

DM825 - Introduction to Machine Learning

Obligatory Assignment 2, Spring 2011

Submission to marco@imada.sdu.dk

Deadline: 15th June 2011 at 9:00

Although it is acceptable that students discuss the assignment with one another, each student must write up his/her homework on an individual basis. Each student must indicate with whom (if anyone) they discussed the assignment.

Exercise 1 – Expectation-Maximization Algorithm

Consider the task of classifying a point in the Euclidean space generated by a mixture of 3 multivariate normal distributions. Each multivariate normal distribution has a distinct mean vector $\vec{\mu}$ and covariance matrix Σ . There are a total of $1000 \cdot 3$ observations and 70% will make the training sample. The R code implementing the generation process is available (`ml_demo.R`) together with the analysis by

- k -means clustering
- k nearest neighbor
- Locally weighted linear regression via kernel
- Regression trees
- Improving trees with Principal Component Analysis
- Linear Discriminant Analysis
- Support vector machines

Your task:

- (a) Implement the Expectation-Maximization Algorithm. Report the R source code in a latex document including it within `\begin{lstlisting}{R}\end{lstlisting}` from the package `listings`.
- (b) Compare the results of the EM algorithm against the other methods on the basis of percentage error on the test data. Report the comparison in a table.
- (c) What are from the theoretical point of view the differences in this specific learning task between the EM algorithm and the linear discriminant analysis? Are the empirical results the same?