Classify the glass samples using Gaussian Mixture Models (GMM)

Ville Sillanpää, k84338 - Lauri Viitanen, 338853
ville.sillanpaa@aalto.fi - lauri.viitanen@aalto.fi

April 12, 2015

1 Introduction

In this project we analyze a glass data set using both unsupervised and supervised machine learning approaches. The glass dataset [1] has been provided by the US Forensic Science Service and is available for download from the UCI machine learning repository. The dataset consists of samples for 6 types of glass (class label 4 is missing in the current data set); defined in terms of their oxide content (i.e. Na, Fe, K, etc). In particular, the dataset contains 214 samples characterized by 9 features and the samples have been categorized into 6 different classes of glass.

The goal of the project is to classify the glass samples using Gaussian Mixture Models (GMM) in unsupervised and supervised approaches. In the unsupervised approach we use GMMs to find clusters from the full data set and compare the clustering results with the ground-truth class labels that are provided with the data set. In the supervised approach we fit a separate GMM for each class label and evaluate the classification performance using the fitted models on the test set. The dataset has been randomly divided into training (75 %) and test sets (25 %) keeping an equal proportion of different class instances in the training and test sets. In addition, we compare the performance of the GMM model with the k-nearest neighbors classifier.

2 Methodology

2.1 Unsupervised Approach

DP-GMM used with Gamma prior on concentration parameter, Gaussian mean on gaussian mean and inverse wishart on covariance matrix. Model inferred using Gibbs sampling with algorithm by (insert name here). For more detailed description of model see (article).

Several runs of 5000 samples were were ran. Below is a scatter plot of 5000th sample from one of the runs. The visualization was produced with multidimensional scaling. PCA was tried first, but since the first two components explained only two thirds of variation, we chose to to MDS instead.

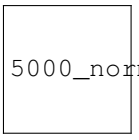

Figure 1: MDS of unsupervised clustering

As you can see from **??**, two dense clouds of points have been identified to belong to three clusters. This result was common to all runs. The points that are spread more sparsely were assigned to various clusters in various runs.

The runs created 3-5 clusters, with cluster assignments switching mostly within these sparsely spread points. This means that the model is able to consistently identify the

two dense clouds as three different clusters, but that the model is more uncertain about the membership of the more sparse points. Based on this model it is difficult to say what clusters the sparse points belong to or that how many true labels these sparse points represent.

Below is a histogram, which compares our model's label distribution to the true distribution of labels

Figure 2: Histogram comparison of predicted versus true labels

As we can see from 2, There are 7 true labels as opposed to the 5 our model found. Some facts: Two large ones are identified relatively well, but remaining clusters are more troublesome. 5 and 7 might be similar. But based on histogram its difficult to say if it is so. Let's look at MDS-visualization of true labels to confirm.

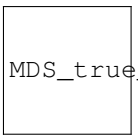

Figure 3: Visualization of data using true labels

As we can see from 3, the lower dense region is more overlapping than what our model predicts. Moreover, quite a few of the more sparse points seem to belong into the label who is very densely concentrated (yellow). Apparently the sparse region contains observations from two additional labels. They seem to overlap so much that it is difficult to see how a gaussian mixture model would be able to find these two labels correcrly. Model and ground truth seem to agree the most about the upper dense cloud.

2.2 Supervised Approach

In the supervised approach, the Gaussian mixture model is fitted to every class label separately. The model complexities range from 2 to 10 components. Class label 4 has no samples in the training data, class label 6 has only six samples and class label 5 has nine samples, meaning that using more than nine classes already drops the number of class labels from seven to four. Only classes 1 and 2 have more than 21 samples.

The GMM is fitted to the training data using the Expectation-Maximization (EM) algorithm. The training halts after 200 iterations or if the log likelihood has not changed enough (0.001 units) during the past few iterations (five). The resulting fitted Gaussians are then applied to the test data of the same classes, producing the log likelihood of the test data being produced by the fitted GMM.

VISUALIZATION HERE

THEN TALK ABOUT THE MODEL COMPONENT COUNT DISTRIBUTION

VISUALIZATION HERE

3 Results

4 Discussion

[1] http://archive.ics.uci.edu/ml/datasets/Glass+Identification

Appendix A

Matlab code for asdf.