Clustering of Patient Clinical Records

Jennifer Nguyen

COSC 3337 - Data Science I

Jingchao Ni

15 November 2024

Task 2: K-means

Results:

```
K-means Clustering Results (k=2):

Percentage of points in each cluster:
Cluster 0: 78.3%
Cluster 1: 21.7%

Overall purity score: 0.679

Purity score for each cluster:
Cluster 0: 0.692
Cluster 1: 0.631

Cluster Cluster 0 has the highest purity: 0.692
```

Analyzing Results:

The K-means clustering results for Task 2 reveal a notable imbalance in the distribution of data points, with Cluster 0 containing 78.3% of the data points (approximately 234 patients) and Cluster 1 containing 21.7% (approximately 65 patients). The overall purity score is 0.679 (67.9%), indicating that about 68% of all patients were correctly grouped based on their death event outcome. While this is better than random (50%), there is still significant room for improvement.

Individually, Cluster 0 has a purity score of 0.692 (69.2%), suggesting it likely represents patients with one particular outcome, probably survival, given its size. Cluster 1 has a lower purity score of 0.631 (63.1%), indicating it is less effective at separating patients based on their outcomes. Cluster 0 has the highest purity at 69.2%, with a relatively small difference in purity between the clusters (about 6 percentage points).

These results suggest that the K-means clustering with k=2 has moderate success in separating patients based on their outcomes. However, the algorithm tends to create uneven clusters, which might not be optimal for identifying two distinct outcomes. The relatively close purity scores between clusters indicate consistent but not exceptional performance across both groups.

Task 3: K-means

Results:

K-Means Experiments Results:

	k	Purity	Silhouette Coefficient
0	2	0.679	0.565
1	10	0.685	0.576
2	30	0.699	0.555
3	50	0.726	0.566
4	100	0.762	0.518

Best k for Purity: 100
Best k for Silhouette Coefficient: 10

Purity changes with increasing k:
k=10: +0.006
k=30: +0.014
k=50: +0.027
k=100: +0.036

Analyzing Results:

Trends in Purity: The purity score consistently increases with higher k values. Starting from a baseline of 0.679 at k=2, it rises to 0.685 at k=10, 0.699 at k=30, 0.726 at k=50, and reaches the highest purity of 0.762 at k=100. This trend is logical because more clusters result in smaller, more homogeneous groups. As k approaches the dataset size, each point could potentially be in its own cluster, leading to higher purity, although this does not necessarily imply better practical clustering.

Trends in Silhouette Coefficient: The Silhouette Coefficient shows a different pattern, peaking at k=10 with a value of 0.576. It then generally decreases as k increases beyond 10, with values of 0.555 at k=30, 0.556 at k=50, and 0.518 at k=100. This suggests that k=10 provides the best balance of cluster cohesion and separation, while larger k values create less well-defined clusters.

Best Values: The best k for purity is 100, while the best k for the Silhouette Coefficient is 10.

Why Purity Changes This Way: Purity increases with k because smaller clusters are more likely to contain points of the same class. As k increases, clusters become more granular, and when k approaches the number of data points, each point could potentially be in its own cluster, leading to perfect purity. However, this does not necessarily mean better clustering in practice because very high k values may lead to overfitting, making the clusters less meaningful for interpretation. The decrease in the Silhouette Coefficient indicates that the clusters become less well-defined.

Practical Implications: While k=100 gives the highest purity, k=10 may be more practical because it has the best Silhouette Coefficient, provides a more manageable number of clusters, and the clusters are likely more meaningful and interpretable.

Task 4: DBSCAN Experiments

Results:

DBSCAN Experiments Results:

	eps	Number of Clusters	Number of Anomalies	Purity
0	0.3	18	146	0.778
1	0.5	22	21	0.701
2	0.7	22	13	0.703

Best eps for Purity: 0.3 (Purity: 0.778)

Detailed Analysis:

eps = 0.3:
- Number of Clusters: 18.0
- Number of Anomalies: 146.0 (48.8% of data)
- Purity: 0.778

eps = 0.5:
- Number of Clusters: 22.0
- Number of Anomalies: 21.0 (7.0% of data)
- Purity: 0.701

eps = 0.7:
- Number of Clusters: 22.0
- Number of Anomalies: 13.0 (4.3% of data)
- Purity: 0.703

Analyzing Results:

The DBSCAN clustering results reveal several insights based on different values of epsilon (eps). For eps = 0.3, there are 18 clusters with 146 anomalies, accounting for 48.8% of the data points, and a purity of 0.778, the highest among the tested values. However, this high number of anomalies suggests that the eps value is too restrictive, creating more compact clusters but struggling with density variations.

For eps = 0.5, the number of clusters increases to 22, with a significant reduction in anomalies to 21 (7.0% of data points) and a purity of 0.701. This represents a more reasonable balance between cluster formation and anomaly detection, as more points meet the density requirements.

For eps = 0.7, the number of clusters remains at 22, with a further reduction in anomalies to 13 (4.3% of data points) and a slight increase in purity to 0.703. This value of eps results in the most inclusive neighborhood size of the three options.

Overall, while eps = 0.3 achieves the highest purity (0.778), it does so at the cost of marking almost half the data as anomalies. In contrast, eps = 0.5 and eps = 0.7 provide similar results with better coverage of the dataset. Between these two options, eps = 0.7 emerges as the optimal choice since it maintains the same number of clusters as eps = 0.5 while achieving both better purity (0.703 vs 0.701) and fewer anomalies (4.3% vs 7.0% of data points). The marginal purity improvement at eps = 0.3 does not justify excluding almost half the dataset as anomalies, making eps = 0.7 the most balanced and practical parameter choice for this dataset.

Conclusion

This study explored different clustering approaches to analyze patient clinical records using both K-means and DBSCAN algorithms. The analysis revealed several key insights about clustering behavior and effectiveness:

Basic K-means clustering (k=2) showed moderate success with an overall purity of 0.679, though it created imbalanced clusters (78.3% vs 21.7% distribution). This suggests that while the algorithm can distinguish between patient outcomes, it may be influenced by underlying data distribution patterns.

Extended K-means experiments (k=2 to k=100) demonstrated a trade-off between purity and cluster interpretability. While purity increased consistently with higher k values (reaching 0.762 at k=100), the Silhouette Coefficient peaked at k=10 (0.576),

indicating that moderate cluster numbers provide better-defined, more meaningful groupings.

DBSCAN clustering revealed the importance of parameter selection in density-based approaches. The optimal choice of eps=0.7 achieved a good balance between purity (0.703) and data coverage, with only 4.3% of points classified as anomalies. This demonstrated DBSCAN's ability to identify meaningful clusters while handling noise in the dataset.

Comparing both approaches, each offers distinct advantages:
- K-means provides more straightforward implementation and interpretation but requires pre-specifying the number of clusters
- DBSCAN offers more flexibility in cluster shape and automatic noise detection but requires careful parameter tuning

For this specific patient clinical records dataset, DBSCAN with eps=0.7 appears to be the more robust choice, as it achieves comparable purity to K-means while better handling outliers and requiring fewer assumptions about cluster structure. This suggests that density-based clustering might be more appropriate for medical data where patient groups may not form well-defined spherical clusters.