COMPSCI5100 Machine Learning & Artificial Intelligence for Data Scientists(M) : Case Study 5 - model selection

Jiahuan Mai - 2549379M

Vaishnavi Balaji - 2549408B

Xinli Wang - 2519768W

School of Computing Science

Sir Alwyn Williams Building

University of Glasgow

G12 8RZ

Friday 2 April 2021

Table of Contents

Introduction: **3**
Methods: **3**
Preliminary work: 3
Gaussian covariance matrices: 3
AIC: 4
BIC: 5
Silhouette: 6
Cross-validation: 7
Results: **8**
Discussion: **10**
Bibliography **11**

Introduction:

For the sample model of distinguishing numbers, different clustering models are used and the most suitable dividing point (number of clusters) and the suitable covariance matrix structure are found. Choose a suitable Gaussian model and use four methods to explore whether the clustering can be successfully divided or which ones are classified into one category. According to sample analysis, it may be difficult to successfully use clustering to make detailed divisions, or the accuracy may be relatively low. Explore the impact of matrix selection or model changes on sample analysis.

Methods:

Preliminary work:

The sklearn learning library covers basic data construction for exercises to test different model methods or other algorithm models. In this K-Means model, the most basic clustering task can be realized. Find the appropriate n_clusters for the data set. The curve of the Elbow Method (EM) further shows and confirms that the inflection point information is the appropriate number of divisions. In addition, the Silhouette Score and BIC methods will be introduced in the remaining part.

Gaussian covariance matrices:

The Gaussian mixture model (GMM) refers to the linear combination of multiple Gaussian distribution functions. In theory, GMM can fit any type of distribution. It is usually used to solve the situation where the data in the same set contains multiple different distributions (Either the same type of distribution but different parameters, or different types of distributions, such as normal distribution and Bernoulli distribution).

When GMM is used for clustering, assuming that the data obeys the Mixture Gaussian Distribution, then it is enough to derive the probability distribution of GMM according to the data; then the K components of GMM actually correspond to K clusters (the task requires Find the number represented by this K). Here, the hyperparameter convariance_type in the GMM model controls the degree of freedom of the shape of each cluster.

- Its default setting is convariance_type='diag', which means that the size of the cluster in each dimension can be set separately, but the main axis of the ellipse boundary must be parallel to the coordinate axis.
- When covariance_type='spherical', the model makes all dimensions equal by constraining the shape of the cluster. The clustering results obtained in this way are similar to the characteristics of k-means clustering, although the two are not exactly the same.

- When covariance_type=’full’, the model allows each cluster to be modeled as an ellipse in any direction.
- When covariance_type=’tied’, all Gaussians share the same covariance matrix.

Here four methods are used here to test, which is very bad here, because the default K is selected as 10 (0-9) which is inappropriate for this moment, and the accuracy rate is relatively low.

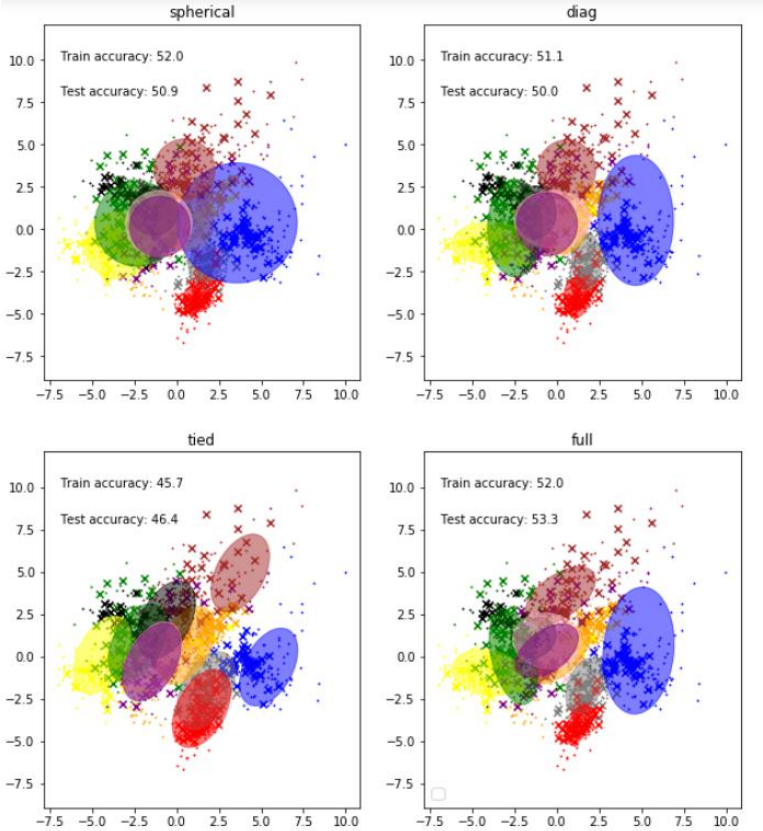

Fig 2.1 Four covariance type with K=10

AIC:

Akaike Information Criterion (AIC) is a standard for evaluating the complexity of statistical models and measuring the goodness of fit of statistical models. In general, AIC can be expressed as: AIC=2k−2ln(L). Increasing the number of free parameters improves the goodness of fitting. AIC encourages the goodness of data fitting but try to avoid overfitting. Therefore, the preferred model should be the one with the smallest AIC value. The method of the Akaike information criterion is to find a model that best interprets the data but contains the fewest free parameters.

Here, the test K is between 2-7 (gmm.aic(data)), and the standard deviation is used to calculate the error.

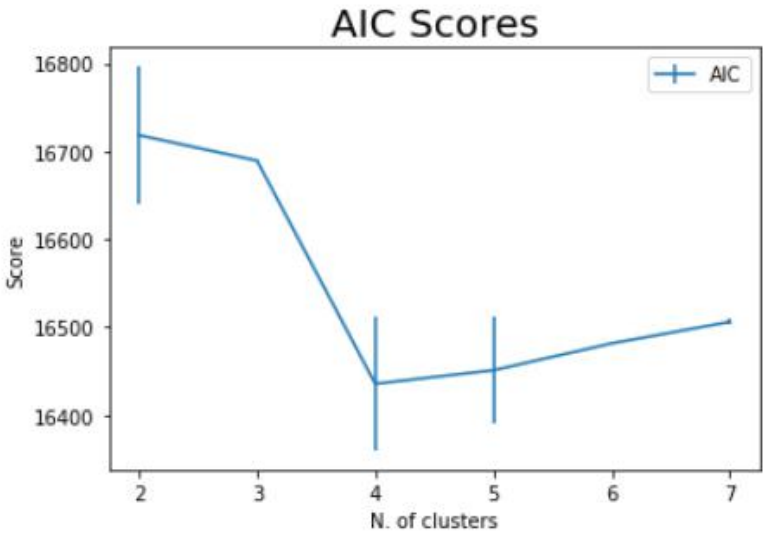

Fig 2.2 AIC Scores

Calculate the gradient value of AIC.

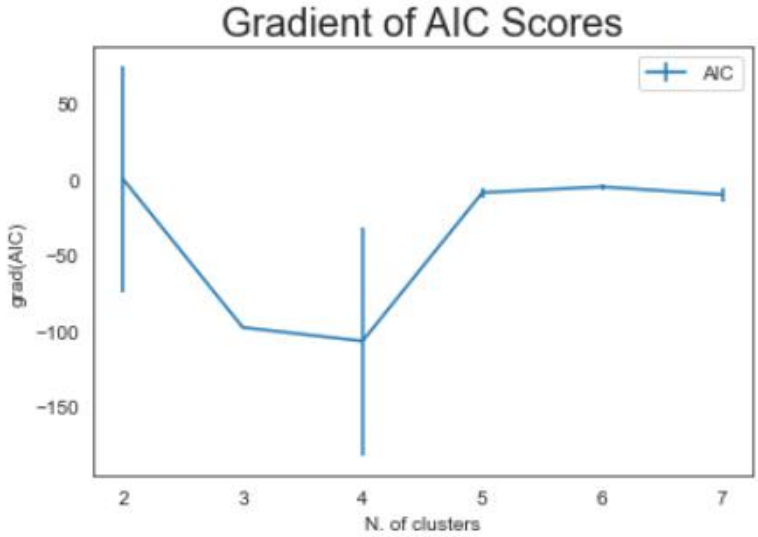

Fig 2.3 Gradient of AIC Scores

BIC:

Bayesian decision theory is an important part of subjective Bayesian induction theory. Under incomplete intelligence, use subjective probabilities to estimate part of the unknown state, then use Bayesian formula to modify the probability of occurrence, and finally use the expected value and modified probability to make the optimal decision. The formula is: BIC=ln(n)k-2ln(L). Similar to AIC, when training a model, increasing the number of parameters, that is, increasing the complexity of the model, will increase the likelihood function, but it will also lead to overfitting. For this problem, both AIC and BIC introduce the number of parameters Related penalty items, the penalty item of BIC is larger than that of AIC, taking into account the

number of samples, when the number of samples is too large, it can effectively prevent the model from being too complicated due to high model accuracy.
With method gmm.bic(data), the implementation process is similar to AIC.

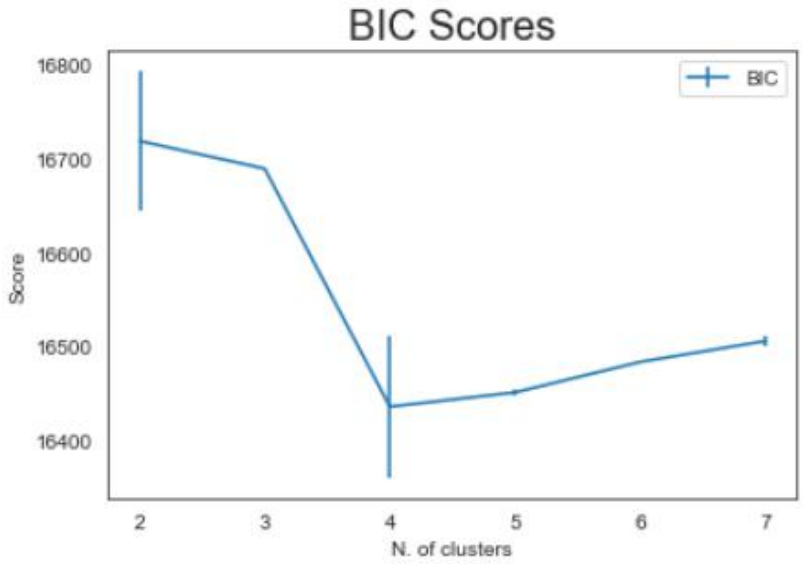

Fig 2.4 Gradient of BIC Scores

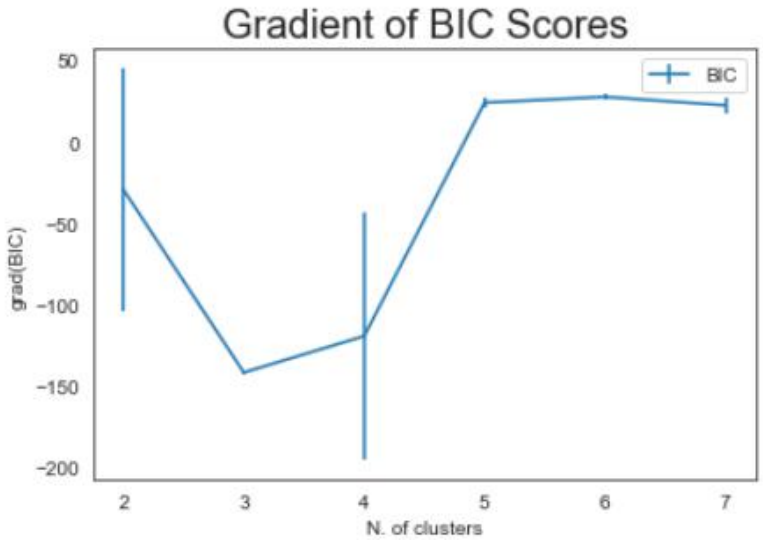

Fig 2.5 Gradient of BIC Scores

Silhouette:

Silhouette Coefficient, which combines two factors of cohesion and separation. It can be used to evaluate the impact of different algorithms or different operating modes of algorithms on the clustering results on the basis of the same original data.

1. Calculate the average distance ai from sample i to other samples in the same cluster. The smaller the ai, the more the sample i should be clustered into this cluster. Call ai the degree of dissimilarity within the cluster of sample i.

The mean value of a i of all samples in cluster C is called the cluster dissimilarity of cluster C.

2. Calculate the average distance bij from sample i to all samples of some other cluster Cj, which is called the dissimilarity between sample i and cluster Cj. Defined as the dissimilarity between clusters of sample i: bi =min{bi1, bi2, ..., bik}
   The larger bi is, the less the sample i belongs to other clusters.
3. According to the in-cluster dissimilarity a i and the inter-cluster dissimilarity b i of the sample i, define the contour coefficient of the sample i:

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \qquad s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, & a(i)<b(i)\\ 0, & a(i)=b(i)\\ \dfrac{b(i)}{a(i)}-1, & a(i)>b(i)\end{cases}$$

Here we use method : metrics.silhouette_score(data, labels, metric='euclidean') .

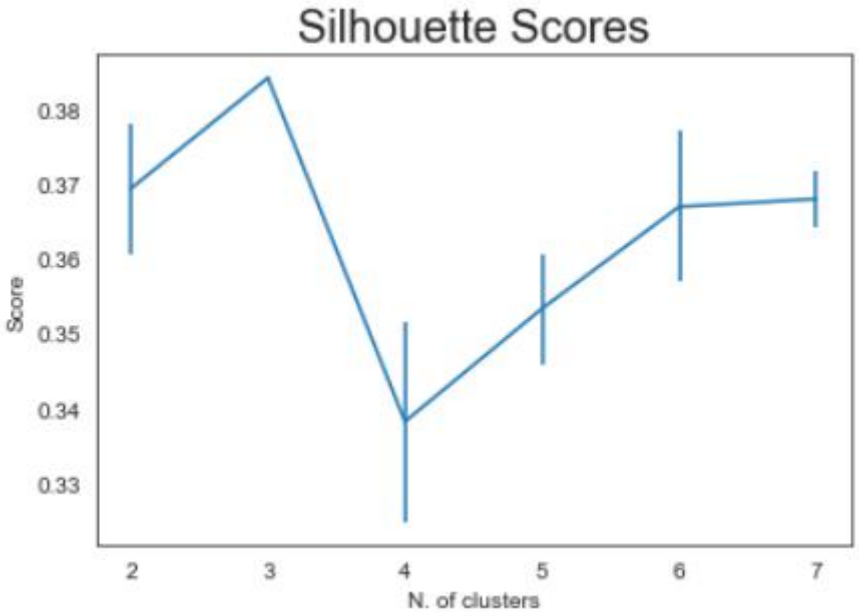

Fig 2.6 Silhouette Scores

Cross-validation:

Cross validation is often combined with web search as a method of parameter evaluation. This method is called grid search with cross validation. Sklearn therefore designed such a class GridSearchCV, which implements fit, predict, score and other methods. As an estimator, using the fit method, in the process:

(1) Find the best parameter

(2) An estimator with the best parameters is instantiated

Here, we have implemented the GridSearchCV method from sklearn.model_selection

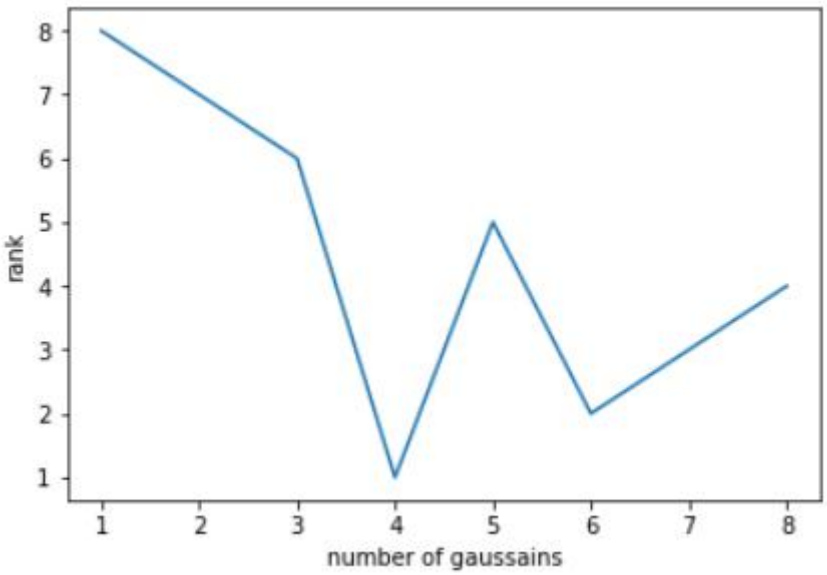

Fig 2.6 Silhouette Scores

Results:

The above method shows that the selection of cluster K performs better at 4. If K is equal to 4, perform GMM model training to find a suitable covariance matrix structure.

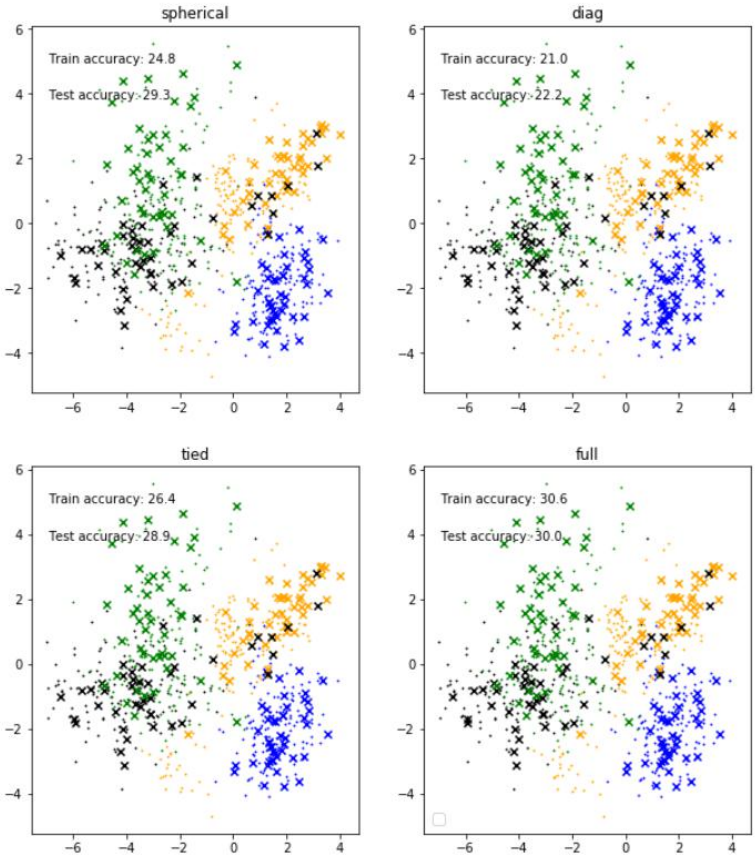

Fig 3.1 Covariance matrix structure

When we adjust the number of samples for training and prediction of the data set, “full”can show relatively high accuracy. But if the seed random function is set, it may change.

The clustering using the Kmeans method is implemented as follows

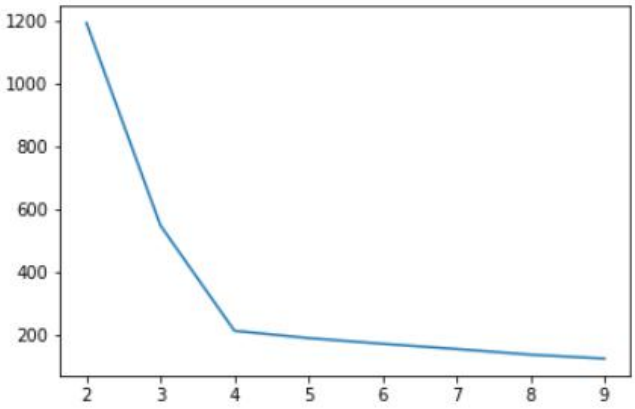

Fig 3.2 Elbow Method for K-Means

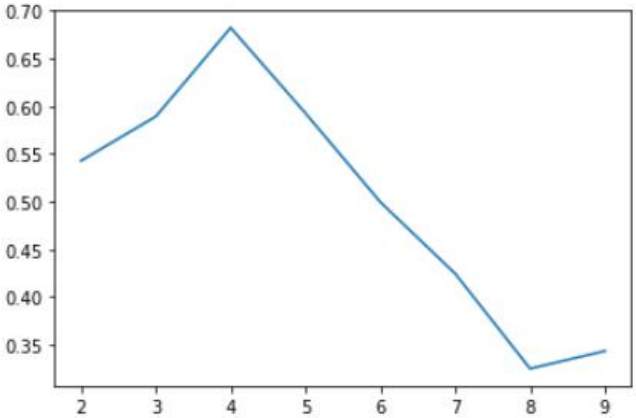

Fig 3.4 Silhouette Score for K-Means

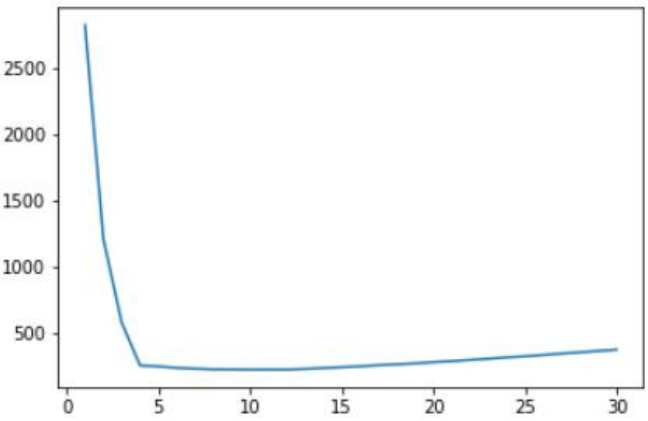

Fig 3.5 BIC for K-means

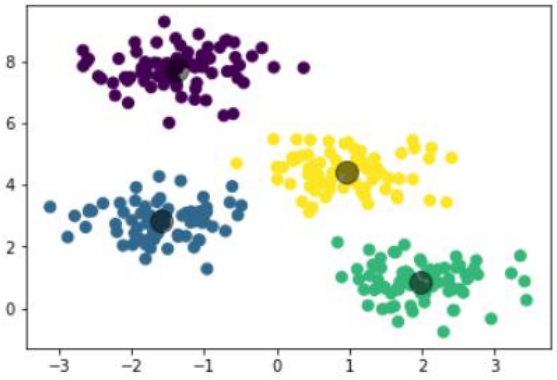

Fig 3.6 Clustering for K-Means

In the case, the kmeans method can easily achieve the clustering effect for the data set with obvious division.

Discussion:

In the ten-digit PCA data graph here, "0,4,6" and "2,3,5,8,9" are the easiest to be classified into one category.

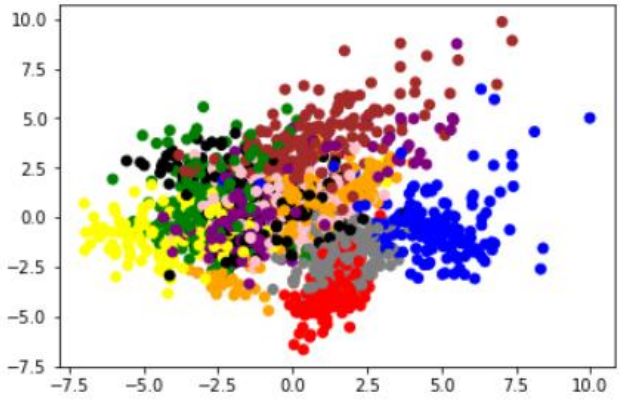

Fig 4.1 Color of numbers

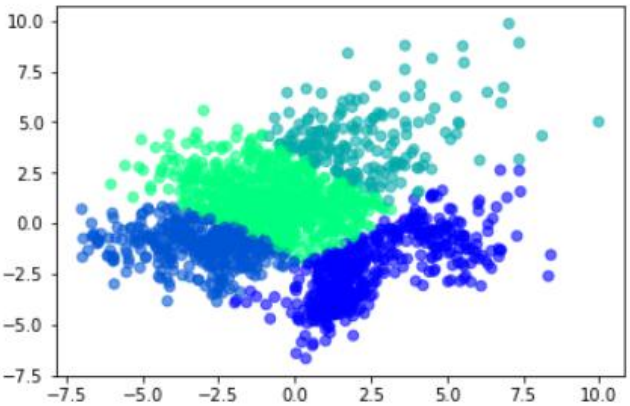

Fig 4.2 Clustering of numbers

For data sets that have many data points with the same characteristics and are interleaved with each other, clustering is still difficult to accurately distinguish each cluster. For data with obvious characteristic differences (similar to having boundaries)

clustering can still show better effect. As a generative model, GMM provides a method to determine the optimal number of components in a data set. Since the generative model itself is the probability distribution of the data set, the model can be used to evaluate the likelihood of the data, and cross- validation can be used to prevent overfitting. Scikit-Learn's GMM evaluator has built-in two standard analysis methods to correct over-fitting: Akaike Information Criteria (AIC) and Bayes Information Criteria (BIC).

Discussion:

1. See, e.g., Ketchen, Jr, David J.; Shook, Christopher L. (1996). "The application of cluster analysis in Strategic Management Research: An analysis and critique". Strategic Management Journal. 17 (6): 441–458. doi:10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G
2. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
3.