Practice of Typical Unsupervised Learning Algorithms

Xingang Liu
AI and Machine Learning Class Fall 2023
Southern University of Science and Technology
Shenzhen, China
2652173921@qq.com

***Abstract*—In this project, the author completed the construction of k-means,Principal components analysis(PCA) and linear autoencoder algorithms, and tested their performance with training sets and pictures. This report states the basic idea of the project and the experimental results.**

I. Introduction

Primary problems and approaches in unsupervised learning fall into three classes: 1. Dimensionality reduction: represent each input case using a small number of variables (e.g., principal components analysis, factor analysis, independent components analysis) 2. Clustering: represent each input case using a prototype example (e.g., k-means, mixture models) 3. Density estimation: estimating the probability distribution over the data space. [1]

K-means ,PCA and autoencoder are classical algorithms of unsupervised learning.

II. Problem formulation

A. Clustering

In practice, the label values of training sets for some classification problems are difficult to obtain or error-prone. In this case, unsupervised machine learning uses such as k-means and improved soft k-means can work well.

B. Dimensionality reduction

In the process of image information transmission, it is often difficult to transmit because of the large amount of data. Therefore, we need to find an algorithm that can compress the amount of data during transmission and minimize distortion after transmission. PCA and linear autoencoder algorithms can achieve such functions.

III. Method and Algorithms

A. K-means

1) Basic model: The author code a class, which takes an passed parameter" k "to represent the final number of categories when it is initialed. When a data set is passed to an instantiated object for classification, k different sample points are first randomly selected as the centers of k classes. Then calculate the Euclidean distance from each sample point to each center, the sample point to which the distance from the center coordinate is the smallest, this point is assigned to the class of the center. Then, the coordinates of all sample points belonging to each class are averaged as the new center coordinates for that class. Recirculating the process of calculating distance classification and updating center center coordinates. When the last two classifications are exactly the same, the loop terminates and the data from the last classification is returned.

2) Optimize: add non-local split-and-merge moves: In the step of updating the classification center coordinates, the farthest distance from the sample points belonging to the class to the class center coordinates in each class is calculated, which is called the radius of the class. The authors set that when the maximum value of all class radii exceeds twice the minimum value, the class center of the minimum radius is replaced by the center coordinate of a random point belonging to the maximum radius class. An upper threshold is set for the number of cycles without changing other steps. The cycle is terminated when the number of cycles reaches the upper threshold or the sample classification does not change before and after two cycles. Returns the result of the last classification.

B. Soft k-means

1) Basic model: On the basis of k-means, the way of updating the classification center coordinates is changed. After calculating the distance from a sample to each center point, like k-means, the distance is fed into the softmax function to

calculate a weight. When updating the classification center coordinates, each new center coordinate is equal to the sum of the coordinates of all sample points multiplied by the classification center weights. The weights and center coordinates are updated. The termination condition is that the weight of the two loops before and after does not change or the number of loops reaches the threshold set by the author. After the loop is terminated, the classification result is output using the classification method of k-means.

2) Optimize: add non-local split-and-merge moves: The same as non-local split-and-merge moves in k-means.

3) Problem and solution:: Problem: When the author used the soft k-means model to complete the clustering of image pixels in Task 7, ”nan” values were easy to appear when the program was running due to the excessive value range of 0-255. Solution: The original pixel matrix is first divided by a value of an appropriate size (about 10-100). If this value is too large, the cluster center gap is too small and tends to overlap. Write this value as ”times”. Then it is sent to soft k-means for clustering. After the clustering is completed, multiply the clustering center by times to get the final clustering results.

C. PCA

The first two dimensions of the pixel matrix are expanded into one dimension, that is, the three-dimensional color pixel matrix is transformed into a two-dimensional data set. In this data set, each row stores three color components of a pixel. Calculate the empirical covariance matrix of the data set and conserve the mean of data set. Then, the ”eigh” function in ”scipy.linalg” library is used to calculate the eigenvalues and eigenvectors of the empirical covariance matrix. The obtained eigenvalues and eigenvectors are rearranged by the size of the eigenvalues using the ”argsort” function. The first k eigenvectors are taken as the principal components of the original data set (that is, a set of bases of the space after dimensionality reduction). The matrix composed of the matrix obtained after subtracting the mean value of the data set is dot multiplied by the eigenvectors, and the coordinates of the previous k eigenvectors in the dimensionally reduced space of the data set are obtained and recorded. The recorded coordinates are multiplied by the corresponding base and then summed, and the mean value of the original data set is added to obtain the reconstructed two-dimensional matrix after PCA compression. Finally, according to the size of the three-dimensional pixel matrix of the original color image, the obtained two-dimensional matrix is transformed into a three-dimensional matrix. loss is represented by the mean square error of the reconstructed 3D matrix and the original pixel matrix. loss is calculated and the pixel matrix obtained after PCA is used to generate an image, which is compared with the original image.

D. Linear Autoencoder

1) Basic algorithms:: Similar to the previous PCA, the first two dimensions of the three-dimensional pixel matrix are first laid out as a column, and the three color channel features of each pixel are retained to obtain a data set in pixels, where each pixel contains the three color channel 3D features. The goal of the linear autoencoder is to train an encoder to reduce the data and a decoder to restore the data after dimensionality reduction. The author’s experimental idea is to apply the MLP training model to train the encoder and decoder (both are parameter matrices), set only a hidden layer as the encoder, use the output layer as the decoder, the output target value is the same as the input value, and use the mean square error of the output value and the target value to represent the loss function. The model is trained, the number of training stops is set to 10000 times, and the encoder and decoder are obtained. The original pixel matrix of the image is put into the model for encoding and decoding, and the output value is generated into the image and compared with the original image.

2) Problem and solution:: Problem:The matrix value obtained after linear autoencoder may exceed the range of color channel 0-255. Solution: Before building, the maximum and minimum values of the pixel matrix are processed, using the where function to assign values that are out of range to 0 or 255.

IV. Experiment Results and Analysis

A. Task 3:

Based on the the wheat seed data set, the results that k equals 3 with different methods are shown as follows.

1) K-means:: The labels and k-means’ classification results of the data set are shown in the figure below:

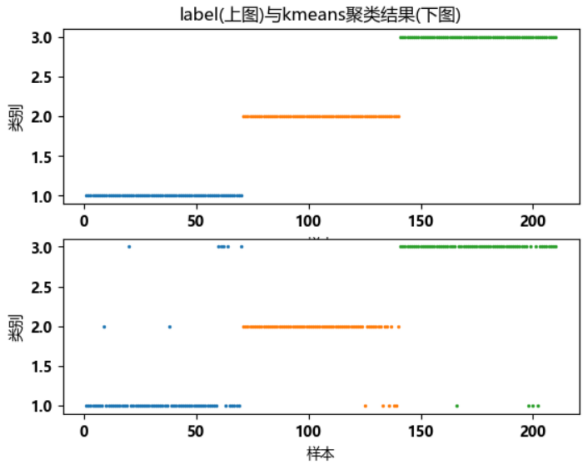

Fig. 1. k-means

Most of 210 samples are classified well in the labels’ standard. It seems that in this case, the result of k-means do

not trapped in a locally optimal solution.

2) Soft k-means:: The labels and soft k-means' classification results of the data set are shown in the figure below:

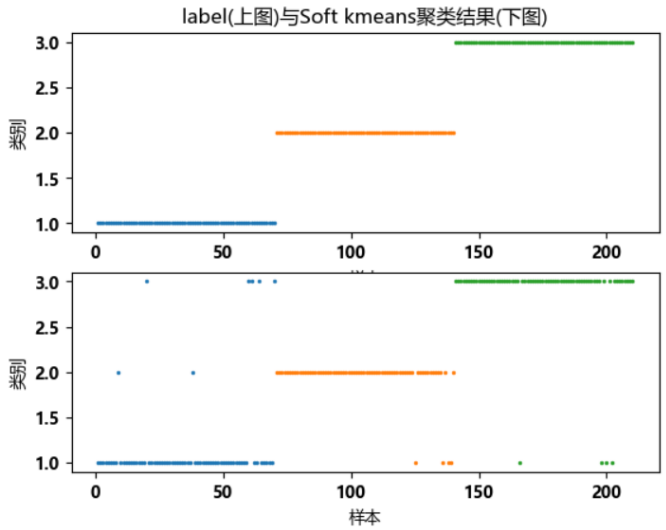

Fig. 2. soft k-means

In this case, compared with k-means, there is no significant difference in the classification results obtained by soft k-means.

3) Performance comparison:: The writer use accuracy compared to the labels to evaluate the performance of k-means and soft k-means. When the initializing category center is the same, the classification results and performance of the two algorithms are shown in the following figures.

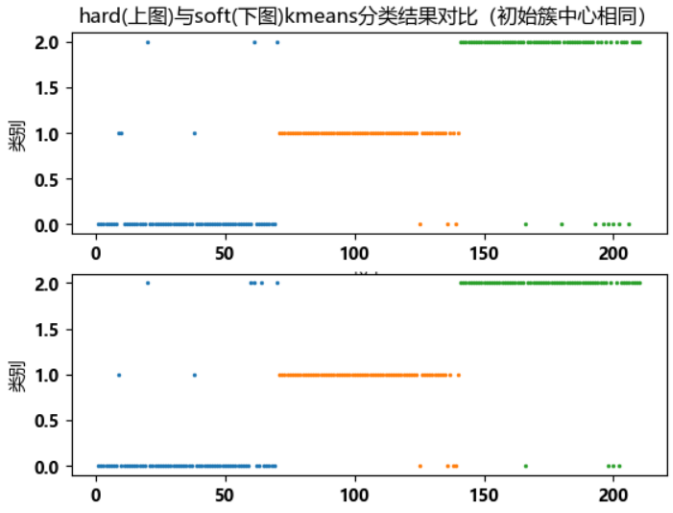

Fig. 3. accuracy

```
kmeans accuracy:0.919047619047619
soft kmeans accuracy:0.9285714285714286
```

Fig. 4. classification results

It is shown that in that case, the performances of the two algorithms are approximately the same.

B. Task 4:

Based on the the wheat seed data set, the results that k equals 10 with different methods before adding non-local split-and-merge moves and after adding non-local split-and-merge moves are shown as follows.

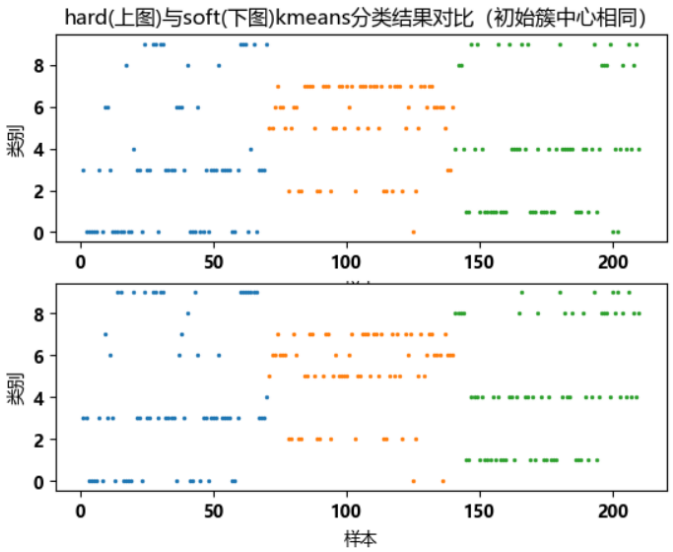

Fig. 5. comparison before merge

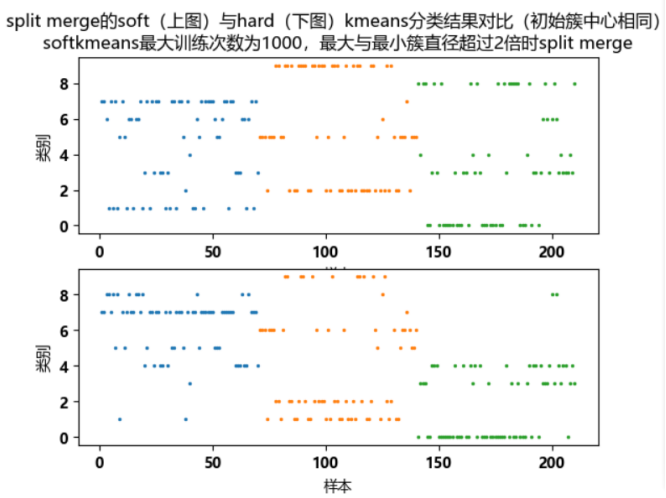

Fig. 6. comparison after merge

When k = 10, Whether non-local split-and-merge moves are added or not, the sample numbers belonging to the same category after classification of the two algorithms, which originally belonged to different categories of data sets, increased significantly compared to that when k = 3. That may because The selection of k value has a significant impact on the accuracy of classification results. When k value is equal to the number of real categories of the sample, the algorithm achieves the best classification effect.

C. Task 7:

1) Results presentation: The comparison of original picture and reconstructed pictures from different dimensionalities and different algorithms are shown in this section.

Fig. 8. PCA(k = 3),loss:2.6149027165521234e-28

Fig. 7. original future

Fig. 9. PCA(k = 2),loss:30.93441966932399

Fig. 10. PCA(k = 1),loss:174.50948073967038

Fig. 12. linear autoencoderk = 2,loss is about 33

Fig. 11. linear autoencoderk = 3,loss is about 0.1

Fig. 13. linear autoencoderk = 1,loss is about 722

Fig. 14. softkmeans(k = 1)

Fig. 15. softkmeans(k = 2)

Fig. 16. softkmeans(k = 3)

Fig. 17. softkmeans(k = 6)

Fig. 18. softkmeans(k = 10)

Fig. 19. softkmeans(k = 20)

2) Analysis:

a) PCA and linear autoencoder: With the decreased dimensionality, the color richness of reconstructed pictures is significantly reduced. Meanwhile when the dimensionality of reconstructed pictures are less than the original picture, the reconstructed pictures of PCA looks a bit weird with some pixels that look greatly different from the original picture. Although the loss of PCA is smaller than that of linear autoencoder with the same imensionality, the same result don't appear in the reconstructed pictures of linear autoencoder. It may because when extracting the principal component of PCA, the feature information corresponding to those sub-components will be directly ignored. The pixel color is only different from the general value in size, so the pixel value will change color directly if it is larger or smaller, which will lead to huge visual difference.

b) Soft k-means: From the results shown, it can be intuitively seen that the number of colors in the reconstructed picture is equal to the k value of soft k-means. When k is small, the main outline of items in the original picture are extracted obviously to the reconstructed pictures(Except k equals 1. When k equals 1, the reconstructed picture consists of just one color which is the mean of the pixels in the original picture).

V. Conclusion and Future Problems

Based on the wheat seed data set, k-means and soft k-means do well when the correct k value is chosen(that is k equals 3). When k equals 3, both algorithms can reach an accuracy of over 90 percent. Besides, during the experiments, no obvious differences are observed from the results of the two algorithms. In the problems of picture compression, the loss of PCA is less than that of linear autoencoder; however, in terms of actual visual effects, the pictures reconstructed by linear autoencoder look more harmonious. If the picture is constructed from soft k-means, the outline of items in the original picture can be extracted well. In the future, the writer can conduct more experiments with different pictures and different data sets.

References

[1] lecture notes