Danmarks Tekniske Universitet

Report 3

Course:
02450: Machine Learning

Author(s):

Anders H. Opstrup (s160148)
Gu Jinshan (s161944)
Huayu Zheng(s162077)

DTU Compute
Institut for Matematik og Computer Science

02450: Machine Learning
Tue Herlau

29 November 2016

Contents

1 Introduction

For this report, the group use the same dataset as project 2, which is the SPAM dataset from http://statweb.stanford.edu/ tibs/ ElemStatLearn/. The dataset is mainly used to detect SPAM emails using attributes such as the frequency of centain words and capital letters.

The group will focus unsupervised learning methods in this report. We will first apply Gaussian Mixture Model and Hierarchical Clustering for clustering the data. Then we use different anomal detection methods to detect outliers. In the association mining part we want to find whether there is correlation between different attributes.

2 Clustering

Clustering is a classification method in unsupervised learning, which we divide data into groups to capture the natural structure of the data. In this section of the report, we are going to cluster the sparm data using both Gaussian Mixture Model (GMM) and hierarchical clustering. In the hierarchical clustering, the cut-off is set at the same number of clusters as astimated by the GMM. Also, the comparision of those two methods will be mentioned afterward.

2.1 Gaussian Mixture Model

We validate the number of clusters for Gaussian Mixture Model based on the EM algorithm using cross-validation.Apart from the cross-validation the optimal number of clusters are sometimes derived by penalizing model complexity based on the Bayseian Information Criteria (BIC) or Akaike's Information Criteria (AIC).

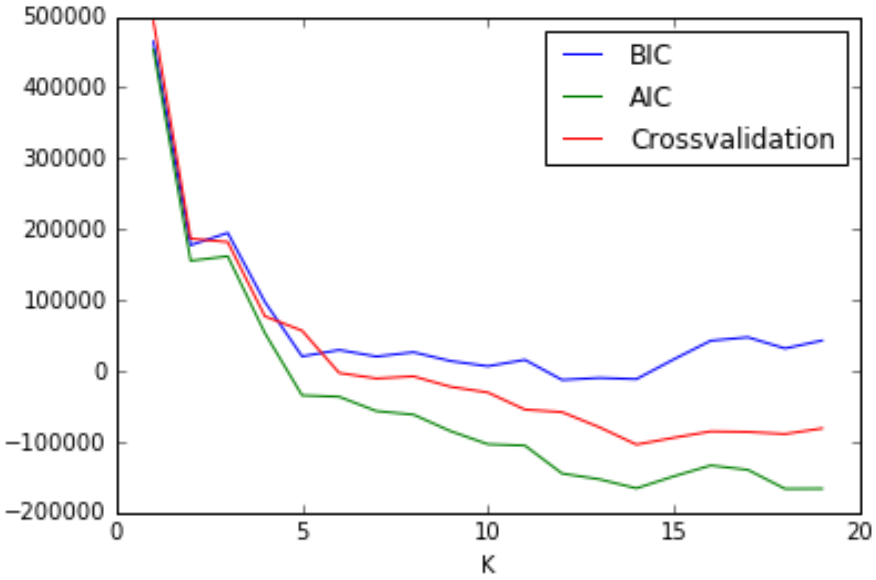

Figure 1: BIC, AIC, and 10-fold crossvalidation

As it shown in fig.1, the group use BIC, AIC and 10-fold crossvalidation to assess the best number of clusters for the SPAM dataset. We compute the three measures for K = 1,...20, and use 10 replicates to avoid bad solutions due to poor initial conditions. Best result will be kepted. The model with lowest AIC and BIC value indicates the model with best trade-off. We can tell from fig.1 that when K = 5,both BIC and AIC reach a elbow, the distortion goes rapidly before that point. In this case, the group considers the ideal value of K would be 5.

We do a dimensionality reduction for the data before we cluster them. The data has been reduced into two componets. The cluster of SPAM data by GMM with 5 clusters, is shown in fig.2. We can see from the fig.2 that few data are clustered into a large cluster with low purity, while other four clusters are locate closly with high purity. The cluster center calculated by GMM is 1.65138013475.

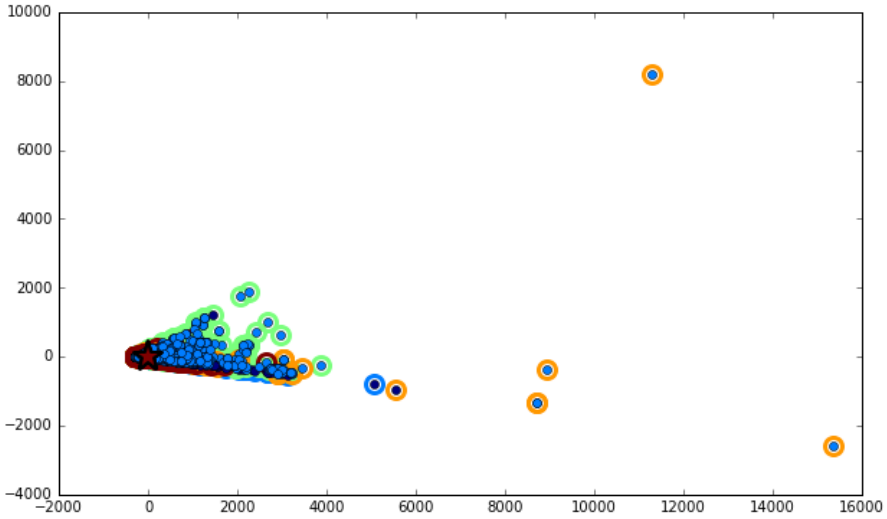

Figure 2: The cluster from GMM with 5 clusters

2.2 Hierarchical Clustering

Hierarchical cluster is defined by a stepwise algorithm which merges two objects at each step,the two which have the least dissimilarity.

Accroding to the clusters we did in GMM, we find out that a few data constitute a big cluster with low purity while other clusters are more centralized with high purity. It turns out that most of the datas should be classified into one cluster.

In this case, the group choose standardized euclidean and minimim dissimilarity (single linkage) to draw the dendrogram, which is shown in fig.3. The branches of this tree are cut at a level where the jump in levels of two consecutive nodes is large. From our dendrogram, we can cut it into 3 classes.

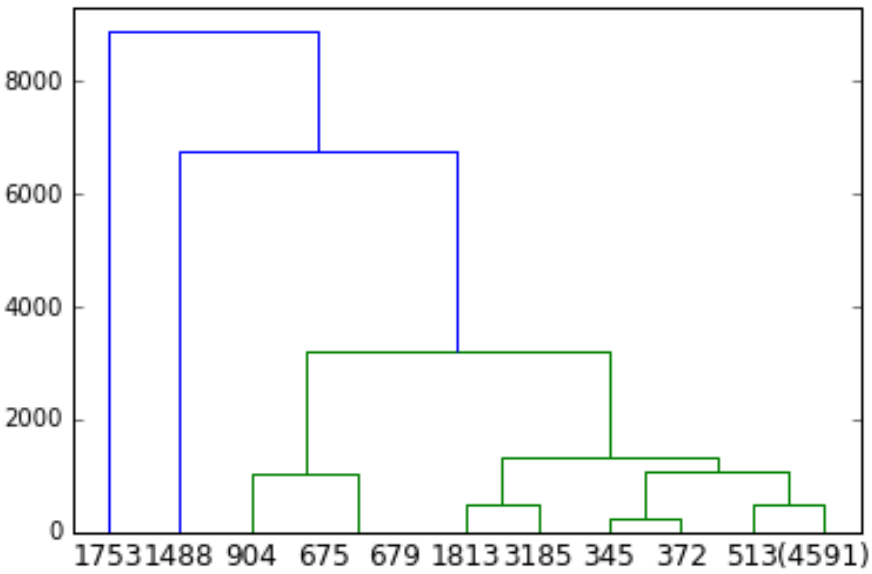

Figure 3: The dendrogram drawn by single linkage cluster analyses of the standardized euclidean distence

3 Evaluation

Comparing the GMM and hierarchical clustering, GMM classifies the data into 5 clusters in which one of them has low purity, while other clusters are so centralized that should be classified into the same cluster. In the hierarchical clustering, two objects merges in each time of the clustering, and it formed a dendrogram which is better cut into three classes. Hierarchical clustering has less error in classification because the data in the same cluster are the clostest, acroding to the standardized euclidean and single linkage. Also the clusters are not as big as the one in GMM, which means the clusters are more specific.

4 Outlier detection/Anomaly detection

We rank the observations using Gaussian Kernel density (leave one-out), KNN density and KNN average relative density seperately. As shown in the following figures, the Gaussian Kernel density seems to be a good model for this data.

Gaussian Kernel density: From the figure we can find that the curve bends between 30% to 40%. Since the propotion of the spam emails in the data is 39.4%. It seems that the Gaussian Kernel density can detect the spam emails as the outliers at a relative high accuracy.

KNN density: The KNN density can only detect few outliers. It is impossible to find spam emails using this density.

KNN average relative density: The KNN average relative density can detect some outliers. However, we cannot infer which part are the spam emails.

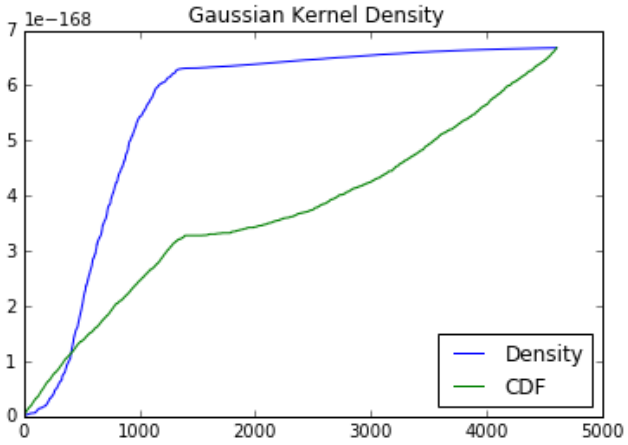

Figure 4: Gaussian Kernel density: Outlier score

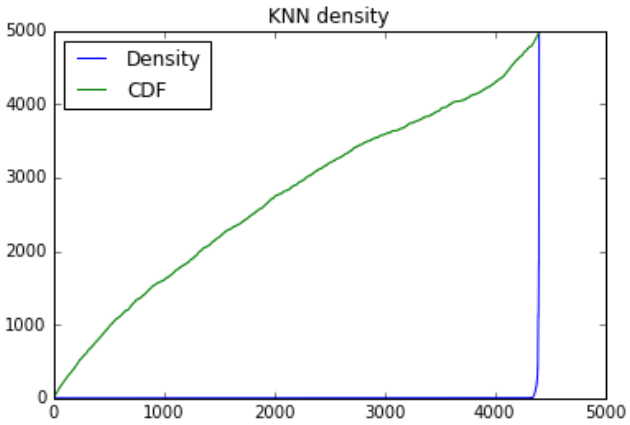

Figure 5: KNN density: Outlier score

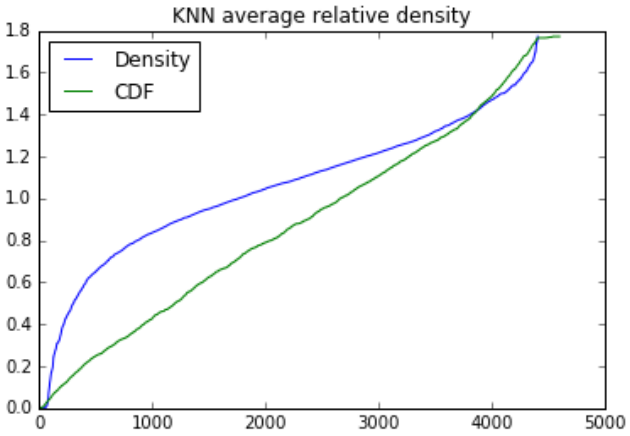

Figure 6: KNN average relative density: Outlier score

5 Association mining

6 Conclusion