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

As described in the conclusion of the first report the group have had some concerns about the dataset (forest fires). Therefor 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 assignments 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 Bayseian Information Criteria (BIC) or Akaike's Information Criteria (AIC).

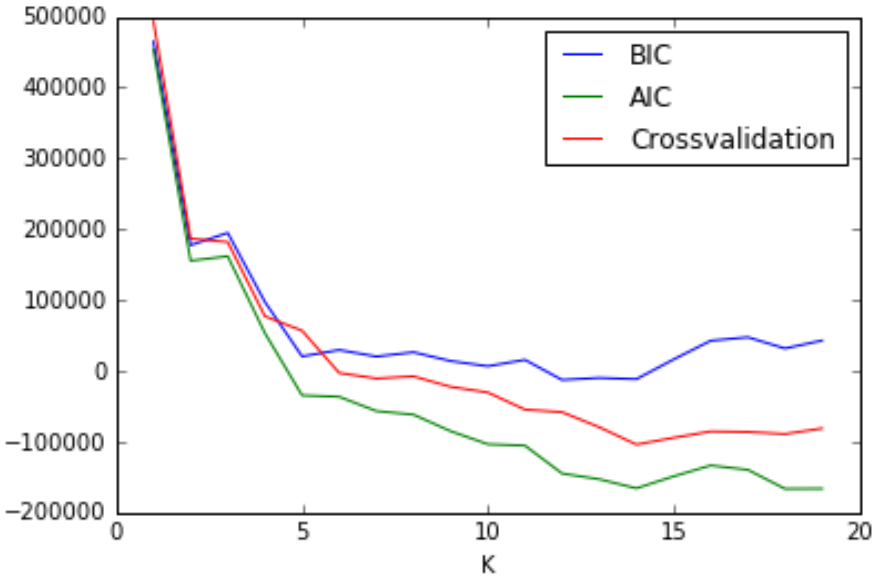

Figure 1: BIC, AIC, and 10-fold crossvalidation

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 did a dimensionality reduction for the data before we did the cluster. The data has been reduced into two componets. The cluster of SPAM data by GMM with 5 clusters, is shown in fig.2.

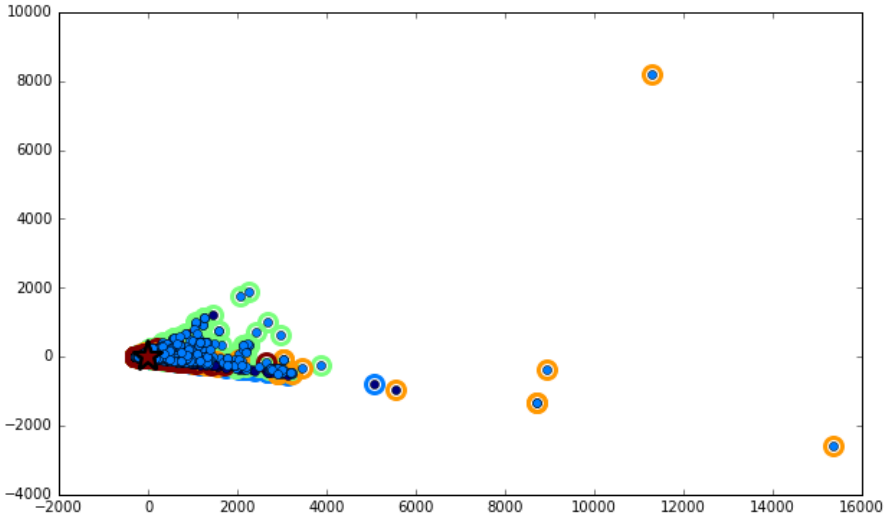

Figure 2: The cluster from GMM with 5 clusters

2.2 Hierarchical Clustering

3 Outlier detection/Anomaly detection

4 Association mining

5 Conclusion