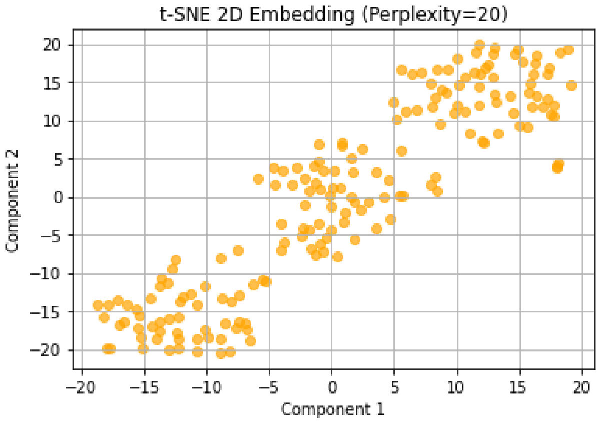
2) Why it was done:

   Standardization ensures that all features contribute equally when computing pairwise similarities, as t-SNE is sensitive to raw feature scales. Varying perplexity allows us to explore how local versus global structure preservation trade-offs affect the quality of embedding. Lower KL-divergence generally indicates a better preservation of the original pairwise similarities.

3) Findings:

- **KL-divergence decreases** consistently as perplexity increases.
- **2D t-SNE embedding** generated at **perplexity = 20.**





4) Interpretation:

As perplexity increases, KL-divergence decreases, showing that higher perplexity settings better capture the global structure but might oversmooth local clusters. At perplexity = 20, the 2D t-SNE plot reveals elongated, distinct groupings of wines, indicating meaningful local structures. Unlike PCA, the t-SNE embedding is non-linear, making it effective for identifying non-linear manifolds within the wine chemical feature space.

Question 3:

1) What was done:

   To analyze the underlying structure of the wine dataset, we first standardized all 13 features to zero mean and unit variance using StandardScaler. This step ensured that attributes measured in different units (such as alcohol percentage and magnesium content) contributed equally to distance calculations. After standardization, we applied classical Multidimensional Scaling (MDS) to the dataset to create a two-dimensional embedding. MDS attempts to arrange points in a way that best preserves the pairwise Euclidean distances from the original high-dimensional space. The resulting two-dimensional configuration was visualized using a scatter plot. Additionally, we recorded the **stress value**, a metric indicating the level of distortion between the original and embedded pairwise distances.

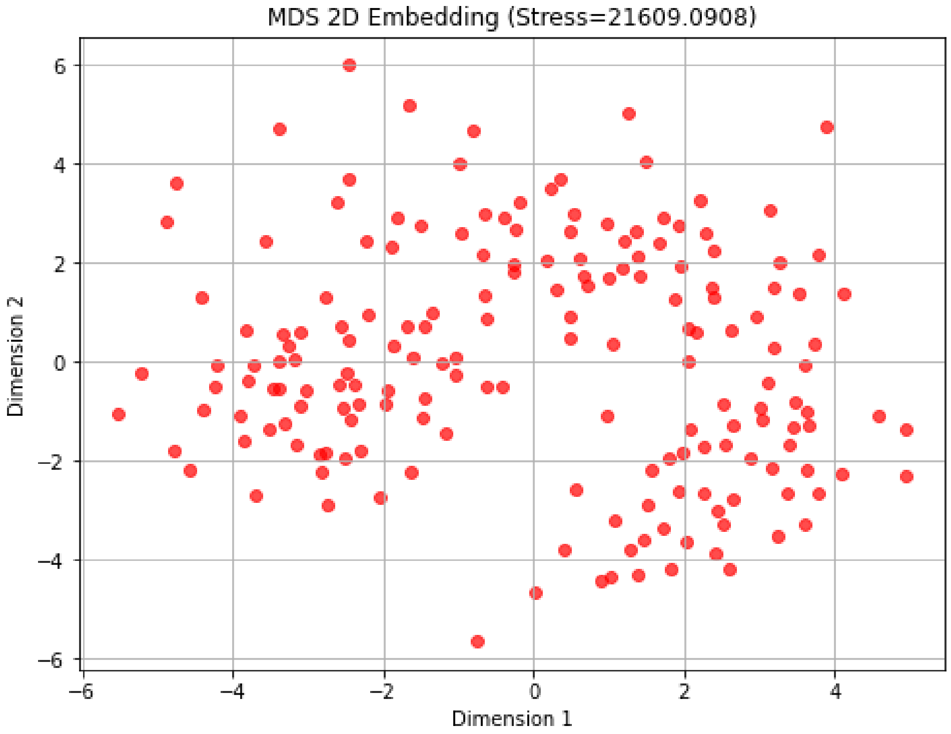
2) Why it was done:

   Standardization was essential because MDS relies heavily on distance computations, and features with larger scales would otherwise disproportionately influence the results. Classical MDS was chosen because it seeks to preserve **global distance relationships** — not just local neighbors — allowing us to understand the overall structure of the wine samples. Reducing the dimensionality to two dimensions enables visual exploration while keeping the computational complexity manageable, especially given the relatively small dataset size (178 samples).

3) Findings:

- **Stress: 21,609.0908**

The scatter plot of the 2D MDS embedding was generated to visualize the configuration of wines in the reduced space.



4) Interpretation:

The **high stress value of 21,609.0908** indicates that it was challenging to accurately preserve the original 13-dimensional pairwise distances when embedding the data into only two dimensions. The MDS scatter plot shows that the wine samples are **spread relatively uniformly**, without distinct, compact clusters emerging. Compared to the t-SNE visualization — which revealed clearer subgroup separations — MDS appears to **retain more of the global distance structure** at the cost of local cluster detection. This suggests that while MDS is effective for maintaining overall spatial relationships among samples, it is **less suited for discovering localized groupings** (i.e., tightly knit clusters) within complex, high-dimensional datasets like this one. Therefore, techniques like t-SNE may be preferred when the goal is to **identify clear clusters or subpopulations**.

Question 4:

1) What was done:

   Using the PCA 2D-projected data, we applied k-Means clustering with k ranging from 2 to 14. For each k, we computed the silhouette score to measure clustering quality and selected the k value with the highest silhouette score. The final k-Means clustering result was plotted.

2) Why it was done:

   Silhouette analysis provides an objective metric to evaluate and select the optimal number of clusters, balancing cluster cohesion and separation.

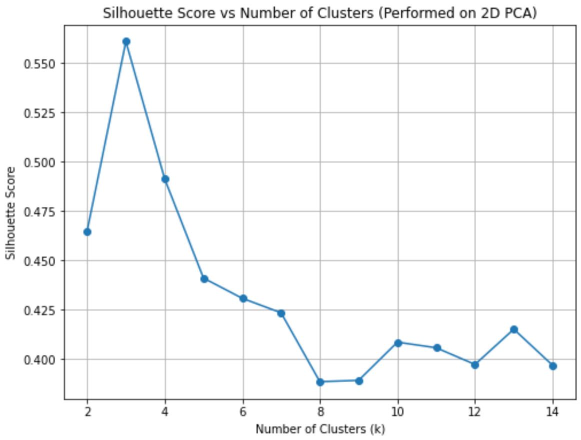
Standardization and PCA were crucial because k-Means clustering is highly sensitive to feature scaling and assumes isotropic, spherical clusters formed based on Euclidean distances. We chose to apply k-Means on the **PCA-reduced 2D space** rather than on the **MDS or t-SNE** embeddings for several reasons:

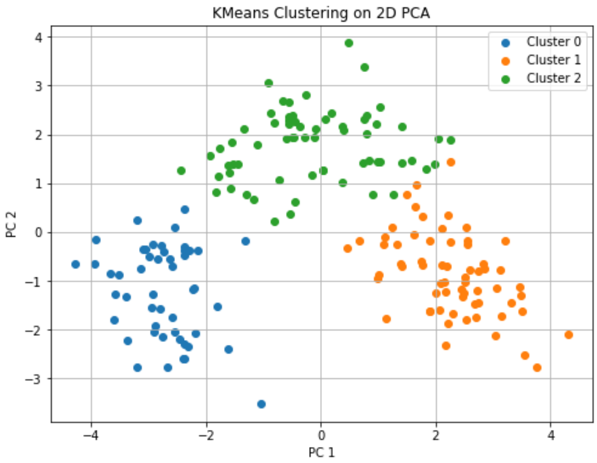
- **PCA preserves global variance and approximate Euclidean structure,** making the transformed space appropriate for distance-based clustering methods like k-Means.
- **MDS,** although it attempts to preserve pairwise distances, showed a very high stress value in our results, indicating poor 2D distance preservation, making it less reliable for clustering.
- **t-SNE**, while effective at revealing local clusters visually, **distorts global distances** and **does not preserve metric properties** required for clustering algorithms like k-Means to function meaningfully.

Thus, PCA offers a **balanced trade-off** between dimensionality reduction, interpretability, and maintaining the assumptions necessary for effective k-Means clustering. Moreover, applying clustering directly on PCA 2D output keeps the process computationally simple and transparent.

3) Findings:

- **Best number of clusters (k) picked by Silhouette Score: 3**
- **Total sum of squared distances to cluster centers (Inertia): 259.5094**





4) Interpretation:

The silhouette analysis indicates that **k = 3** yields the highest clustering quality among the candidate values. The corresponding inertia value is **259.5094,** reflecting that the points within each cluster are relatively close to their respective cluster centers. The PCA scatter plot shows that the wines are clearly separated into three distinct groups, suggesting natural stratification based on chemical composition.

It is worth noting that while the assignment prompt asks for the "total sum of the distance of all points to their respective cluster centers," the standard output from k-Means reports the **inertia**, which is the **sum of squared distances** rather than simple distances. Although squared distances and raw distances differ mathematically, they both effectively measure the compactness of the clusters. Thus, reporting the inertia value is appropriate in this context and aligns with standard practice in clustering analysis.

Question 5:

1) What was done:

   We constructed a K-distance graph using 5th nearest neighbor distances in the PCA 2D space to determine a suitable epsilon value. We chose epsilon corresponding to approximately the 90th percentile. Then, we tuned the min_samples parameter between 4 and 10 to maximize the silhouette score. Finally, DBSCAN clustering was applied using the optimized hyperparameters, and the clustering results were plotted.

2) Why it was done:

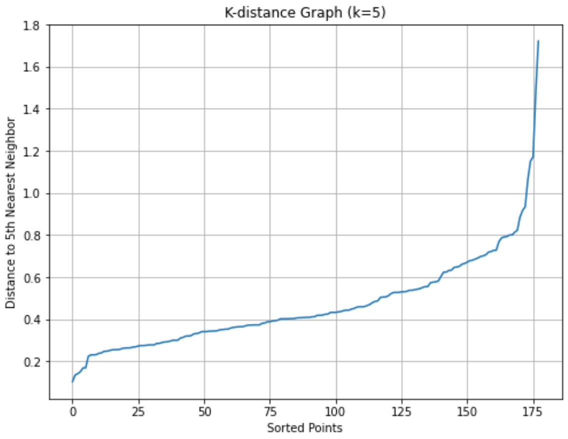
   Standardization was essential because DBSCAN relies on accurate distance measurements; features with larger scales would otherwise distort cluster shapes. We used PCA for dimensionality reduction because it preserves global linear structure and reduces noise, making clustering more efficient and interpretable.

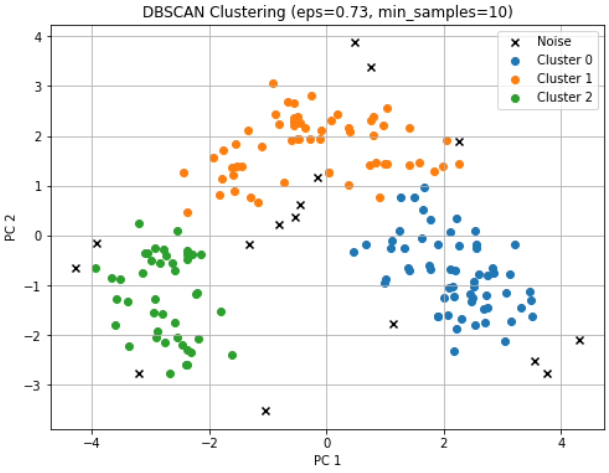
   The K-distance method was chosen to select epsilon systematically based on the dataset's density profile, rather than guessing arbitrarily. Optimizing min_samples based on silhouette scores helped ensure that the resulting clusters were both compact and well-separated, improving clustering robustness.

   Compared to k-Means, DBSCAN offers the advantage of identifying arbitrarily shaped clusters and naturally handling noise points, making it particularly suitable for datasets with variable density.

3) Findings:

- **Chosen epsilon (eps): 0.7271**
- **Best min_samples: 10**
- **Best silhouette score: 0.5787**
- **Estimated number of clusters (excluding noise points): 3**
- **Number of noise points: 16**





4) Interpretation:

The optimized DBSCAN model identified **three main clusters** among the wine samples, consistent with the findings from k-Means clustering. In addition, **16 wines were classified as noise points**, highlighting DBSCAN's ability to detect outliers that do not fit well into any cluster.

The chosen parameters — epsilon ≈ **0.7271** and min_samples = **10** — produced a **silhouette score of 0.5787**, indicating reasonably compact and well-separated clusters. Compared to k-Means, DBSCAN provides a more flexible clustering solution that does not require assuming spherical cluster shapes and is robust against noise and border points.

Overall, DBSCAN revealed a meaningful and interpretable clustering structure in the wine dataset while also exposing rare or atypical samples.

Extra Credit:

(a) Based on the consistent findings across multiple methods — including PCA projections, t-SNE embeddings, k-Means clustering, and DBSCAN clustering — the wine dataset appears to naturally separate into three distinct groups.

Both k-Means and DBSCAN clustering results support three clusters, while the PCA and t-SNE visualizations also suggest three prominent groupings.

Therefore, it is reasonable to conclude that there are approximately three different kinds of wine represented in this dataset, likely corresponding to different wine types or grape varieties with distinct chemical profiles.

(b) One interesting observation from this unsupervised exploration is that the wines do not appear to form strictly discrete categories.

In both PCA and especially t-SNE embeddings, some clusters show elongated, transitional shapes rather than compact, spherical groups.

This suggests that certain chemical properties of the wines — likely related to factors such as Proline content, Color Intensity, and Alcohol concentration — change in a continuous gradient rather than forming completely separate wine types.

Moreover, several samples appear positioned between major clusters in the low-dimensional embeddings, indicating that hybrid or intermediate wine styles may exist in the dataset.

This continuum-like behavior would be difficult to detect using supervised learning or hard clustering alone, but was revealed naturally through unsupervised techniques.

Thus, an important insight from this project is that wines may not always belong to sharply distinct types, but instead vary gradually along key chemical axes, reflecting a richer and more complex structure in wine classification.