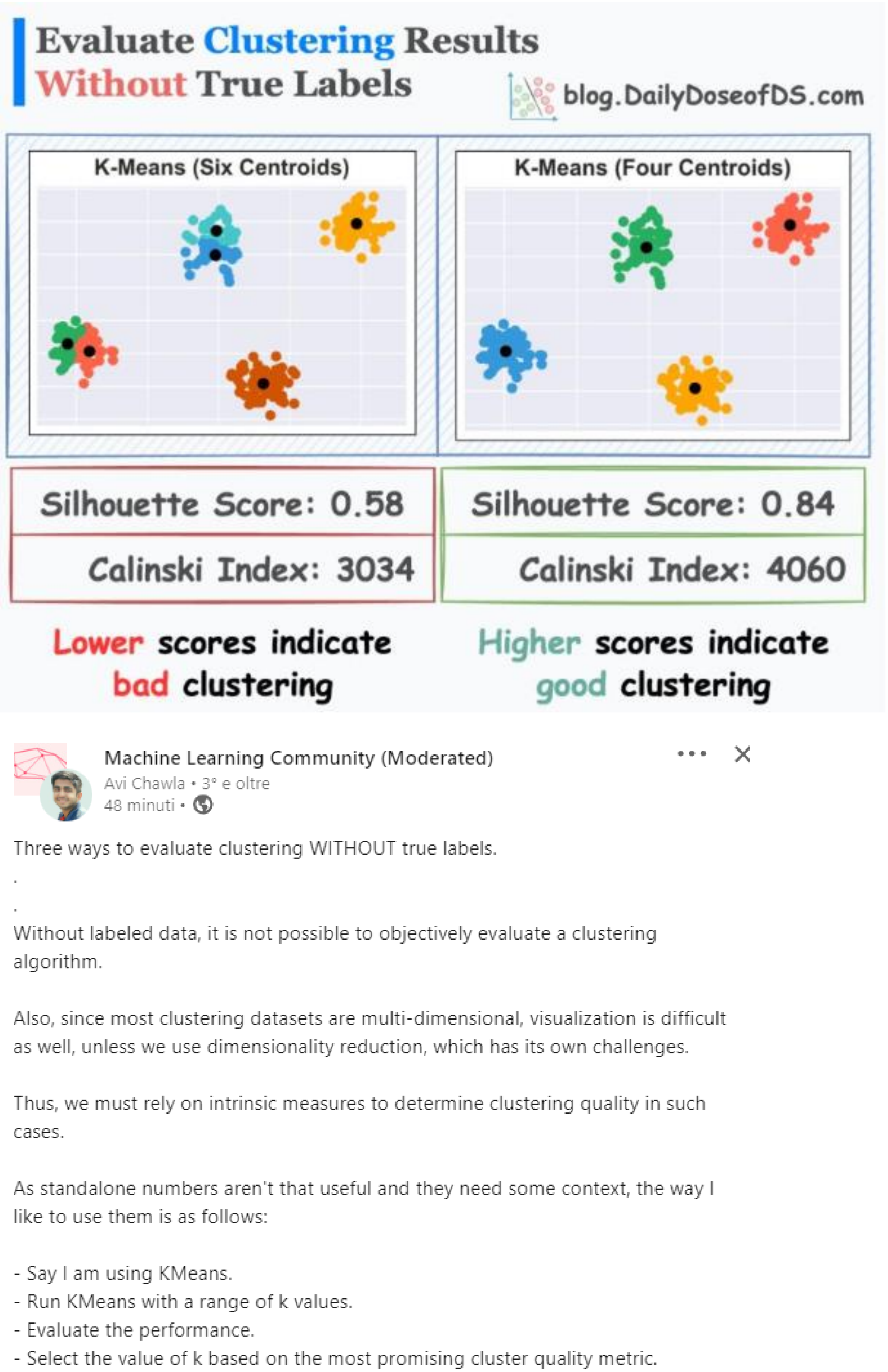

Evaluate Clustering Results Without True Labels

blog.DailyDoseofDS.com

Silhouette Score: 0.58	Silhouette Score: 0.84
Calinski Index: 3034	Calinski Index: 4060

Lower scores indicate bad clustering

Higher scores indicate good clustering

Machine Learning Community (Moderated)

Avi Chawla • 3° e oltre

48 minuti •

Three ways to evaluate clustering WITHOUT true labels.

.

.

Without labeled data, it is not possible to objectively evaluate a clustering algorithm.

Also, since most clustering datasets are multi-dimensional, visualization is difficult as well, unless we use dimensionality reduction, which has its own challenges.

Thus, we must rely on intrinsic measures to determine clustering quality in such cases.

As standalone numbers aren't that useful and they need some context, the way I like to use them is as follows:

- Say I am using KMeans.
- Run KMeans with a range of k values.
- Evaluate the performance.
- Select the value of k based on the most promising cluster quality metric.

Here are three metrics that can help.

1) **Silhouette Coefficient**:
The Silhouette Coefficient indicates how well each data point fits into its assigned cluster.

The idea is that if the distance to all data points in the identified cluster is small but that to the nearest cluster is large, this intuitively indicates that the clusters are well separated and the results can be reliable.

Measuring it across a range of centroids (k) can reveal which clustering results are most promising

2) **Calinski-Harabasz Index**:

The problem with Silhouette Coefficient is that it has a quadratic run-time relation with dataset's size.

Calinski-Harabasz Index is similar to the Silhouette Coefficient but it is computationally less expensive.

- A → sum of squared distance between all centroids and overall dataset center.
- B → sum of squared distance between all points and their specific centroid.
- metric is computed as A/B (with a scaling factor).

If A>>>B, this means that:
- The distance of centroids to the dataset center is large.
- The distance of data points to their specific centroid is small.
- Thus, it will result in a higher score, indicating that the clusters are well separated.

> The reason I prefer the Calinski Index over the Silhouette score is that:
> - It is relatively much faster to compute.
> - And it still makes the same intuitive sense for interpretation as the Silhouette Coefficient.
>
> 3) **DBCV**
> The issue with both the Silhouette Coefficient and the Calinski Index is that they are typically higher for convex (or somewhat spherical) clusters.
>
> So using them to evaluate arbitrary-shaped clusters, like those obtained with density-based clustering, can produce misleading results.
>
> DBCV is a better metric in such cases.
>
> Simply put, DBCV computes two values:
> - The density within a cluster
> - The density overlap between clusters
>
> A high density within a cluster and a low density overlap between clusters indicate good clustering results.