Problem 2.1 (Feature Selection) (25 P.)

Given the following dataset for a binary classification problem consisting of 4 instances and 3 features each:

	Feature A	Feature B	Feature C	Class
Instance 1	-2.0	-1.0	0.0	-1.0
Instance 2	-2.0	-1.0	1.4	-1.0
Instance 3	2.0	1.0	-1.0	1.0
Instance 4	-0.2	-0.1	-1.0	1.0

Compute for each subset of features ($\{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}$) the average feature-class correlation $\overline{|r_{fc}|}$, the average feature-feature correlation $\overline{|r_{ff}|}$, and the "Merit".

Answer the following questions:

a) Which subset of features would be chosen by the algorithm $NaiveFS$ for $k = 2$?

b) Which subset(s) of features with cardinality 2 would be optimal according to the merit?

c) Which subset of features with arbitrary cardinality would be optimal according to the merit?

Note: To solve you can either write a python program **yourself** (submit with solution) or compute the solution manually ("per Hand").

Problem 2.2 (Principal Components Analysis) (25 P.)

Given the following dataset

	Feature A	Feature B
Instance 1	0	-5
Instance 2	1	-5
Instance 3	1	-6
Instance 4	2	-4
Instance 5	3	-2
Instance 6	3	-3
Instance 7	4	-3

Conduct the principal component analysis (PCA) manually. Give the following intermediate results:

a) The 7×2 data matrix (X)

b) The mean-corrected data matrix

c) The 2×2 covariance matrix (C_{emp})

d) Eigenvalues and (normalized) eigenvectors of the covariance matrix (Λ, W)

e) The data set in in the new coordinate system which is given by the principal components (i.e. the eigenvectors) (Y)

Note: Eigenvalues of a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ can be determined by finding the roots ("Nullstellen") of the characteristic polynomial which is given by

$$\chi(\lambda) = \det(\mathrm{A} - \lambda \mathrm{I}) = \det \begin{bmatrix} a - \lambda & b \\ c & d - \lambda \end{bmatrix} = (\mathrm{a} - \lambda)(\mathrm{d} - \lambda) - \mathrm{bc} = \lambda^2 - (\mathrm{a} + \mathrm{d})\lambda + (\mathrm{ad} - \mathrm{bc}).$$

The solutions are $\lambda_{1,2} = \frac{a+d}{2} \pm \sqrt{\frac{(a+d)^2}{4} + bc - ad} = \frac{a+d}{2} \pm \frac{\sqrt{4bc+(a-d)^2}}{2}$.

The eigenvector v_λ corresponding to a eigenvalue λ can be computed by solving the equation $(A - \lambda I)v_\lambda = 0$ for v_λ ($v_\lambda \neq 0$).

On the hand-in date, **16.11.2016**, you must hand-in the following: [1]

a) a text file stating how much time you (all together) used to complete this exercise sheet

b) your solutions / answers / code

for problem 2.1 and 2.2.

[1]upload via StudIP (if there are problems with the upload contact me **beforehand**: krell@uni-bremen.de)