Instructor's Name: Pekka Marttinen
Project Assistant: Muhammad Ammad-ud-din
Course Title: Machine Learning: Advanced Probabilistic Methods
19th March, 2015

T-61.5140-Course Project:

As a course project, you will 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. 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.

Unsupervised Approach: Use GMMs to find clusters from the full data set (combining training and test sets) and compare the clustering results with the ground-truth class labels that are provided with the data set. In this approach class labels should not be used in the modeling phase and part of the problem is to find a suitable number of distinguishable clusters in the data (one label equals one GMM component). Choose a suitable model selection technique (e.g., BIC/cross validation) to determine the optimal model during the training and choose appropriate evaluation criteria for investigating the results.

Supervised Approach: Your task is to fit a separate GMM for each class label and evaluate the classification performance using the fitted models on the test set. In this approach, class labels will used to train class-specific GMM models that will used to predict labels for the test set. Use your own choice of model selection technique (e.g., BIC/cross validation) to choose the optimal model during training and the evaluation criteria for the test set. In addition, compare the performance of the GMM model with the k-nearest neighbors classifier (Note, in order to perform a fair comparison, the samples in different folds of the training set should be the same for the two methods).

Summarize your results and findings in a short report (max. 4 pages, excluding the title page).The report should contain a title page (with your names and student numbers) and three sections, namely 1) Introduction, 2) Methodology (including a description of the data set, the models used, and details on your training procedures, model selection and evaluation criteria), 3) Results and Discussion. Figures and/or tables can be used to present the results. Remember to justify your choices and critically review the results.

The project can be done in pairs. You can use all freely available Matlab code in your solution, e.g., code implemented during the course or demos accompanying the course book.

The report should be returned by Sunday, April 19th, by email to *muhammad.ammad-ud-din'at'aalto.fi.*

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