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 19, 2015

1 Introduction

$$d(x,y) = \begin{cases} 2, & if x > y \\ 1, & if x < y \\ 0, & if x = y \end{cases} \tag{1}$$

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.

After fitting the model, we choose calculate the AIC and BIC values for the fitted model.

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.

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.

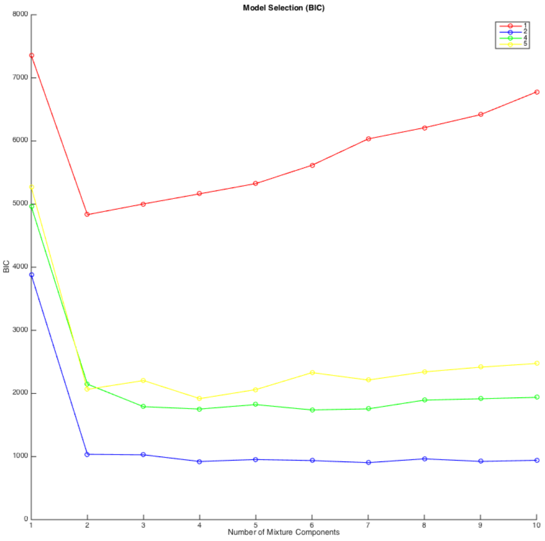

Figure 1: BIC values for the unsupervised model selection, using different data reduction and number of components.

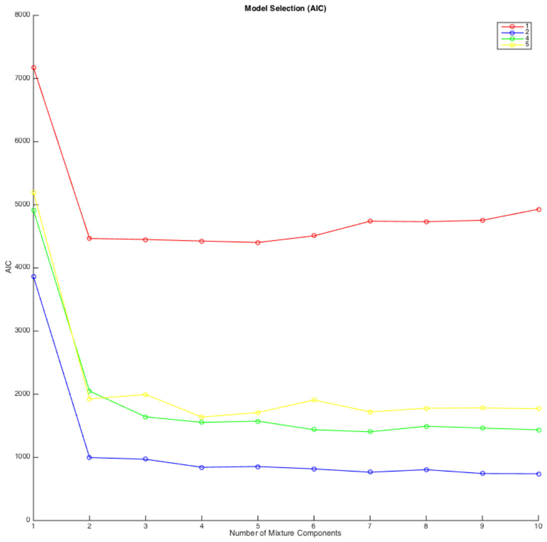

Figure 2: AIC values for the unsupervised model selection, using different data reduction and number of components.

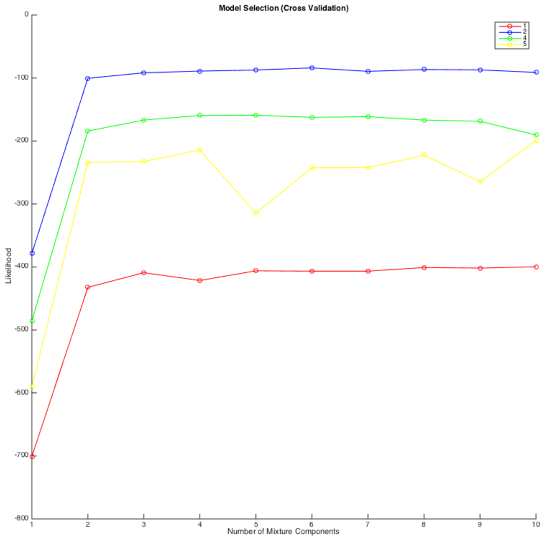

Figure 3: Likelihoods for the different numbers of mixture components, using cross-validation.