Pattern Recognition
Lab week 1

Useful matlab functions:
`axis square`, `corrcoef`, `load`, `plot`, `hold on`, `hist`, `bar`, `length`, `find`, `rand`, `scatter`, `sprintf`, `sum`, `xor`, `normpdf`, `normcdf`

Guidelines for lab reports:

- Always give a (short) explanation of what you are doing.
- Do not forget to include your Matlab programs, the results of your programs, and an interpretation of these results.
- Put large pieces of Matlab code in an appendix.
- One should be able to understand plots independently, be sure to label axes, add a legend for colors, etc.
- Refer to all plots, tables, code blocks, etc. in your report.
- If your print gray-scale make sure the colors used in the plots are distinguishable.

Assignment 1: *3D distributions, data visualization, pair-wise correlation coefficients* Consider the 24×3 array in the file `lab1_1.mat`. Each row is a three-dimensional (3D) feature vector. The first element of such vector is the height of a person in centimeters; the second element is the age in years; the third element is the body weight in kilograms.

1. Compute the pair-wise correlation coefficients between the features (elements of the vectors). The correlation coefficient of two features is equal to their covariance divided by the square root of the product of their variances, see the file 'statistics reminder'.
2. Create 2D scatter plots of the data points with x- and y-axis being:
    - plot A: the two features for which the correlation coefficient is largest,
    - plot B: the two features for which the correlation coefficient is second largest.

   What conclusions can you draw from these plots?

Assignment 2: *statistical decision theory, decision criterion, confidence interval, false acceptance, false rejection, hypothesis testing, errors of type I and II, normal distribution, missing binary features* Before you start this assignment perform these steps, you do not need to include them in your report:

- Consider the two-dimensional binary arrays in files `person01.mat` to `person20.mat` . Each row of such an array `person[i].m` is a binary feature vector of 30 elements that is extracted from an iris image of a person that we call here person[i] ($i = 1 \dots 20$). Hence, each row is a 30-dimensional binary iris code of that person. There are 20 such iris codes of each person in the corresponding file person[i]; each row of the array is one such binary iris code.
- Take a closer look at the rows of one such array and notice that two rows can differ in only a few positions (bits). Compare now two rows that come from two different files person[i] and person[j]. Notice that two such iris codes differ in about 15 positions.

Discuss the following points in your report:

1. The Hamming distance (HD) of two binary iris codes is the number of positions (bits) in which the two codes (binary feature vectors) differ.

   Compute two sets S and D of 1000 HD values each as follows:

   a) For set S: Choose randomly one of the files person[i].mat, $i = 1 \dots 20$.

      Choose randomly two rows in that file. Compute the HD of these two rows. Normalize the HD by dividing it by 30. (If you use a Matlab function that computes the normalized HD, you need not divide by 30.)

      Repeat this process 1000 times to obtain 1000 HD values.

      Hint 1: Use the function **`sprintf`** to generate the filename for a person. For example **`sprintf`**`('person%02d.mat',3)` gives `'person03.mat'`.

      Hint 2: Create a string array containing strings `'person01.mat'`, `'person02 .mat'` etc, using the function `char`, to be able to load a random file.

   b) For set D: Choose randomly two different files person[i].mat and person[j].mat, $i = 1 \dots 20$; $j = 1 \dots 20$; $i \neq j$. Choose randomly one row from each of these two files. Compute the HD of these two rows. Normalize the HD by dividing it by 30. Repeat this process 1000 times to obtain 1000 HD values.

2. Plot the histograms of S and D in one figure with different colors. Make sure to use bins of the same size for the two histograms and to use an appropriate number of bins. How much do the two histograms overlap?

3. a) Compute the means and the variances of the sets S and D.

   b) Estimate the probability that two bits (in the same position) of the iris codes of two different persons are different.

   c) Estimate the value of the number of statistically independent bits in the iris code.

   See the slide 'Statistics reminder' in the slide file on missing binary features for the relation between the above mentioned quantities.

4. Add to the plots of the previously computed histograms plots of two normal distributions (Gaussian functions) with the above computed means and variances. Find an appropriate way to scale the normal distribution curves so that they fit well the histograms. Explain and justify your way of scaling.

5. The distribution associated with the set S is the class-conditional probability density function that we measure a given HD value for two iris codes of the same person. The distribution associated with the set D is the class-conditional probability density function that we measure a given HD value for two iris codes of two different persons.

   To compute the decision criterion either use Matlab's `normcdf` function and iterate to find the correct value, or use $\sqrt{2}\cdot$**`erfinv`** of the exact confidence interval. The false acceptance rate is the value of the integral of the normal distribution corresponding to the set D for HD $< d$, where d is the value of the decision criterion. False rejection rate is the value of the integral of the normal distribution corresponding to the set S for HD $> d$.

   a) Estimate the value of the decision criterion for which the false acceptance error is 0.0005.

      False acceptance occurs when the iris codes of two different persons are declared to be sufficiently similar so that one can assume that they come from the same person.

   b) For that value of the decision criterion, determine the false rejection rate.

      False rejection occurs when two iris codes of the same person have a HD which is above the decision criterion so that they will wrongly be assumed to come from two different persons. (Note that here the terms acceptance (of an impostor) and rejection (of an authentic person) are related to the alternative hypothesis stating that two iris codes which are compared come from the same person, the zero hypothesis being that they come from two different persons. False acceptance and false rejection thus correspond to an error type I and II, respectively, in terms of statistical decision theory and hypothesis testing.)

6. Consider the iris code given in the file `testperson.mat`. This file contains an iris code of which some bits are missing. These missing bits have the value 2 instead of 0 or 1. To which of the 20 persons whose iris codes are stored in files `person01.mat` to `person20.mat` does this iris code most likely belong to? What is the significance level of your decision?

   Hint: Excluding the bits with a value 2, find the person that has the lowest normalized Hamming distance to the testperson; call this minimum normalized Hamming distance HDmin. The significance level which corresponds to HDmin is defined as the probability that the comparison of the iris code of the testperson with the iris code of a different person will result in a HD such that HD$\leq$HDmin, i.e. the concerned significance level is equal to the integral of the tail of the distribution for HD$\leq$HDmin. To compute its value, make use of the *theoretical* expression for the parameters of the normal distribution of the normalized iris code HD for the given number of *available* (i.e. non-missing) bits.