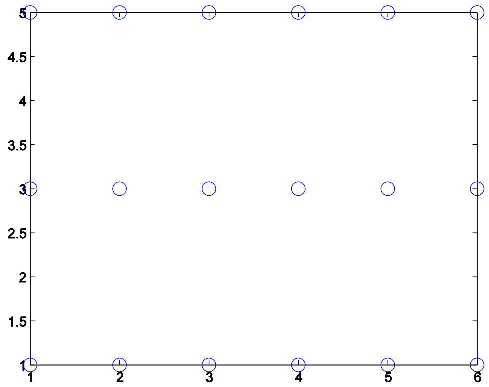
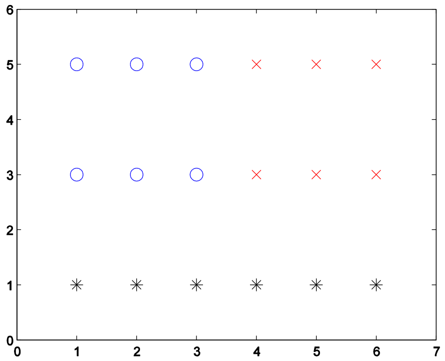
Practising for the first half of M33SAD

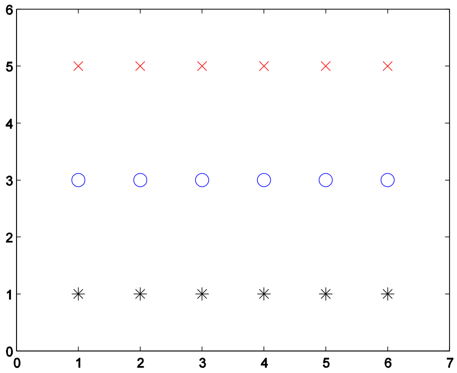
1. Figure 1 depicts input data for clustering. Figures 2 and 3 correspond to clustering using k-means (with Euclidean distance) and using hierachical clustering (single linkage, Euclidean distance). Choose which of the figures corresponds to algorithm k-means and which to hierarchical clustering.



Obrázek 1: Vstupní data

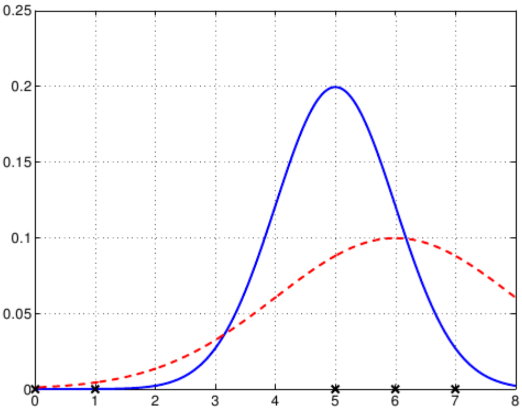


Obrázek 2

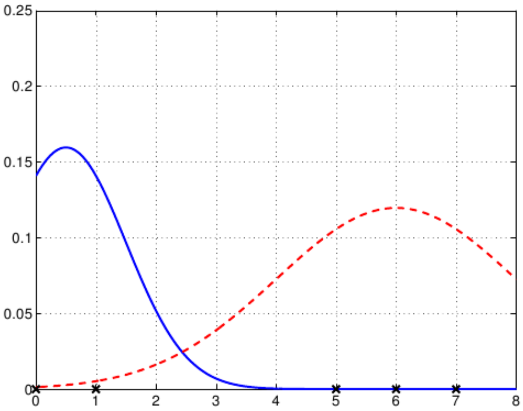


Obrázek 3

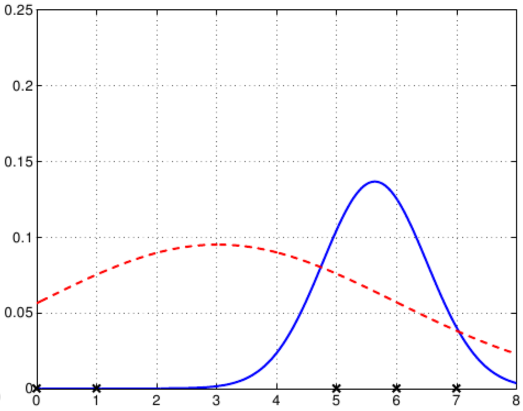
2. Estimate parameters of the mixture of 2 gaussians using EM algorithm. The density of mixture is given by: $f(x,\vartheta)=\alpha N(x;\mu_1,\sigma_1^2)+(1-\alpha)N(x; \mu_2,\sigma_2^2)$. Figures shown below illustrate steps of EM algorithm (the horizontal axis corresponds to parameter *x*, the vertical axis to value of probability density, observations are marked by crosses). A random initialization step (*init)*, a first optimalization step (*step1*) are shown in 2 of the figures below. The third figure is an additional unrelated figure. Figures are ordered randomly. Choose which of the figures corresponds to the mentioned steps: *init* and *step1*. Explain.



a)



b)



c)

3. Let us have a transaction database. Let us assume that the only frequent itemsets of size 3 are the following: *{a,b,c}, {a,b,d}, {b,c,d}, {a,c,d}, {b,c,e}*. Decide which of the following itemsets cannot be frequent: {a,b,c,d}, {a,b,c,e}, {b,c,d,e}.

4. Let us have a transaction database shown in Table 1. Find all of the association rules with support at least 50% and confidence more than 60%.

Transaction	**Items**
T1	beer, bread
T2	bread, peanut butter
T3	beer, milk
T4	bread, jam, peanut butter
T5	bread, milk, peanut butter

Table 1

5. Let us have an alphabet of two symbols *{a,b}*. Let us assume the task of undirected sequence mining. Answer the following questions:
    - How many different undirected sequences of length 3 are there?
    - Sketch how you would generate different sequences of length 4. Show at least one duplicate sequence of length 4.
    - In case of sequences of length 3 you have assured that the only frequent sequences are *{aab,bab,bbb}*. Which sequences of length 4 can be still frequent? Why?