

Foundations of Information Retrieval

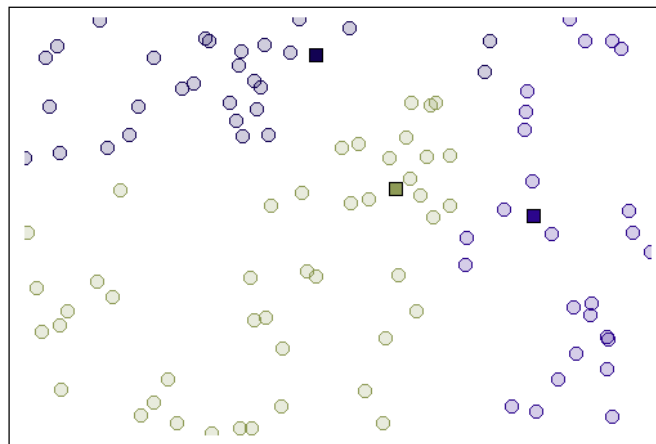
Exercise 9

Exercise session: 08.01.2015

Questions to: Helge Holzmann, holzmann@L3S.de

1. Supervised learning and Naive Bayes classifier

1. The figure below shows a state of the k -means algorithm with $k=3$. The squares represent centroids and circles represent the data points. The color encoding corresponds to the current cluster assignment.
 - a. What phase of the algorithm has just finished and what phase is going to follow next?
 - b. Sketch the changes that will be performed by the k -means algorithm in the next step.



2. The documents that use different terms (e.g. automobile, VW Golf, etc.) for one concept (i.e. car) will be most probably assigned to the same cluster if using term vectors for clustering. Explain why.
3. Explain similarities and differences between classification and clustering. Pay specific attention to the following aspects:
 - a. Training of the system
 - b. Number of cluster / classes
 - c. Presentation of the results