Exercise 1-2Tasks in Data Mining

In the following you have some use cases, where we need to apply a Data Mining method. Determine which kind of Data Mining task (Clustering, Outlier Detection, Classification, ...) it is and decide wheater it is super-vised or unsupervised.

a) Computer Aided Diagnosis: Patients, who suffer on Leukemia (blood cancer) could be separated into two types: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The therapy differs heavily between those two groups and the therapy for AML is harmful for ALL patients (and vise versa).To distinguish these two groups the scientists look at gene expression data. The gene expression of the new patients is compared to the gene expression of patients where the type is known. – **Classification, supervised**

b) Man and maschine: Modern Web search engines answer user requests, which contain only one or just a few search terms. As a general rule, the query returns a huge set of search results. This set is sorter based on relevance by a ranking algorithm. It is up to the user how many links he wants to visit. The challenging part is twofold. Determining the content of a page and the correct ranking of helpful pages. Ambigious words are another challenging aspect here. For example the word “golf” could refer to a car,a sport or to a geographic region. Data Mining techniques are used to optimize the ranking quality and grouping the result set into meaningful terms. – **clustering unsupervised**

c) Recommender systems: An online shop wants to determine interesting goods for registered customers.These goods are suggested to the user after the user logs in into the online platform. The payment history indicate the preferences of each customer. For example, someone who bought the book “Lord of therings” is also interested in the movie. This algorithm could also be used to suggest similar goods to already selected items. – **association rules minig, unsupervised**

d) Identification of important supplier: An online trading platform wants to know, which supplier are themost important ones in terms of turnover and/or revenue. The algorithm supports the questioning of strengthen a relationship to the suppliers. The algorithm could make suggestions for founding a logistic centre close to the suppliers or even a acquisition of the supplier to shorten the delivery time. – **outlier detection, unsupervised**

Exercise 1-3 K-Means definition

What is the usual definition of an optimal clustering for the K-means algorithm?

Goal: Partitioning in kcluster, that our objective function gets minimized (quality criteria: compactness)

•Central assumptions:–Number of clusters is known (input parameter)

Characteristics of the clusters: compactness

Compactness: Distance of all objects within the cluster to a cluster representative (representative=centroids of assigned points) is minimal. –Compactness criteria leads to a spherically shaped clusters

