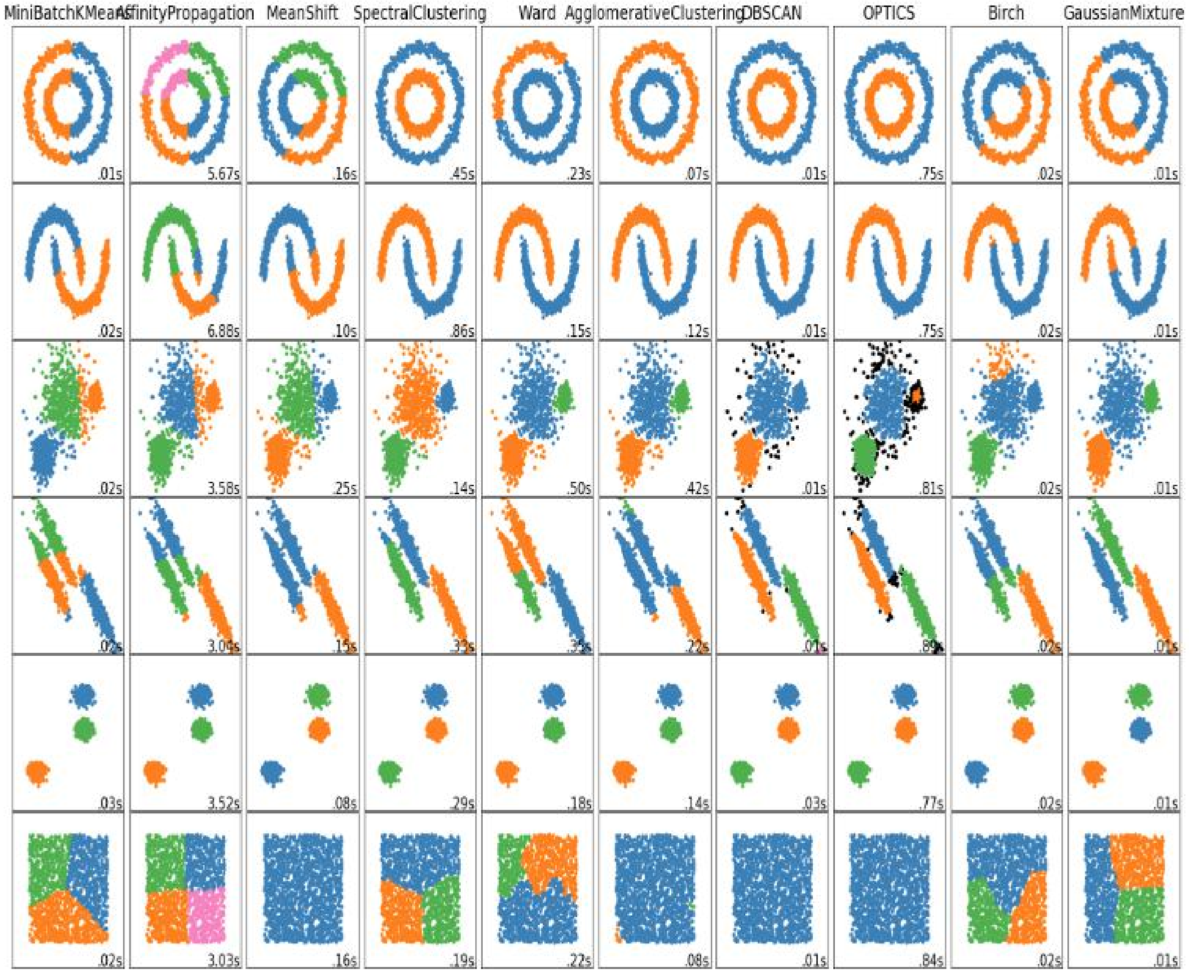
Homework2: Clustering with sklearn

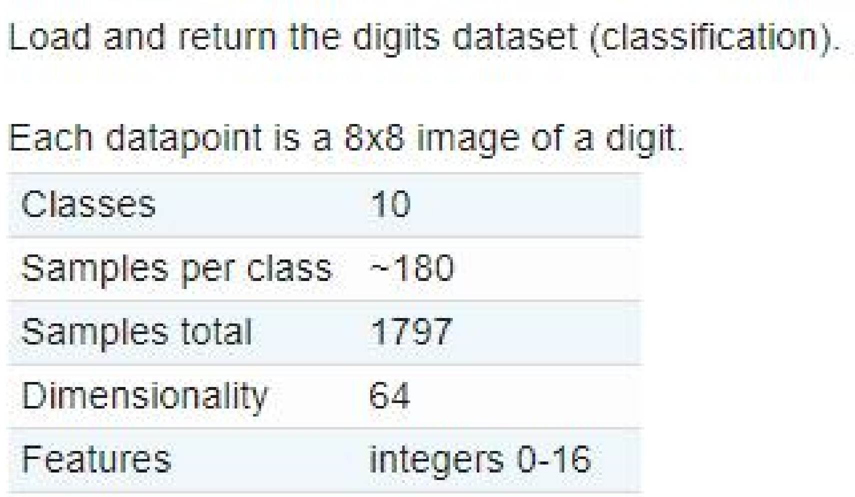


A comparison of the clustering algorithms in scikit-learn

https://scikit-learn.org/stable/modules/clustering.html#

Homework2: Clustering with sklearn

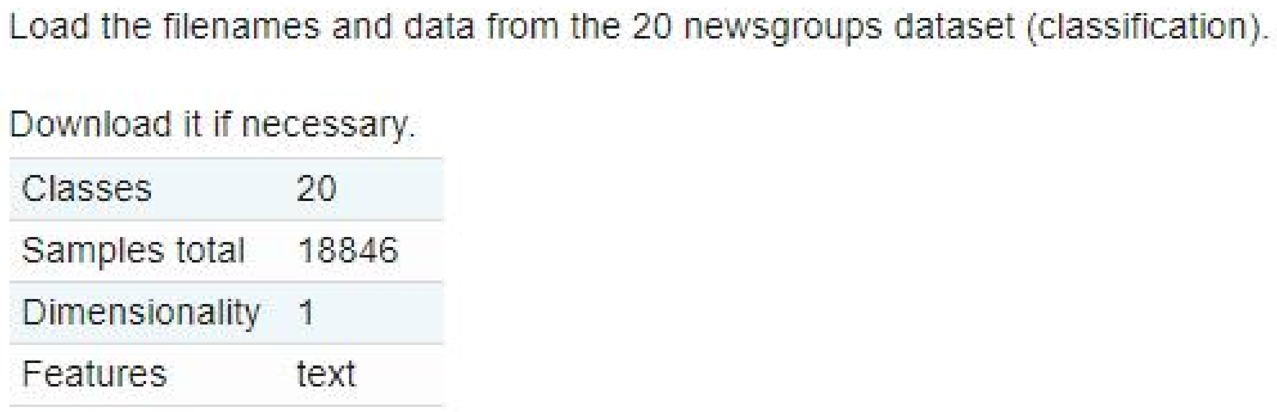
- Datasets
  - sklearn.datasets.load_digits
  - sklearn.datasets.fetch_20newsgroups



Load and return the digits dataset (classification).

Each datapoint is a 8x8 image of a digit.

Classes	10
Samples per class	~180
Samples total	1797
Dimensionality	64
Features	integers 0-16



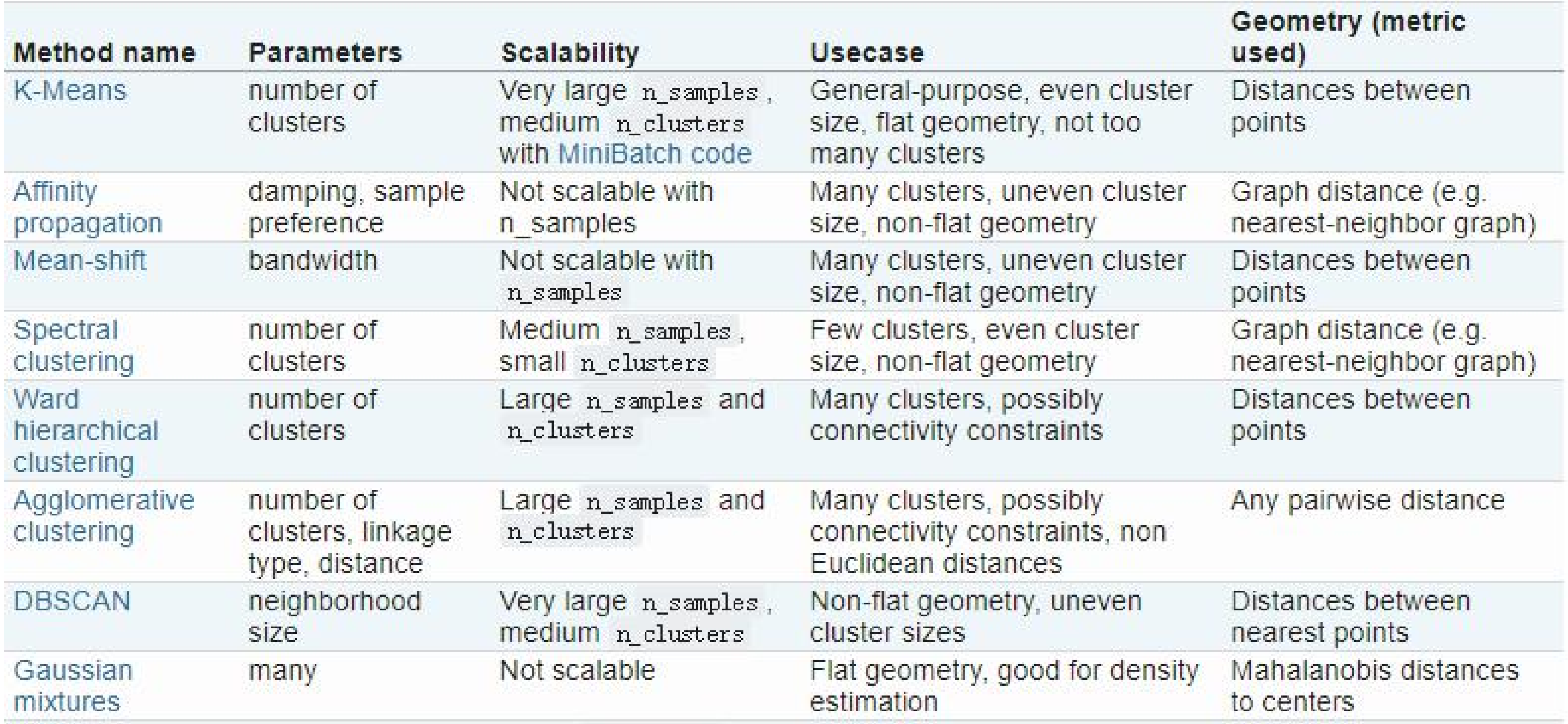
Load the filenames and data from the 20 newsgroups dataset (classification).

Download it if necessary.

Classes	20
Samples total	18846
Dimensionality	1
Features	text

Homework2: Clustering with sklearn

- 测试sklearn中以下聚类算法在以上两个数据集上的聚类效果。



Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large n_samples, medium n_clusters with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium n_samples, small n_clusters	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large n_samples and n_clusters	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large n_samples and n_clusters	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large n_samples, medium n_clusters	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers

https://scikit-learn.org/stable/modules/clustering.html#

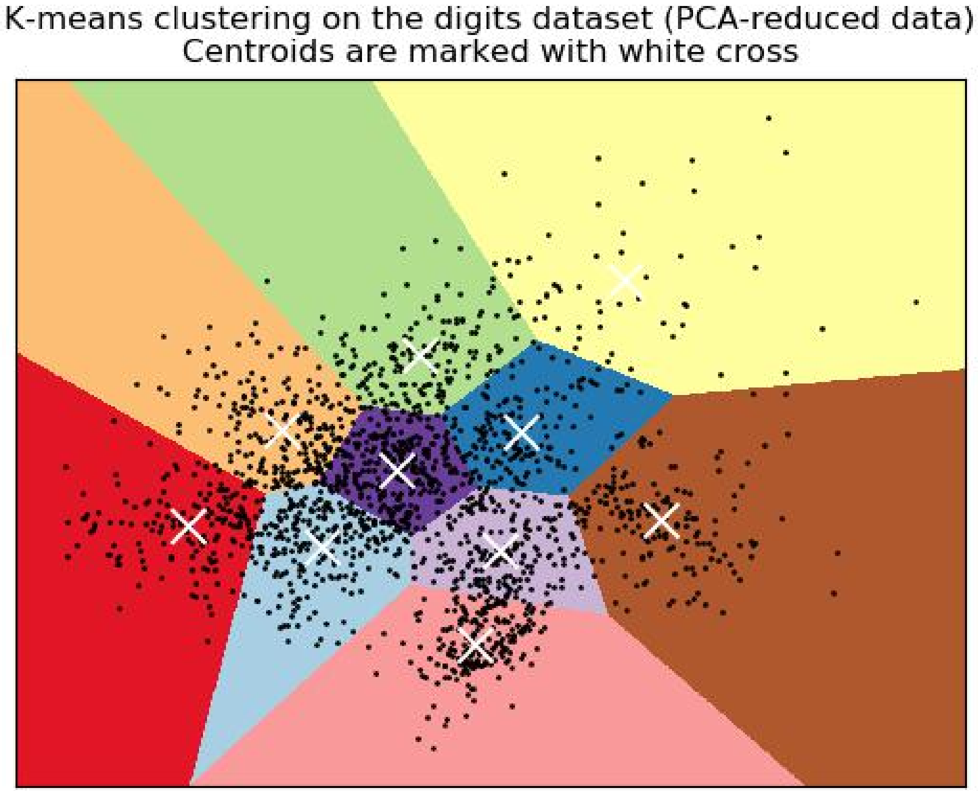
Homework2: Clustering with sklearn

- Evaluation
  - labels_true and labels_pred
    - >>> from sklearn import metrics
    - >>> labels_true = [0, 0, 0, 1, 1, 1]
    - >>> labels_pred = [0, 0, 1, 1, 2, 2]
  - Normalized Mutual Information (NMI)
    - >>> metrics.normalized_mutual_info_score(labels_true, labels_pred)
  - Homogeneity: each cluster contains only members of a single class
    - >>> metrics.homogeneity_score(labels_true, labels_pred)
  - Completeness: all members of a given class are assigned to the same cluster
    - >>> metrics.completeness_score(labels_true, labels_pred)

https://scikit-learn.org/stable/modules/clustering.html#

Homework2: Clustering with sklearn

- Examples
  - A demo of K-Means clustering on the handwritten digits data
    - https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py
  - Clustering text documents using k-means
    - https://scikit-learn.org/stable/auto_examples/text/plot_document_clustering.html#sphx-glr-auto-examples-text-plot-document-clustering-py


K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross