

Introduction to Intelligent Systems 2018

Lab 2

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

Instructions

You should hand in a brief report, in which you provide answers to all the problems, with the same structure of the present document (Assignment 1, Assignment 2, etc.). The code developed to solve the problems has to be included in the report, commented, and opportunely referenced.

The plots have to be self-consistent, with clear title, labels and captions. They have to be sufficiently commented in the text.

The grading will be based on the following aspects (ordered acc. to importance):

- completeness of the report and correctness/plausibility of the results
- readability, layout, formatting of the text
- appearance of the plot, e.g. axis labels, quality of the caption
- language/English issues

A grade equal to 10 can be achieved by solving all the problems correctly and matching the above conditions. The bonus problem 3.5 can be evaluated to compensate for eventual deficits of above conditions.

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

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] ($i = 1, \dots, 20$). 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.

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`, $i = 1, \dots, 20$. 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`, $i = 1, \dots, 20; j = 1, \dots, 20; i \neq j$. 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.
3. Plot the histograms of S and D in one figure with different colors. 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. (BONUS) The distribution associated with the set S is the class-conditional probability density function that we measure a certain 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 certain HD value for two iris codes of two different persons. Estimate the value of the decision criterion for which the false acceptance error is approximately 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 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.)

(*hint: Either use Matlab's `normcdf` function and iterate to find the correct value, or use $\sqrt{p} \times \text{erfinv}$ 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 $HD > d$.)*