MATH/CSCI 485 Assignment #2: PCA and Dimensionality Reduction Report

Introduction

This report analyzes the Wine Quality dataset using Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction. The dataset contains physicochemical properties of wine samples along with their quality ratings.

Data Preprocessing

The analysis combined both red and white wine datasets, with basic statistics showing:

- 6,497 total samples
- 11 features with varying ranges and distributions
- Quality ratings ranging from approximately 3-9
- No missing values were found

All features were normalized to ensure they were on the same scale for dimensionality reduction.

PCA Analysis

Explained Variance

PCA was applied to the normalized dataset, revealing that:

- The first principal component (PC1) explains 27.5% of the variance
- The first two principal components together explain 50.2% of the variance
- The first three principal components together explain 64.4% of the variance

This indicates that we can capture a significant portion of the dataset's variability with just 2-3 principal components, though there is still substantial information loss.

Trade-off between Dimensionality Reduction and Information Loss

When reducing from 11 dimensions to 2 dimensions, we lose approximately 49.8% of the information in the dataset. This is a significant trade-off that must be considered when interpreting the results. The benefit is a much more interpretable visualization and reduced computational complexity for subsequent analyses.

When using 3 dimensions, we retain 64.4% of the information, which is a reasonable compromise between dimensionality reduction and information preservation. Each additional dimension provides diminishing returns in terms of explained variance, as shown in the cumulative explained variance plot.

Feature Contributions

The heatmap of feature contributions shows that:

- PC1 is primarily influenced by density, alcohol, and total sulfur dioxide
- PC2 is strongly associated with volatile acidity, fixed acidity, and pH

This suggests that these chemical properties play a significant role in differentiating wine samples.

2D and 3D Visualization

The 2D PCA projection reveals:

- A clear separation between red and white wines
- Quality ratings spread across the projection, with some clustering
- Overlapping distributions suggesting complex relationships between features and quality

The 3D PCA projection adds some additional separation, but the gain in explained variance (14.1%) comes at the cost of more complex visualization.

t-SNE Analysis

The t-SNE visualization shows:

- Much more distinct clustering than PCA, particularly for wine types
- Local structures and neighborhoods are better preserved
- Quality ratings appear to have some correlation with the clusters, but the relationship is non-linear

Comparison between PCA and t-SNE

Interpretability and Clustering

- **PCA**: Provides a global view of the data with clear axes that can be interpreted in terms of the original features. The linear projections maintain global relationships but may not capture local structures well.
- **t-SNE**: Excels at revealing local structures and clusters, making it superior for visualization of complex data. However, the axes have no direct interpretation, and distances between well-separated clusters may not be meaningful.

How PCA and t-SNE Handle High-Dimensional Data Differently

1. **Linear vs. Non-linear**:

- PCA is a linear technique that finds orthogonal directions of maximum variance
- t-SNE is non-linear and can capture complex manifolds in the data

2. **Global vs. Local Structure**:
   - PCA preserves global structure and maximum variance
   - t-SNE preserves local neighborhoods and similarities between points
3. **Deterministic vs. Stochastic**:
   - PCA is deterministic, always producing the same result for a given dataset
   - t-SNE is stochastic, potentially producing different visualizations on each run
4. **Interpretability**:
   - PCA components have clear meaning in terms of original features
   - t-SNE dimensions have no direct interpretation
5. **Computational Complexity**:
   - PCA is computationally efficient ($O(n^2)$)
   - t-SNE is more computationally intensive ($O(n^2 \log n)$)

Key Observations from Visualizations

1. **Wine Type Separation**:
   - Both PCA and t-SNE clearly separate red and white wines
   - t-SNE shows more distinct boundaries between wine types
2. **Quality Distribution**:
   - Neither method shows perfect separation by quality
   - t-SNE reveals more local structure that might correlate with quality
3. **Feature Importance**:
   - PCA heatmap shows that different chemical properties contribute differently to the principal components
   - Alcohol, density, and sulfur dioxide appear particularly important
4. **Information Preservation**:

    - The first two PCs preserve only about half of the variance
    - This suggests complex relationships that can't be fully captured in low dimensions

Conclusion

Both PCA and t-SNE offer valuable but complementary insights into the Wine Quality dataset:

- PCA is better for understanding global structure and feature relationships
- t-SNE is superior for visualization and identifying clusters

For this dataset, a hybrid approach is recommended:

1. Use PCA to understand feature contributions and global structure
2. Use t-SNE for visualization and cluster identification
3. Consider using the first 3-5 PCs (capturing ~80-90% of variance) for subsequent machine learning tasks

The analysis demonstrates that while dimensionality reduction inevitably loses information, thoughtful application of these techniques can reveal important patterns while making the data more manageable for both visualization and analysis.