Data Mining: Practical Assignment #2

Due on Thu & Fri, April 24-25 2014, 10:15am-13:15 & 14:15am-17:15

Task 1

You are given the following recorded data:

	Observed	Expected
Female	50	
Male	60	

The first table depicts the students visiting a lecture at the UKE. Under the assumption that the gender is uniformly distributed, fill in the gaps in the table and then calculate a χ^2 test.

	Science Fiction	Zombie	Animation	Love	Total
Shy	20	6	30	44	
Extrovert	180	34	50	36	
Total					

The second table shows observation of a specific movie genre distinguished between two character types. Here, also fill in the values you need to perform a χ^2 test. Discuss, what does the result shows you?
Hint: Write down your calculations as you are not allowed to use Matlab for that. Further, you find χ^2 tables in the internet. The degree of freedom is 1 for the first data set and 3 for the second. Also assume 5% significance.

Task 2

For the next task, open the BrainIQ.mat. The data consists of nine variables derived from neurological experiments with monozygotic twins. (For further information, refer to *ReadMe.txt*.) Use Matlab to make a simple correlation analysis. For that, use the built-in functions *corrcoef* and *corrplot* and according help pages for function call. What do the coefficients tell you? Are there any correlations? What could be interesting questions, which could be answered by the underlying data and the correlation analysis?

Task 3

Given the following tiny images of size 2×2 pixels. The brightness value is shown on each pixel.

Regard these five images as five data vectors and compute a PCA on them. This means, you compute the following: *(i)* the mean vector *(ii)* the covariance matrix of the data, *(iii)* the eigenvalues of the covariance matrix and *(iv)* the corresponding eigenvectors.

Task 4

On the MIN-CommSy you find a zip file with a Matlab script for face recognition using PCA. Look at the data and into the code and try to understand it (you may comment the code).
Which pre-processing is being made? How many eigenfaces are being used, and why this number?
Run the example program. Describe the mean and the eigenfaces. How good is the reconstruction of the persons who are part of the training data and of those who aren't part of it?