Useful matlab functions: `load`, `plot`, `hold on`, `hist`, `bar`, `length`, `find`, `rand`, `char`, `sum`, `xor`, `normpdf`

Introduction to Intelligent Systems
Lab week 1

Assignment 0: Get acquainted with Matlab. Explore the mentioned 'useful matlab functions' by typing `help <function>`.

Assignment 1: *1D distributions, empirical data, histogram and probability density function, decision criterion, classification error.*
Given are two sets of measured lengths (in cm) of men (`length_men`) and women (`length_women`) in the file lab1_1.mat.

1. Plot histograms of both sets in one figure.
2. Now choose the decision criterion at 170 cm. How many men are classified incorrect? And how many women?
3. What decision criterion should be used to minimize total number of misclassifications (sum over men and women)?

Assignment 2: *2D distributions, data visualization, decision boundary.*
In the file lab1_2.mat a two dimensional array is given, consisting of measurements of the length (in cm) and the hair length (in cm) of 200 people.

1. Plot the length versus the hair length.
2. The measurements originate from 100 women and 100 men. Given the fact that in general men have shorter hair than women and men are taller, where would you draw the decision boundary (for example use a simple graphics editor to sketch it, or just plot a line between two points on top of your plot in matlab) and why?

Assignment 3: *class conditional probabilities, priors, posterior probabilities, Bayes decision rule, minimum error classification.*
Let's assume we obtained the class conditional probability of salmon and sea bass, describing the distribution of the length of the two classes. These are given by `p_salmon` and `p_seabass` at length `l` (in cm), all in the file lab1_3.mat.

1. Given that sea bass is caught 3 times as often as salmon, calculate the posterior probabilities (don't forget to normalize to 1 for all lengths) and plot them.

2. Now a new fish is measured, which turns out to have a length of 8 cm. According to your posterior probabilities, how would you classify this fish? And what if it's length is 20 cm?

Assignment 4: *statistical decision theory, decision criterion, confidence interval, false acceptance, false rejection, hypothesis testing, errors of type I and II, normal distribution.* **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] ($i = 1 \ldots 20$). 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.**

1. 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.

2. 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, $i = 1 \ldots 20$. Choose randomly two rows in that file. 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. (hint: 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, $i = 1 \ldots 20$; $j = 1 \ldots 20$; $i \neq j$. 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.

3. Plot the histograms of S and D in one figure with different colours. How much do the two histograms overlap?

4. Compute the means and the variances of the sets S and D. Add to the histograms of the previous question (4.3), plots of two normal distributions (Gaussian functions) with these means and variances. How well do the normal distributions fit the histograms?

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. 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. 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 asumed to come from two different persons. (Note that here the terms acceptance (of an imposter) and rejection (of an authentic person) are related to the alternative hypothesis stating that two iris codes which are compared come form 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.)
(hint: Either use Matlab's normcdf function and iterate to find the correct value, or use $\sqrt{2}\times$erfinv of the exact confidence interval (for an explanation on how to use these values, look for 'error function erf' in connection with normal distribution, for example in Wikipedia or Mathworld). 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 $> d$.)