Classify the glass samples using Gaussian Mixture Models (GMM)

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

April 11, 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

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.