STAT 6650 Homework 4

María Victoria Liendro
Master's in Data Engineering
mzl0201@auburn.edu

Spring 2025

1) b) In Table 1 and 2 we have the top 3 principal components. For PC1, we have a positive characterization in Fixed Acidity and Citric Acid (0.657 and 0.516), in addition is negatively characterized by pH (-0.334). Meaning it is a dimension focused in wine acidity. For PC2, we have a strong characterization of Alcohol (0.758). Finally, for PC3 has a strong positive characterization of Free Sulfur Dioxide and Total Sulfur Dioxide (0.758 and 0.522).

1) c) Considering both Figure 1 (silhouette score) and Figure 2 (error score), the best option is R=5. We can see that the error is not as bad as 6, but the slope has decrease in comparison to 3 or 4. And the silhouette score is better than 6.

1) d) In Figure 3, 4 and 5 we have the ROC curve per class for each model. For the original model the highest score was for Class 0 with an AUC of 0.94, while lowest score was for Class 3 with an AUC of 0.69. The average AUC was 0.81. For the PCA model, the highest score was also for Class 0, 0.96, and the lowest also for Class 3, 0.70. The average score was 0.83. For the NMF model the highest score was for Class 5, 0.98, followed by Class 0, 0.96, while the lowest was again for Class 3 0.66. The average score was 0.84. It is interesting to see that Class 3, when the quality is medium, was the hardest to predict, but those extreme, like 0 or 5, were always the best predictions. Among these, the NMF model got the best performance.

2) a) First, we get f.

$$V = \frac{S * a^d}{d}$$

$$V_{inner} = \frac{S * (a-\epsilon)^d}{d}$$

$$V_{shell} = V_{total} - V_{inner} = \frac{S * a^d}{d} - \frac{S * (a-\epsilon)^d}{d}$$

$$V_{shell} = \frac{S}{d}(a^d - (a-\epsilon)^d)$$

$$f = \frac{V_{shell}}{V_{total}} = \frac{\frac{S}{d}(a^d - (a-\epsilon)^d)}{\frac{S*a^d}{d}}$$

$$f = \frac{V_{shell}}{V_{total}} = 1 - \left(1 - \frac{\epsilon}{a}\right)^d$$

Now, we show that for any ϵ, f tends to 1 as the dimensions tend to infinity. For the fraction of the volume f:

$$f = 1 - \left(1 - \frac{\epsilon}{a}\right)^d,$$

where $0 < \epsilon < a$.

As $d \to \infty$: Since $0 < \frac{\epsilon}{a} < 1$, the term $1 - \frac{\epsilon}{a}$ satisfies $0 < 1 - \frac{\epsilon}{a} < 1$. When raised to the power d (where $d \to \infty$):

$$\left(1 - \frac{\epsilon}{a}\right)^d \to 0.$$

Thus:

$$f \to 1 \quad \text{as} \quad d \to \infty.$$

Even for a small fixed ϵ, the fraction of volume $[a - \epsilon, a]$ dominates the entire volume in high dimensions. This illustrates the "curse of dimensionality," where most of the hypersphere's mass concentrates near its surface.

2) c) In Figure 6 we can see how as the dimensions increase, the f value approaches 1. This means that a small neighborhood, radius ϵ, with enough dimensions, it starts to take over all the space. In high-dimensional settings, even a small local region can cover most of the volume.

Table 1: Principal Component (Part 1)

	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides
PC-1	0.516	-0.215	0.657	0.057	0.071
PC-2	-0.062	-0.281	0.253	-0.049	-0.073
PC-3	-0.162	-0.179	0.205	0.136	0.017

Table 2: Principal Component (Part 2)

	Free Sulfur Dioxide	Total Sulfur Dioxide	Density	pH	Sulphates	Alcohol
PC-1	-0.075	-0.011	0.334	-0.334	0.129	-0.047
PC-2	-0.168	-0.204	-0.435	0.087	0.066	0.758
PC-3	0.758	0.522	-0.116	0.019	0.107	0.093

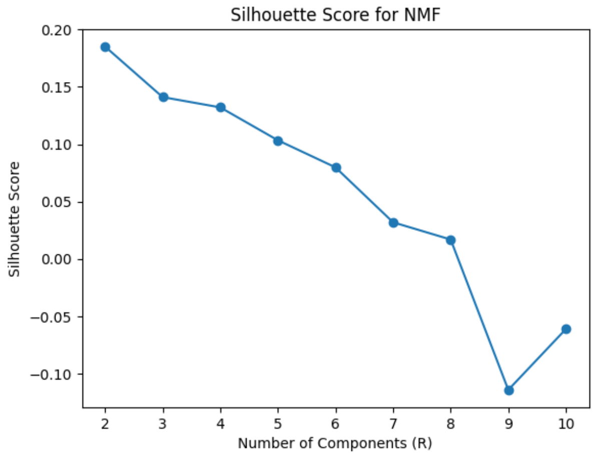

Figure 1:
Silhouette score.

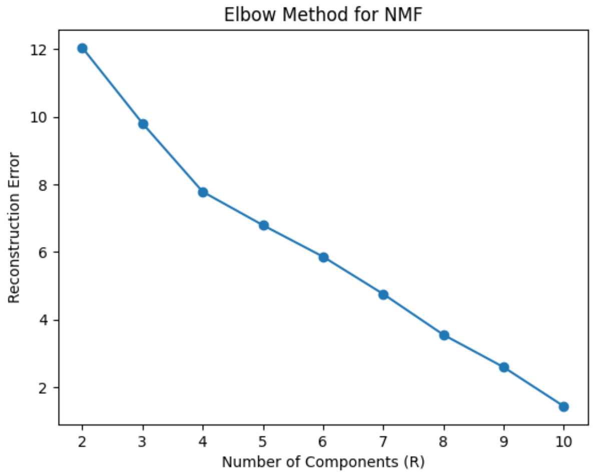

Figure 2:
Score for different R.

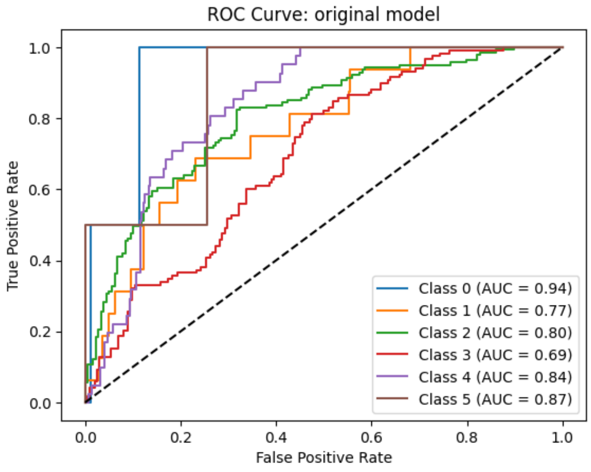

Figure 3:

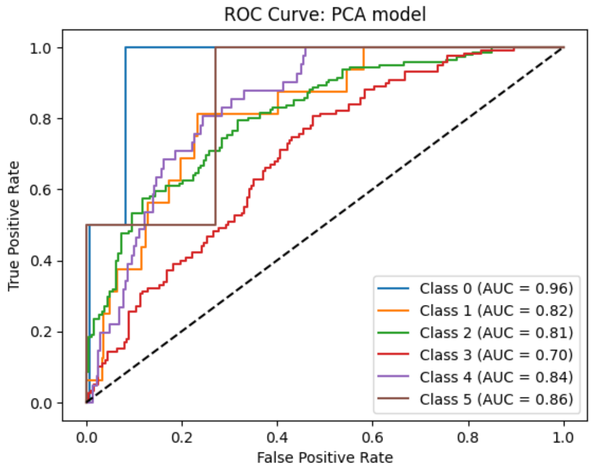

Figure 4:

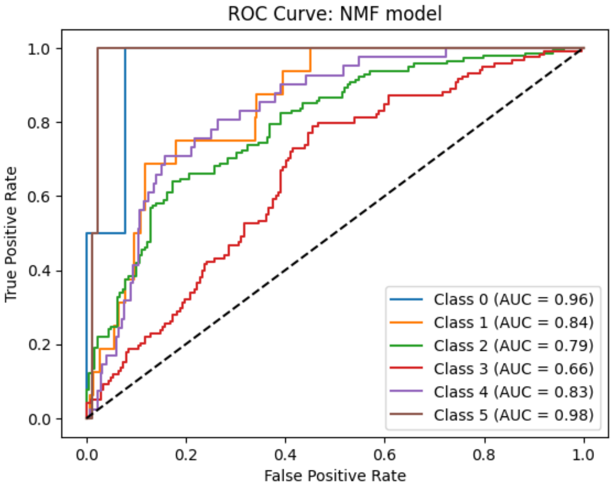

Figure 5:

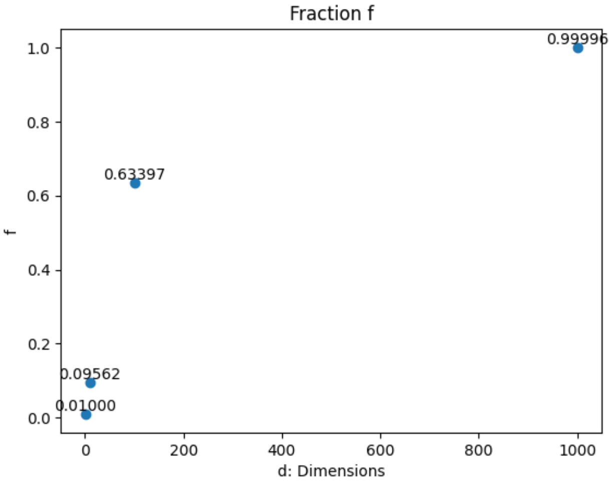

Figure 6: f values