**Data Mining Exercise 5: 20.02.2020**

**Remember to enroll yourself on the examination of the current course no later than 8 days before an examination date!**

1. Iris.txt file contains variables that are highly correlated. In this case principal component analysis (PCA) can be used in reducing data dimension. Run pca() using Matlab on Iris data. How many percent the two first principal components explain from the whole data variance?

2. Let us consider a task of hand written text recognition. A sample image is text.tif. Grayscale image can simply be considered as a matrix that has *h*=(height) rows and *w*=(width) columns. A single pixel value p in coordinates (*i*,*j*) is an integer on closed interval [0 255]. The value of 0 is usually considered black and the value of 255 is white. (a) Draw a histogram of pixel values and propose a "feature" that could be used in separating the background from the text. (b) Propose a "feature" that could be used to find line spaces between the text rows. Hint. An image can be considered as a set of rows or set of columns.

3. Calculate the entropy of image text.tif. How would you interpret the result?

4. Create a Bayes rule that gives us a posterior probability for which a given pixel value from image text.tif belongs to paper background or text foreground. You can freely select "threshold" value that tells the background and foreground of the image apart. Note we can use probabilities of pixel values directly, so no probability density estimation is necessary.

5. If merely nominal (excluding binary) variables are included, the encoding of data can be prepared for some data mining algorithms as neural networks as follows. To simplify, let all variables have the same dimensionality of m (relatively small) for their categories. A value can only be one of the m alternatives. We encode each value with a bit sequence of length m, in which the other bit values are equal to 0 except one that is equal to 1. The latter corresponds to the nominal value, which a case includes for the pertinent variable. This encoding was used for amino acids, which forms a set of {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, when m=20. For instance, if a value occurs to be C, the second bit is 1 and all the others are 0's. Encode polyproline type II (PP II) of amino acid sequence chunk GGKAPAMM. This encoding is valid as neural network input. What is its disadvantage regarding the number of input variables (nodes) required by a neural network? How many input nodes (number of bits) are needed? See all distance measures considered in the lecture material. Which one of them would be suitable for binary data considered here? How many different symbol sequences are there if all possible amino acids are scanned for as long sequences as the preceding sample GGKAPAMM? Give also its approximation in the form of x·10y. (Actually, PP II structures are not searched for with mere sequences, but other information is also required, which results in more complex search problems.)

6. Video clip Video.avi captures a process of calcium intake in human cardiomyocytes derived from stem cells. Study Matlab script below and propose a data mining method for extracting the interesting information from the process.

```
%Place the video clip in same folder from which you run this
script.

v = VideoReader('Video.avi');
while hasFrame(v)
video1 = readFrame(v); %Read one video frame.
fr=rgb2gray(video1); %Change it into gray scale image.
%You are free to consider fr as a matrix that has
%rows and columns.
imshow(fr);
end
```