K-Means

Pan Mengyu, ZHANG kai

October 2020

1 Introduction

We use the Optical Recognition of Handwritten Digits Data Set to learn how to use K-Means method. This data set was used in 1998 [3] to valid their method. In this data set, there are 3823 training data and 1797 test data. For each data, it contains 64 input and 1 class attribute.

2 method

K-Means is an unsupervised learning method proposed by J. MacQueen in 1965 [2]. It aims to partition n samples into k clusters in which each sample belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. It has been successfully used in segmentation, computer vision, and astronomy among many other domains.
This method has 4 steps:

1. At first it randomly initializes the clusters. Then it calculates the center of each cluster. After centers are calculated, it repeats the step "assignment" and the step "update".
2. In the step "assignment", it calculates the distance between each point and each center and assign the points to the nearest cluster.
3. In the step "update", it calculates the centers of the new clusters.
4. If the new centers don't have much difference with the previous centers, it stops the repetition. If not, the loop still goes on.

According to its four steps, we have set some of its critical parameters before we use kmeans method. In our code, we set number of clusters and we set the random state to initialize the method
However, in the dataset, there are a number of features with huge data. In this case, we need to reduce dimensionality and we choose principal component analysis (PCA)[1]. Its principal is to find the main part of the dataset to represent the original one. It uses the linear transformation to reduce the dimension of the original data.

3 Result

We applied the K-Means algorithm to the aforementioned dataset and evaluated the cluster performance. The training data includes 3823 images and the test data contains 1797 images. Due to the high dimension of the input data, we utilized the PCA to reduce the number of dimensions, from 64 to 25. Then, we leveraged the K-Means algorithm to cluster 10 classes. In the evaluation procedure, we adopted two methods to measure the performance, the cluster metrics, and the classification metrics.

The cluster metrics do not need the label information and it is trivial to count the number of correct classification points and errors. Accordingly, it defines separations of the data similar to some ground truth set of classes or satisfying some assumption such that members belong to the same class are more similar than members of different classes according to some similarity metric. In the module of scikit-learn [4], it provides several metrics, such as homogeneity score, completeness score, v measure score, adjusted rand score(ARI), adjusted mutual info score(AMI), silhouette score, etc. In our experiment, we leveraged several metrics to evaluate the performance, as shown in Figure 1.

method	inertia	homogeneity	completeness	v-meas	ARI	AMI	silhouette
K-Means	2478785	0.743	0.761	0.752	0.663	0.750	0.291

Figure 1: Evaluation result of cluster metrics

The classification metrics mean to evaluate the cluster results as classification results that count the number of correct classification samples. Before calculating the accuracy, we constructed the project between cluster labels and the ground-truth attributes. We took the class with most samples as the label in each cluster. After alignment, we evaluated the accuracy of the clustering algorithm. Besides, we analyzed the effect of dimension reduction on the accuracy, as demonstrated in Figure 2. It can be seen that when the n_components equals 60, the algorithm achieves the best result. When $n\_components = 25$, it obtains the second-best performance. When taking the computation cost into consideration, we concluded that the $n\_components = 25$ was the most appropriate parameters, which indicated that with 25 dimensions, the K-Means algorithm worked the best in this digit number dataset.

To visualize the details of clustering result, we set the n_components equal to 2 so that the results could be draw in an image. As shown in Figure 3, the K-Means algorithm clustered the whole dataset into 10 classes.

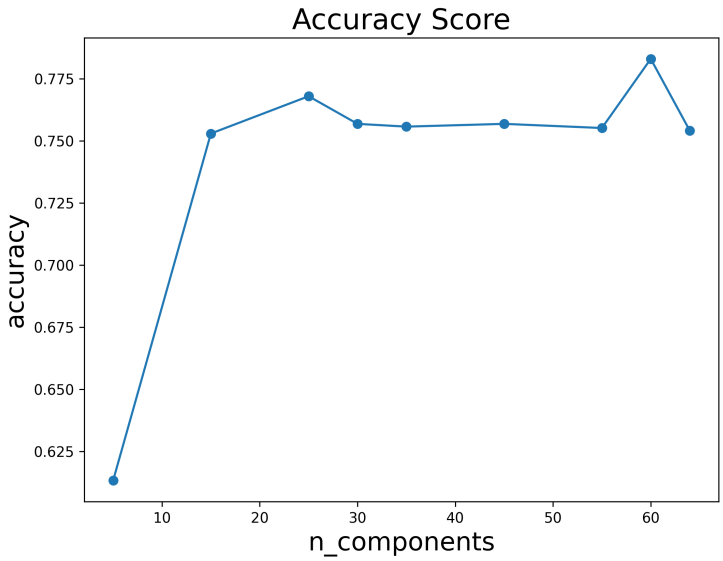

Figure 2: Accuracy scores with different dimensions

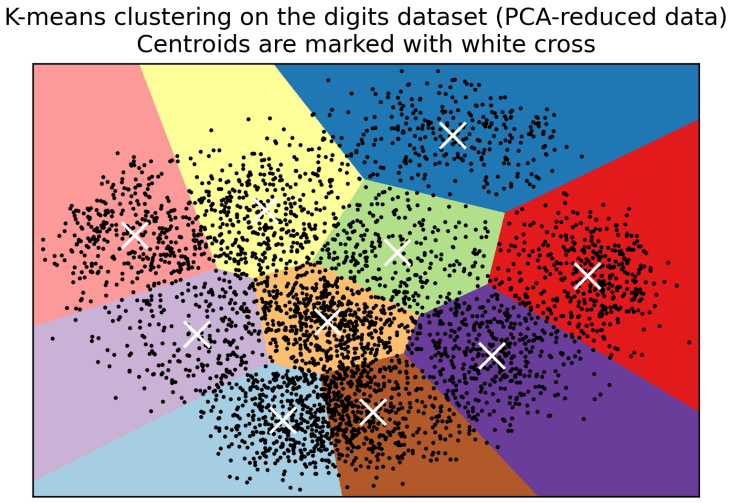

Figure 3: Visualization result of 2 dimensions data

References

[1] Karl Pearson F.R.S. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 6.559–572 (1901).

[2] J. MACQUEEN. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1.281-297 (1967).

[3] *Optical Recognition of Handwritten Digits Data Set*. [EB/OL]. `http://archive.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits`. 1998.

[4] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.