Data Mining: Practical Assignment #2

Due on Thu & Fri, April 20-21 2017, 10:15am-13:15 & 14:15am-17:15

Task 1

You are given the following recorded data:

	Observed	Expected
Female	42	
Male	54	

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

	Science Fiction	Zombie	Animation	Love	Total
Shy	20	12	24	44	
Extrovert	180	28	56	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.

1. What does the results show you?
2. Explain in the context of the task the terms: *null hypothesis (H0)*, *alternative hypothesis (H1)*, and *significance level*.

Hint: You are not allowed to use Matlab for that, so solve the task by hand. You are not expected to put that into your documentation file, but explain the result of your calculation. You will find χ^2 tables in the Commsy. The degree of freedom (dof) is 1 for the first data set, and 3 for the second. Assume 5% significance level (often referred to as α). For supplementary material check the CommSy!

Task 2

In the CommSy you can find the data2.zip that contains data and scripts for the following tasks. For the next task, open the *task2/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 Matlab functions *corrcoef* and, if you want to visualize the correlation, *corrplot* (ignore the main diagonal).

1. What do the coefficients tell you? What are the correlations between the variables?
2. Can you support the statement, that a big head size is positively linked to the IQ?
3. Explain in the context of the task the difference between correlation and causation.
4. Find another example which shows that correlation does not imply causation (not related to this task).

Task 3

In the lecture you learnt about data normalization and smoothing strategies for data preprocessing. Given the data, perform the following:

1. Min-max $[0; 1]$ and zscore normalization for the data vector: $50, 100, 200, 300, 400, 600, 1000, 1100$.
2. zscore normalization for the *fitness* data. How are the two variables correlated (Pearson coefficient)? Does a plot support your result?
3. Smooth the image grayvalues by *bin mean* (bin depth: 3):
   $30, 150, 50, 50, 50, 55, 60, 80, 80, 150, 99, 99, 160, 180, 190, 200, 200, 220, 255, 80, 80$

You are not expected to note down all of your calculations in the document, but explain your steps and report on the results .
Hint: Note the difference between a vector and a matrix in your normalization calculations. For the normalization you can use some of the Matlab built-in functions.

Task 4

In the zip file you can find another set of Matlab scripts for face recognition using PCA (check the ReadMe!). Look at the data and into the code and try to understand it (you may comment the code).

1. Which pre-processing is being made?
2. How many eigenfaces are being used, and why this number?
3. 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?

@home Task

To prepare the next tutorial, your homework will be to learn about the following topics:

- Data representation for classification, especially have a look at confusion matrices. In addition, you should be familiar with the quantities you can derive from them, e.g. False Positive Rates (FPR) and their computation
- Association rules, especially the apriori-algorithm
- Decision Trees: Entropy, Information Gain and their computation

Recommended literature:

1. Han J. & Kamber, M. Data mining: Concepts and techniques. Elsevier/Morgan Kaufmann, Amsterdam, 2006.
2. Kantardzic, M.: Data mining : concepts, models, methods, and algorithms. Wiley, NY, 2011.