T-61.5140 Machine Learning: Advanced Probabilistic Methods
Course project

Kristian Hartikainen (222956)
kristian.hartikainen@aalto.fi

Risto Vuorio (84525R)
risto.vuorio@aalto.fi

April 20, 2015

1 Introduction

Our task in the course project was to utilize what we have learned on the course about unsupervised and supervised learning, different model selection methods and gaussian mixture models on a data set from US Forensic Science Service. The dataset contains samples of 6 different kinds of glasses with 9 different attributes to characterize them.

In the unsupervised approach we take the data and try to fit a model to it without prior knowledge of the data classes. The classifier is generated by trying to fit Gaussian Mixture Models with different numbers of clusters and then selecting among the generated models.

In the supervised approach we know the clustering of the training data and can use that information to generate the model. The knowledge of the classes is used to fit GMM for each of the classes separately. The classifier itself works so that the data is compared to each of the GMMs and the one which gives the highest likelihood.

2 Methodology

2.1 Unsupervised Approach

We did the unsupervised model selection by using few different 'training-validation-testing' setups discussed in the lecture 7 slides. Three different selection criteria were used: BIC, AIC, and cross-validation. BIC and AIC both penalize the likelihood of the model, where as the cross-validation is often used when the goal is prediction, that is, how accurately a model will perform in practice.

We started by fitting 10 different Gaussian Mixture models, with 1-10 mixture components, to the using the combined training data and test data (complete_data in the code), and choosing the 'best' one by minimizing the AIC and BIC values or maximizing the likelihood for the cross-validation. The cross-validation was done by using fold-count of 5 and 10, however the results were pretty much identical so we chose to omit the results for fold-count 10 and only consider the case of 5.

It seems like the multidimensionality of the data might cause the poor performace of the model selection. Thus we also did the model selection by first applying dimensionality reduction (principal component analysis) for the data. We ran the same tests for the data reduced to 2, 4 and 5 dimensions. The first two principal components explain 73.94% of the data, first four components 94.92% and the first five components 98.23% of the total variance.

As we can see from the table in the results section, the dimensionality reduction makes the model selection to be much more likely to hit the correct ballpark (6 mixture components).

The results for model selection criteria (AIC, BIC, and likelihood for cross-validation) are plotted in the figures in the results section. Each plot contains the results for the complete dataset, and data that is reduced to 2, 4 and 5 dimensions.

2.2 Supervised Approach

We did the supervised Gaussian Mixture model fitting by first training a mixture model with the training data for each class label, such that we end up having 6 different Gaussian Mixture models. After that we use those models to predict labels for the test data, and choose the class that maximizes the probability of the point being in that Gaussin model.

We got 44.44% accuracy with this tactic, and comparing it to the k-nearest-neighbour classifier using 5, 10, 15, 20 and 25 nearest neighbours, we see that kNN classifier works much better, resulting in 68.52%, 59.26%, 59.26%, 59.26%, 61.11% classification accuracy.

We think that this has something to do with our very bad approach, or the multidimensionality of the data.

3 Results and Discussion

```
PCA dimensions  AIC BIC CV
None 2 3 10
2 7 10 6
4 6 7 5
5 4 4 10
```

Table 1: Resulting number of Gaussian Mixture model components.

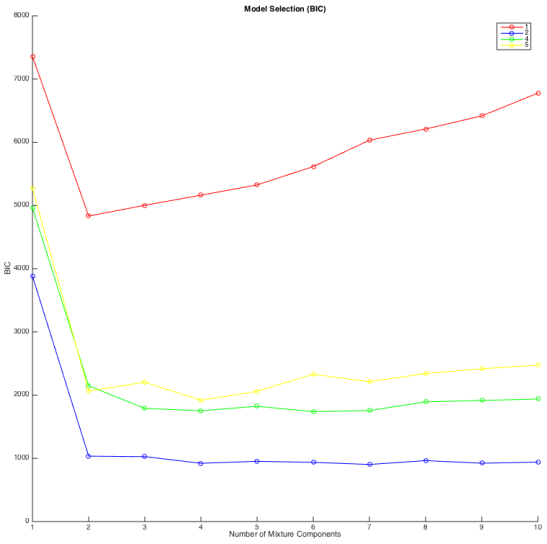

Figure 1: BIC values for the unsupervised model selection, using different data reduction and number of components.

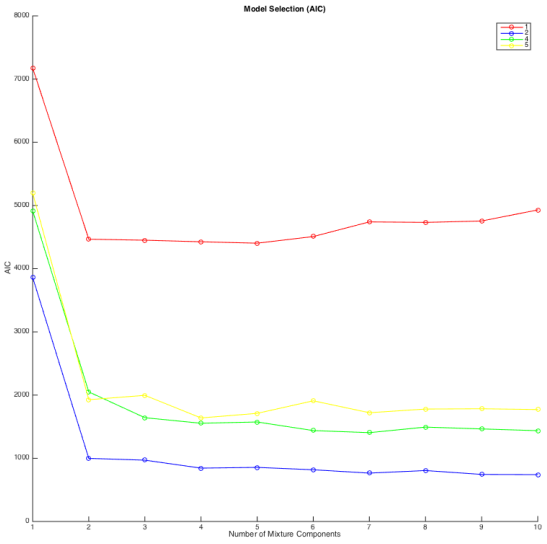

Figure 2: AIC values for the unsupervised model selection, using different data reduction and number of components.

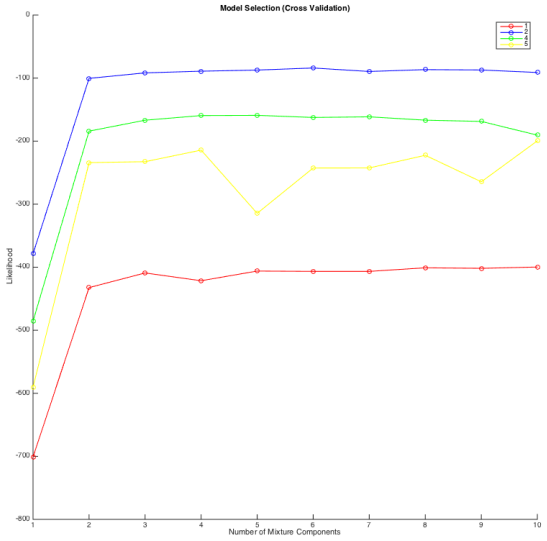

Figure 3: Likelihoods for the different numbers of mixture components, using cross-validation.