**Handling Sparse Datasets**: you can do z transformations or PCA, otherwise-

* **Lloyd's algorithm**: (also known as Voronoi iteration or relaxation) is an algorithm for finding evenly spaced sets of points in subsets of Euclidean spaces and partitions of these subsets into well-shaped and uniformly sized convex cells.
  + It repeatedly finds the centroid of each set in the partition and then re-partitions the input according to which of these centroids is closest
  + The mean operation is an integral over a region of space, and the nearest centroid operation results in Voronoi diagrams
  + ***Voronoi diagram*** is a *partition of a plane* into regions close to each of a given set of objects. In the simplest case, these objects are just finitely many points in the plane (called seeds, sites, or generators). For each seed there is a *corresponding region* consisting of all points of the plane closer to that seed than to any other. These regions are called *Voronoi cells*.
    - ***Natural element method*** *(NEM) is a meshless method to solve partial differential equation, where the elements do not have a predefined shape as in the finite element method, but depend on the geometry*
* ***k*-means clustering**: a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.
  + minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances
  + ***k*-medians** clustering is a cluster analysis algorithm. It is a variation of k-means clustering where instead of calculating the mean for each cluster to determine its centroid, one instead calculates the median. This has the effect of minimizing error over all clusters with respect to the 1-norm distance metric, as opposed to the squared 2-norm distance metric
  + ***k-medoids*** or ***partitioning around medoids (PAM) algorithm*** is a clustering algorithm reminiscent of the k-means algorithm. Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster
    - *k*-medoids chooses data points as centers (medoids or exemplars) and can be used with arbitrary distances, while in k-means the center of a cluster is not necessarily one of the input data points.
    - k-medoids chooses data points as centers (medoids or exemplars) and can be used with arbitrary distances, while in k-means the center of a cluster is not necessarily one of the input data points.
* **Dynamic time warping:** dynamic time warping (DTW) is one of the algorithms for measuring similarity between two temporal sequences, which may vary in speed.
  + DTW is a method that calculates an optimal match between two given sequences (e.g. time series) with certain restriction and rules
  + The optimal match is denoted by the match that satisfies all the restrictions and the rules and that has the minimal cost, where the cost is computed as the sum of absolute differences, for each matched pair of indices, between their values
  + The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in time series classification. Although DTW measures a distance-like quantity between two given sequences, it doesn't guarantee the triangle inequality to hold.
* ***k*-nearest neighbors**: a non-parametric method used for classification and regression
  + In *k*-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
  + In *k*-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors
  + Neighbors are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known.
  + Condensed nearest neighbor (CNN, the Hart algorithm) is an algorithm designed to reduce the data set for *k*-NN classification
* **Large margin nearest neighbor (LMNN)** classification is a statistical machine learning algorithm for metric learning. It learns a pseudometric designed for k-nearest neighbor classification. The algorithm is based on semidefinite programming, a sub-class of convex optimization.
* **Spectral clustering** techniques make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions
  + the similarity matrix may be defined as a symmetric matrix A, where *Aij*>0 represents a measure of the similarity between data points with indices *i* and *j*.
  + use a standard clustering method (there are many such methods, k-means is discussed below) on relevant eigenvectors of a Laplacian matrix of A
  + ***Laplacian matrix***, sometimes called admittance matrix, Kirchhoff matrix or discrete Laplacian, is a *matrix representation of a graph*. Together with Kirchhoff's theorem, it can be used to calculate the number of spanning trees for a given graph.
* **Neighborhood components analysis**: a supervised learning method for classifying multivariate data into distinct classes according to a given distance metric over the data
  + It aims at "learning" a distance metric by finding a linear transformation of input data such that the average leave-one-out (LOO) classification performance is maximized in the transformed space
  + *Leave-one-out (LOO) classification: predicting the class label of a single data point by consensus of its k-nearest neighbors with a given distance metric*