In data processing, if there are many feature dimensions, it usually affects the effect of machine learning models. Because redundant features will bring some noise and affect the calculation results; irrelevant features will increase the amount of calculation and consume time and resources. So we usually re-transform the data and run the model again. The purpose of data transformation is not only to reduce dimensionality, but also to remove the correlation between features and discover some potential feature variables.

PCA is a way to reduce the dimensionality of data with the least loss of information. The result is to project the feature space of the original data (n d-dimensional samples) into a smaller subspace, and express the information as well as possible.

Kmeans and DBSCAN are two common clustering algorithms in unsupervised machine learning.

# Kmeans Algorithm:

Kmeans algorithm devides the given dataset into K clusters according to the distance between data points. Let the points in the cluster be as close as possible, and the distance between the clusters should be as large as possible.

It is easy to understand, algorithm complexity is low, but The K cluster value needs to be set manually, and the results is sensitive to the initial cluster center, different selection methods will get different results; sensitive to outliers. Common methods for selecting K values are elbow method or Silhouette scores.

DBSCAN algorithm is a density-based clustering algorithm, which generally assumes that the class can be determined by the tightness of the sample distribution. By grouping closely connected samples into a class, a clustered class is obtained. By dividing all groups of closely connected samples into different categories, we get the final result of all clustering categories.

It can find outliers while clustering; can be used to any shape of dataset, but the clustering convergence time is longer if the sample set is large; parameter adjustment is more complicated and different parameter combinations have great impact on the final clustering effect.