**Case 5: Model selection**

**(Model-based approaches)**

**Lin Chengli 2529061L**

**Tang Mingkai 2550578T**

# Introduction

## Introduction to case study

The k-means algorithm solves the clustering problems, and the outstanding advantage of this algorithm is that it is simple and easy to use, and the amount of calculation is not much. However, often being too simple is also a disadvantage. Because the method based on the distance model as the clustering standard may not always be successfully applied. To solve these shortcomings, a method of clustering with a statistical mixture model will play a better role-Gaussian Mixture Model (GMM). This clustering method obtains the probability that each sample point belongs to each class, rather than judging that it belongs to a class completely, so it is sometimes called soft clustering. (jteng, 2014)

Many parameter estimation problems use the likelihood function as the objective function. When the training data is sufficient, the accuracy of the model can be continuously improved, but at the cost of increasing the complexity of the model, it also brings a very common problem in machine learning—overfitting. Therefore, the model selection problem seeks the best balance between model complexity and the model's ability to describe the data set (likelihood function). (Lu,2014) Clear the statement of predictive/exploratory question.

There are three questions need to be proved and explored. Firstly, which is the best Feature Selection among the four methods in lesson. The below part demonstrates three of them: wrapper, filtering, and Embedded, comparing with their accuracy. Next, finding the most useful features and remove useless ones, then return a matrix with features and participants. Third question is whether the result is good or not, from the aspect of accuracy, specificity and sensitivity, which can be judged from the numerical value, because they represent the performance of the model.

## 1.2 Clear the statement of predictive/exploratory question.

There are four methods (BIC, AIC, Silhouette, and cross-validation) to check the best option of number of clusters in the same dataset. The first question that needs to be explored is the value of four methods, which represent the performance of the different number of clusters in one method and appear the most efficient number of clusters in diverse methods.

# Methods

The raw data is shape by 1797, 64 and 64 is an 8x8 picture which store a digit 0-9.

Chart, scatter chart

Description automatically generatedThe main idea is that we apply KMeans to find out the most suitable clustering number this digit is. The optimal number should be 10 while the digit is from 0 to 9. For each result, we plot a scatter to show how the cluster being placed. The scatter is plot by the data go through the PCA function and reduce the dimension to the 2-D plot like bellow’s figure:

First part, we do the silhouette score to calculate the highest number.

For silhouette score, python has a package called silhouette\_score from sklearn. We iterate cluster number from 2 to 15 to find out the best score.

We set the random state to 30 and fit the kmeans. Then, predict the digits.data, which is the raw data, and compare the prediction and the true number to calculate the score. In the end, plot the scatter of the best cluster.

Second part, we apply the cross validation with the package from sklearn called cross\_val\_score. We iterate the cluster number from 1 to 15 to find out the best number of the clustering for kmeans. We judge the cluster is good or not by accuracy. The highest accuracy number would be the best clustering number for this data.

Third and fourth part, aic and bic we import a package from sklearn called GaussianMixture. We firstly set the boundary name lowest\_bic and lowest\_aic as the infinite. Then, we iterate our components from 1 to 30.

For each time, we create a GaussianMixture named gmm and then fit the model with the raw data (digits.data). In the end, we would get the best number with the proper cluster.

# Result

3.1 At first, we show the cluster with ten cluster which is calculate by the cross-validation value. The graph bellow shows the best cluster number is 10 which is also the optimal number of the data.

A picture containing application

Description automatically generated

Background pattern

Description automatically generated

Above figure shows the optimal clutering with 10 cluster.

3.2 The result of the silhouette predict is that cluster with 9 is best solution. Below is the scatter to show what cluster had been grouped

Scatter chart

Description automatically generated cluster with red circle is being grouped because center of two digits is too close.

3.3 aic has the result of clustering the best cluster with 11.

A picture containing text, stationary

Description automatically generated

Chart, scatter chart

Description automatically generatedThe graph above shows that aic have the split the cluster 2 to 2 and 4, so it has 11 clusters in this graph.

A picture containing text, stationary, writing implement

Description automatically generated 3.4 The result of bic is located at 10 cluster which is the same as the optimal cluster.

The scatter shows the cluster with 10.

Background pattern, scatter chart

Description automatically generated

Compare with these four methods to calculate the best number of clusters of Kmeans function, Bic and cross validation have the optimal solution with 10 clusters. However, for silhouette which is grouped the cluster into 9 and aic split into 11.

# Discussion

Only Cross validation and BIC’s optimal number of clusters are 10, while that of AIC is 11 and that of Silhouette is 9. Obviously, the BIC and Cross Validation have better performance on processing that dataset. Between BIC and AIC, we find that 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. Both AIC and BIC introduce penalty items related to the number of models parameters. The penalty term of BIC is larger than that of AIC. When the number of samples is too large, it can effectively prevent the model from being too complicated due to the high accuracy of the model. Therefore, BIC get the better performance in that experiment.

The method of cross validation got the highest score, because it uses the part of data set to test and check the accuracy and obtain the highest score. Comparing to Silhouette, cross validation has more times to check and train the model and all samples are used as training set and test set, and each sample is verified once, which ensure to catch high accuracy.

# Reference

1. Jteng. 2014. Gaussian mixture clustering. Retrieved form <https://blog.csdn.net/liudongdong19/article/details/80968396>
2. Lu. 2014. Model selection of AIC and BIC. Retrieved form <https://blog.csdn.net/jteng/article/details/40823675>
3. scikit-learn developers .2020. sklearn.manifold.TSNE Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html>
4. scikit-learn developers.2020. sklearn.metrics.precision\_score¶ Retrieved form

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html>

1. scikit-learn developers .2020. sklearn.metrics.GaussianMixture Retrieved form cross\_val\_score

[https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html#sklearn.mixture.GaussianMixture](https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html" \l "sklearn.mixture.GaussianMixture)

1. scikit-learn developers .2020. sklearn.metrics. cross\_val\_score Retrieved form

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html>