Master Thesis: Clustering sequential data using Mixture of Hidden Markov Models

Nowadays, the amount of information about the customer behavior is richer than ever. Not only are we able to record statical characteristics of the person but also all his actions across multiple channels: On-line, Off-line, POS. This rich landscape of data assets is the key of GfK Market Research practice.

If handled with care this data can provide interesting facts and insights, for example information about the orientation process of customers: What are they searching for, how are they searching and in what order, etc. In order to get access to this kind of information it is essential to identify underlying patterns and structures in the sequential databases. Sequence segmentation can be a suitable solution for this purpose giving the possibility to identify differences and similarities between sequences and to analyze them in a holistic way.

In order to create more actionable insights more advanced techniques must be employed. Common practice in Market Research is to assume that customers are in "hidden" state while performing recorded activities. One wants to infer from the data the number of those states and the most common customer path across those states. Given that there is also heterogeneity across the customers' behavior we also assume the presence of distinct segments in the data.

The aim of this master thesis is to extend Markov Model sequence segmentation to the previously described problem with hidden states. Therefore, two different approaches should be compared and applied on the provided GfK datasets:

i. Two - step procedure: Hidden states should be detected from the behavioral data using classical segmentation techniques like hierarchical clustering, k-means or similar. Resulting sequence data should then be clustered using Mixture of Markov Models (MMM).
ii. One – step procedure: using Mixture Hidden Markov Model (MHMM) to simultaneously predict the hidden states and discover the clusters.

For both approaches one can find a R package *seqHMM* suitable though it is not necessary to perform the corresponding analysis using this particular package. Appropriate datasets will be provided by GfK.

Requirements:

- Studies in mathematics, statistics, informatics or related
- Solid mathematical statistics background
- Programming skills, preferable R

Company: GfK SE **Division**: Marketing and Data Sciences
Time period: tba
Compensation: tba
Supervisor (GfK): Jakub Glinka <jakub.glinka@gfk.com>, Sandra Romeis <sandra.romeis@gfk.com>