



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

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Clustering</b>	<b>2</b>
2.1	Gaussian Mixture Model . . . . .	2
2.2	Hierarchical Clustering . . . . .	3
<b>3</b>	<b>Evaluation</b>	<b>4</b>
<b>4</b>	<b>Outlier detection/Anomaly detection</b>	<b>5</b>
<b>5</b>	<b>Association mining</b>	<b>7</b>
<b>6</b>	<b>Conclusion</b>	<b>8</b>

# 1 Introduction

As described in the conclusion of the first report the group have had some concerns about the dataset (forest fires). Therefore the group decided to change the dataset. The dataset was too small to make any interesting analysis. The group have chosen the SPAM dataset from <http://statweb.stanford.edu/tibs/ElemStatLearn/>. This dataset has around 50 attributes, compared to 12 the forest fires dataset has and the SPAM dataset has around 5000 records where the forest fires dataset only has around 500.

This report covers the second assignment for course *02450: Introduction to Machine Learning and Data Mining*. The assignment's purpose is to solve a relevant regression problem and classification problem for our dataset. The group decided to predict the variable  $word_{freq_{credit}}$  which is the frequency of the word credit used in an email, which could be an indication of SPAM mail. For the classification problem the group decided to identify SPAM from not-SPAM.

## 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 Bayesian Information Criteria (BIC) or Akaike's Information Criteria (AIC).

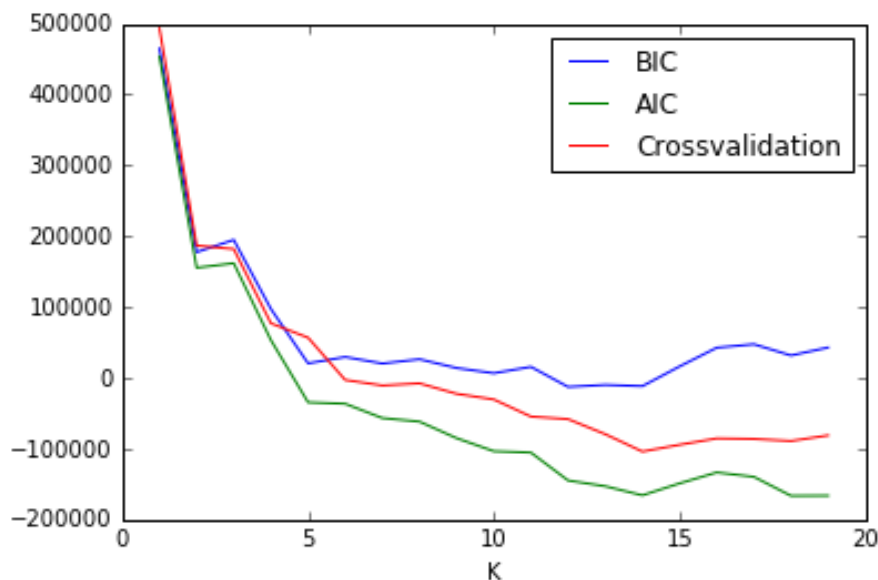


Figure 1: BIC, AIC, and 10-fold crossvalidation

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  $K = 1, \dots, 20$ , and use 10 replicates to avoid bad solutions due to poor initial conditions. Best result will be kept. 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.

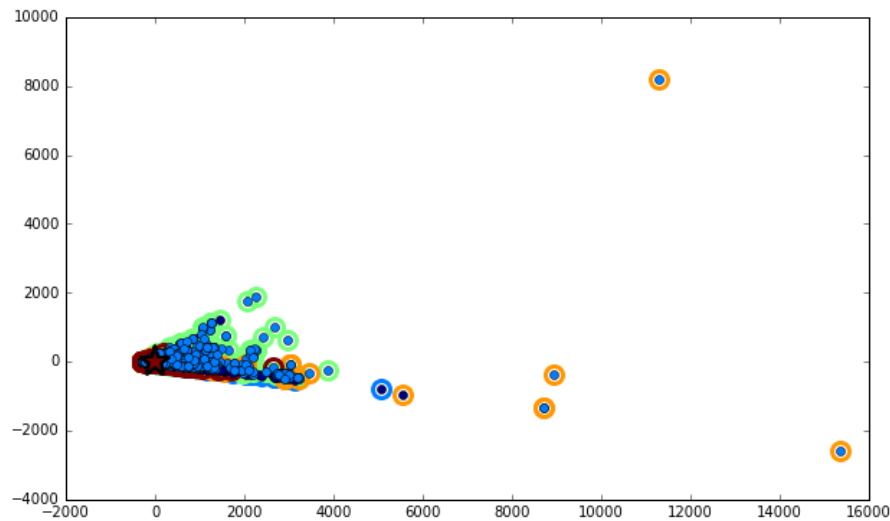


Figure 2: The cluster from GMM with 5 clusters

## 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.

According 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 data should be classified into one cluster.

In this case, the group choose standardized euclidean and minimum 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.

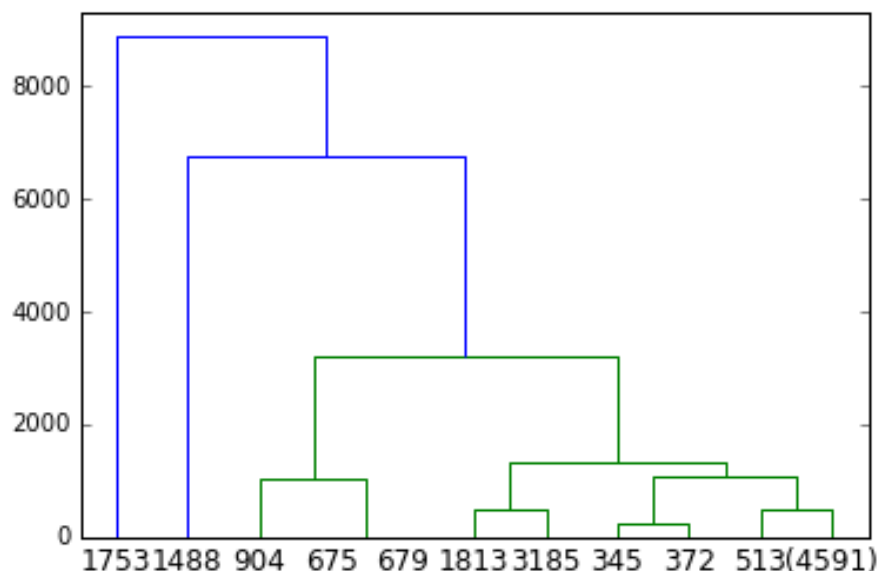


Figure 3: The dendrogram drawn by single linkage cluster analyses of the standardized euclidean distance

### 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 closest, according 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 separately. 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 proportion 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.

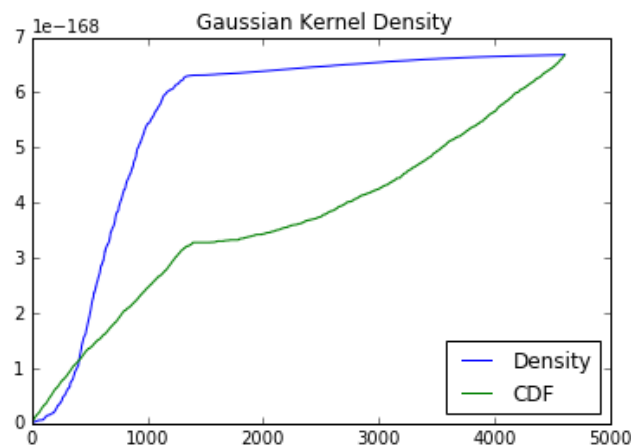


Figure 4: Gaussian Kernel density: Outlier score

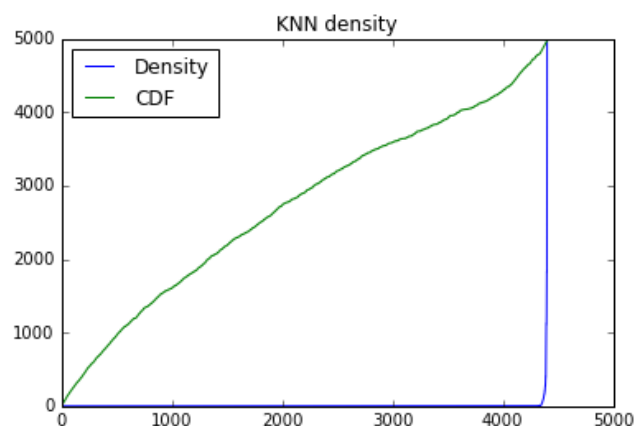


Figure 5: KNN density: Outlier score

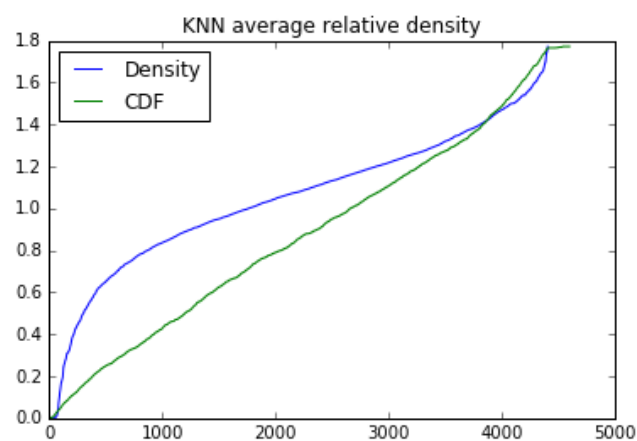


Figure 6: KNN average relative density: Outlier score



## 5 Association mining

## 6 Conclusion