THINKFUL

# Clustering Algorithms

DATA SCIENCE

# Agenda

- Evaluating clustering performance
- Choosing the right algorithm

# Warm Up

Consider the following questions:

- Intuitively, what characteristics do well-formed clusters exhibit?
- What are some common considerations that come up when clustering? (Think back to the advantages and disadvantages sections of previous lectures.)

UP NEXT

# Evaluating Clustering Performance

# Overview

Evaluating the performance of clustering algorithms is very difficult!

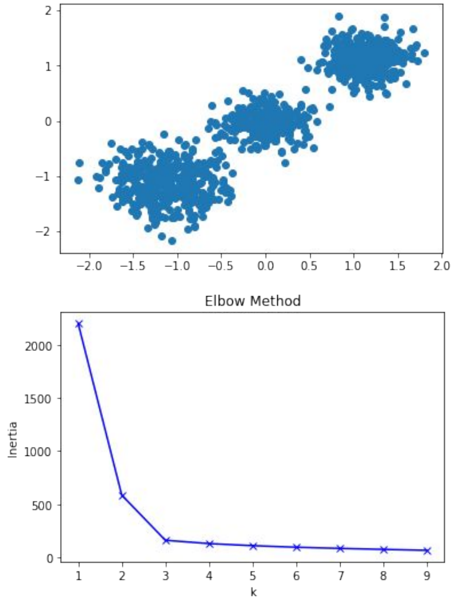
- No universal, unambiguous definition of what a “cluster” is
- Generally speaking, no labels to guide evaluation (*internal evaluation*)
- A few popular internal evaluation techniques include: **elbow plots, silhouette scores, and Davies-Bouldin scores**
- If labels are available (*external evaluation*), a wide range of evaluation criteria are available

# Elbow Method

A heuristic that examines how an evaluation metric of interest changes as a function of the number of clusters

Most common metric is within-cluster SSE, sometimes called *inertia*

- Goal is to find the “elbow”, at which point additional clusters do not explain very much additional variance in the data
- “Elbow” is not a precisely defined term, and is consequently very ambiguous and open to interpretation
- Only appropriate for centroid-based clustering algorithms

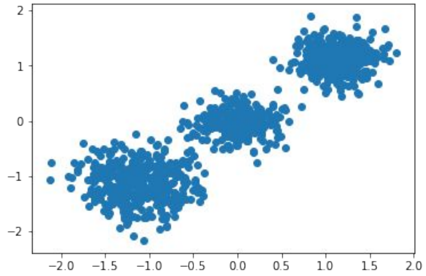


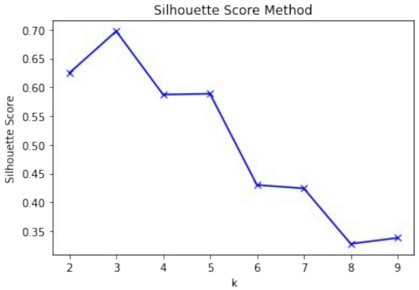

A synthetic dataset and the corresponding elbow plot for k-means.

# Silhouette Score

A heuristic that compares cluster cohesion and cluster separation

- More precisely, it compares the mean intra-cluster distance to the smallest mean extra-cluster distance
- Ranges from -1 (worst) to +1 (best)
- High silhouette values indicate that the points in each cluster are generally close to each other and far from the points in other clusters
- Can plot silhouette scores for individual samples; good clustering will show high values for most points within each cluster
- Assumes that “good” = “compact”, which is not always applicable or appropriate



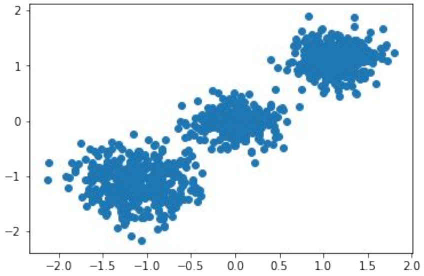

A synthetic dataset and the corresponding silhouette score plot for k-means.

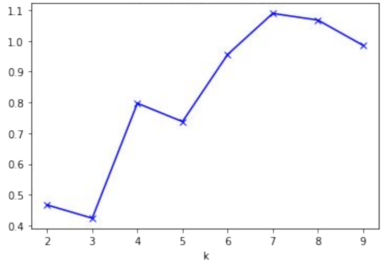
# Davies Bouldin Score

- An alternative to the silhouette score that also compares cluster cohesion and separation
- Unlike silhouette score, examines distances from points to centroids (instead of to each other) and distances between centroids (instead of to nearest cluster)
- Exact formula:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j} \frac{S_i + S_j}{D_{i,j}}$$

- Minimum score is 0 (best)
- Low scores indicate that the points in each cluster are generally close to their respective centroids and that cluster centroids are relatively far apart
- Like silhouette scores, assumes that ”good” = ”compact", which is not always applicable or appropriate




A synthetic dataset and the corresponding davies bouldin score plot for k-means.

# External Evaluation

- Although we usually do not have labeled data when clustering, there may be scenarios in which do have labels. Below are some common metrics for labeled data:
- Purity: the extent to which clusters contain a single class
- Rand Index: consider all pairs of samples as having a TP, FP, FN, or TN outcome (i.e. if a pair of observations have the same label and are assigned to the same cluster $\rightarrow$ TP; if they have different labels and are assigned to different cluster $\rightarrow$ TN, and so on). The Rand Index is simply the proportion of correct decisions:

$$\frac{TP + TN}{TP + FP + FN + TN}$$

# External Evaluation

- Many other similarity measures that should be familiar from classification and/or similarity contexts, such as F- score, Jaccard similarity, etc.
- Information-theoretic measures such as Mutual Information, which quantifies the mutual dependence or information between two random variables

$$\frac{TP + TN}{TP + FP + FN + TN}$$

UP NEXT

# Clustering Evaluation from a Business Perspective

# Clustering: A Business Perspective

- The objective of clustering is to group similar things together so that we can more easily make decisions about them.

- Clustering allows us to make decisions at the group level instead of having to make decisions about every single entity individually.

- The purer a cluster (i.e. the less overlap there is with other clusters), the easier it is to decide what to do with it.

# Clustering: A Business Perspective

- In business, it is often the case that we don’t have labels for our clusters to begin with.
- In such cases, it is useful to summarize the data after clustering has been performed and determine appropriate labels for each cluster.

| | Count | SeniorCitizen | Partner | Dependents | tenure | Gender_Female | Gender_Male |
|---|---|---|---|---|---|---|---|
| **Non-Senior Females** | 2915 | 0.000000 | 0.565695 | 0.336535 | 32.226415 | 1 | 0 |
| **Senior Females** | 568 | 1.000000 | 0.767606 | 0.073944 | 32.621479 | 1 | 0 |
| **Senior Males** | 574 | 1.000000 | 0.766551 | 0.085366 | 33.963415 | 0 | 1 |
| **High Tenure Males with Dependents** | 1027 | 0.000000 | 0.483934 | 1.000000 | 38.440117 | 0 | 1 |
| **Low Tenure Males with Partners & no Dependents** | 1948 | 0.000000 | 0.588296 | 0.000000 | 29.028747 | 0 | 1 |

UP NEXT

# Choosing A Clustering Algorithm

# Overview

There are several aspects of our problem we must consider when choosing the “best” clustering algorithm

- **Size of data:** many algorithms are asymptotically $O(n^2)$ and do not scale well with large datasets
- **Similarity measure:** is my data amenable to a proper distance metric (e.g. L-norms), or do I need to use looser similarity measures due to the presence of categorical data (or other considerations)?
- **Uncertainty:** do I care about how confident the clustering assignments are?
- **Prediction:** do we need to make predictions on new data?
- **Cluster shapes**: do my clusters have similar and simple shapes (e.g. all roughly spherical), or could their shapes be highly irregular and variable from cluster to cluster?
- **Number of clusters**: do I know this upfront or not?

# Size & Type Of Data

For increasingly large datasets:

Affinity Propagation, Mean Shift, Hierarchical Clustering, GMM, K-Medoids

DBSCAN

K-means

Less Suitable

More Suitable

For arbitrary similarity measures (often used with categorical or mixed data):

K-means, Mean Shift, GMM

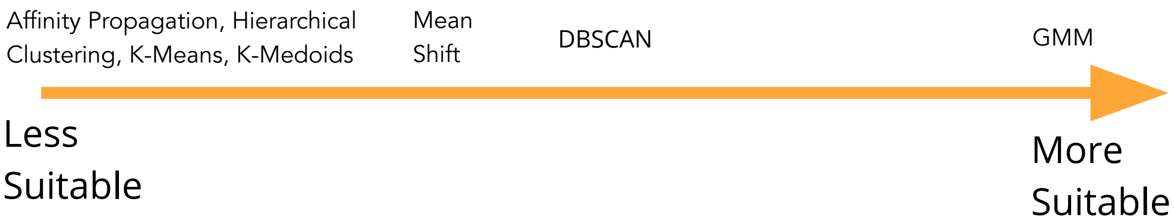
K-medoids, DBSCAN, Affinity Propagation, Hierarchical Clustering

Less Suitable

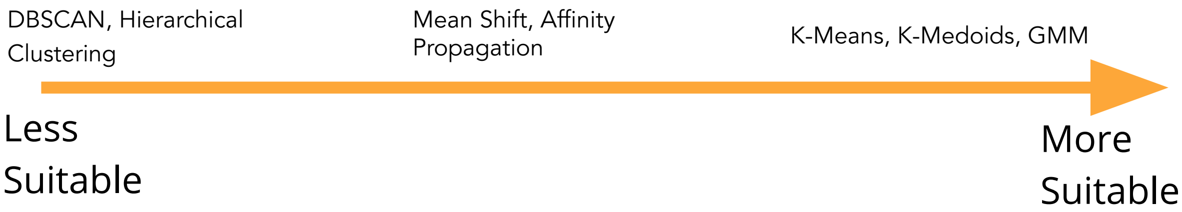
More Suitable

# Uncertainty & Predicting New Points

If you need to quantify confidence in cluster assignment:




If you need to predict cluster assignments of new points:



# Cluster Shapes

## If you expect your cluster shapes to be irregular and/or variable:

K-Means,
K-Medoids

GMM

Hierarchical Clustering, Mean Shift,
Affinity Propagation

DBSCAN

Less
Suitable

More
Suitable

## If you don’t know how many clusters are present upfront:

K-Means, K-Medoids, GMM

Mean-Shift, Affinity Propagation, DBSCAN,
Hierarchical Clustering

Less
Suitable

More
Suitable

# Summary

- Evaluating the performance of clustering algorithms without labeled data (internal evaluation) is very difficult, and the commonly used methods for doing so are fairly constrained
- If labels are provided, there are many evaluation metrics, some of which may look familiar from classification settings
- There are many factors to consider when choosing an algorithm, such as data size, feature types, cluster shapes, etc.

## EXERCISE

1. Jupyter notebook
2. Data set: Financial well-being survey results
3. Isolate the “score” subset of questions
4. Create an elbow plot using K-Means clustering and use it to determine an acceptable range of values for k
5. Create a silhouette plot using K-Means and identify the optimal value for k

THINKFUL

# Thank You