Mean Shift Clustering

Use mean shift clustering to determine the optimal number of clusters.

Chapter Goals:

- Learn about the mean shift clustering algorithm

A. Choosing the number of clusters

Each of the clustering algorithms we've used so far require us to pass in the number of clusters. This is fine if we already know the number of clusters we want, or have a good guess for the number of clusters. However, if we don't have a good idea of what the actual number of clusters for the dataset should be, there exist algorithms that can automatically choose the number of clusters for us.

One such algorithm is the mean shift clustering algorithm. Like the K-means clustering algorithm, the mean shift algorithm is based on finding cluster centroids. Since we don't provide the number of clusters, the algorithm will look for "blobs" in the data that can be potential candidates for clusters.

Using these "blobs", the algorithm finds a number of candidate centroids. It then removes the candidates that are basically duplicates of others to form the final set of centroids. The final set of centroids determines the number of clusters as well as the dataset cluster assignments (data observations are assigned to the nearest centroid).

In scikit-learn, the mean shift algorithm is implemented with the `MeanShift` object (part of the `cluster` module). Since the algorithm doesn't require us to pass in the number of clusters, we can initialize the `MeanShift` with no arguments.

The code below demonstrates how to use the `MeanShift` object.

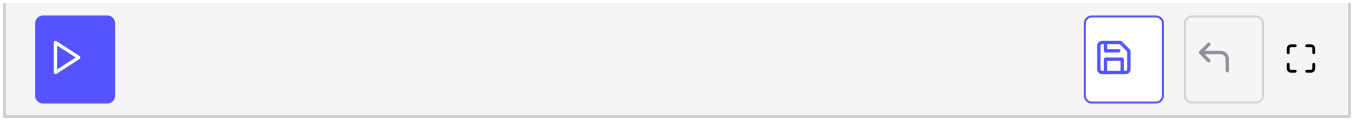
```
from sklearn.cluster import MeanShift
mean_shift = MeanShift()
```

```
mean_shift = MeanShift()
# predefined data
mean_shift.fit(data)

# cluster assignments
print('{}\n'.format(repr(mean_shift.labels_)))

# centroids
print('{}\n'.format(repr(mean_shift.cluster_centers_)))

new_obs = np.array([
  [5.1, 3.2, 1.7, 1.9],
  [6.9, 3.2, 5.3, 2.2]])
# predict clusters
print('{}\n'.format(repr(mean_shift.predict(new_obs))))
```



Since mean shift is a centroid-based algorithm, the `MeanShift` object contains the `cluster_centers_` attribute (the final centroids) and the `predict` function.