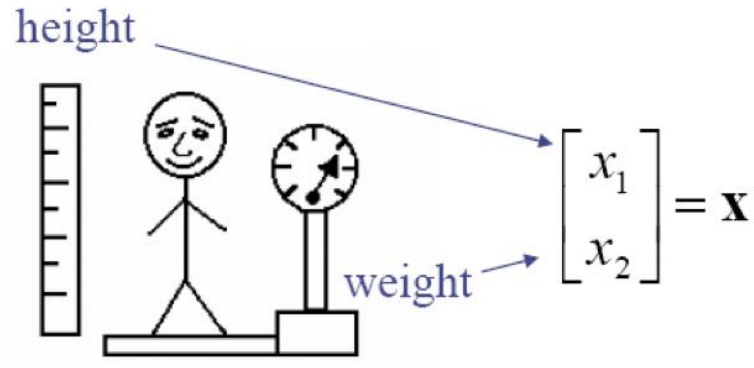
Biotec Lab 1: Pattern Recognition

Introduction

The goal of this exercise is to assess the performance of different pattern recognition techniques applied on biometric data. In particular, the height and weight of a group of people will be used to estimate their gender.



Tools and kick off

Completing the quiz

If you have a GDrive account, you can complete this report online by generating a copy (File > Copy…). If you prefer to complete it locally with your word processor, you can export it to *.odp, .rtf, .doc*… (File > Download as…). You will be asked to deliver a PDF version of it at the end of the session on the dedicated task on Atenea.

Matlab

This lab exercise will require the usage of *Matlab* to load the data and run the provided scripts. You can access the computers in lab D5-004 with your username *biotec11x*, where *x* corresponds to the ID shown on your desk. Ask the professor for your password.
Once you have logged in, copy to your home directory the contents of the *lab1* folder. Finally, launch *Matlab* and set your working directory to your copy of *lab1*.

Features

1. A dataset of labelled observations is provided in the *data.mat* file. Load it by running the following instruction from the Matlab interactive shell:

```
load data.mat
```

2. The provided data contains information about real people who have provided their height, weight and gender. View the contents of the data matrix, which appears in the *Workspace* window within Matlab. What is the name of the matrix containing all data ?

Answer: data

3. Display the provided data on the shall by just typing the name of the matrix. Which information contains each column of the data matrix ?

Answer: data(:,1)=height; data(:,2)=weight; data(:,3)=gender;

4. According to your intuition, which label (0 or 1) corresponds to *females* and which one to *males* ?

Answer: The taller and heavier persons are 0, so 0 corresponds to male and 1 to female.

5. How many observations are included in the dataset ? The size of a matrix can be checked with the *size()* comand.

Answer: 34 observations

Classification with Least Mean Squares (LMS)

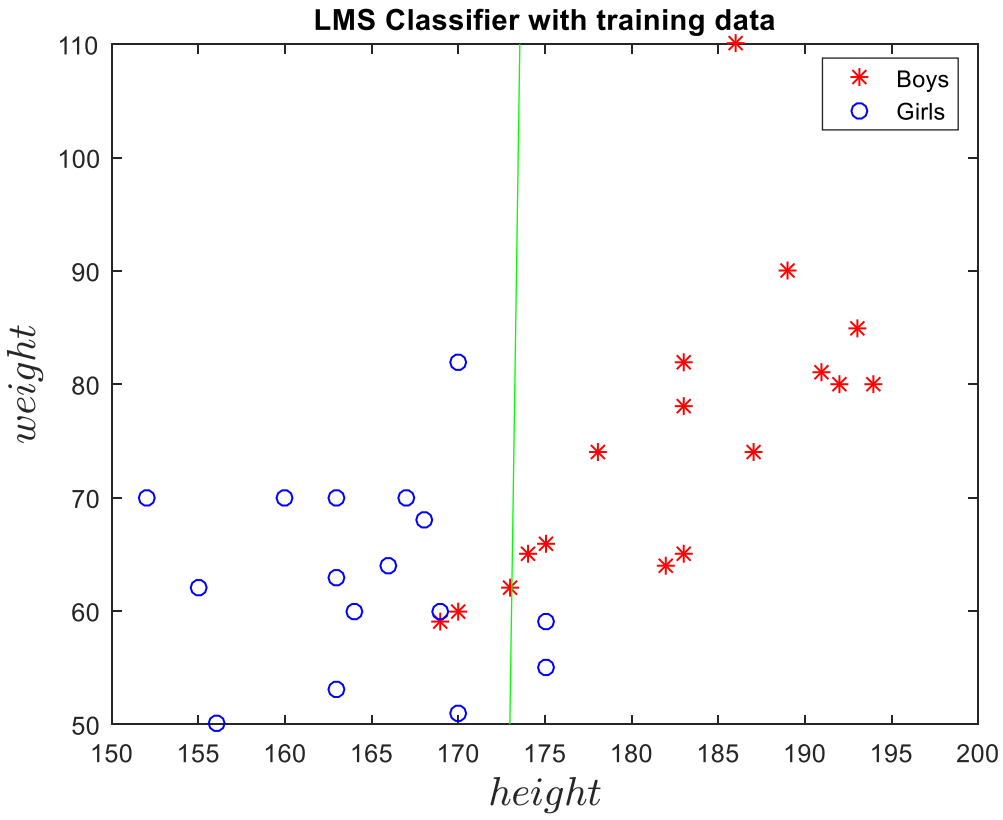
The Least Mean Squares (LMS) classifier defines a linear boundary in the feature space between the different classes.

6. Launch script *lms1.m* by just typing:

```
lms1
```

7. The plotted figure shows in green the boundary that it is estimated by considering all the observations included in the dataset. The observations are also plotted in a 2D feature space. Save the figure that is plotted and add it below:

Answer:



8. Compare the values plotted in the figure with your intuition about the interpretation of the gender labels. Was you intuition correct ?

Answer: Yes it was.

9. In addition to the figure, the Command Window provides the ratios of false boys and girls identifications, related to two popular measures in biometrics. What are these measures ? What is their definition ?

Answer: Rate of false acceptance (FAR), here it is 0.17, which means that 17% of the boys are defined as girls according to the classifier. The other one is Rate of false rejects (FRR), here it is 0.13, which means that 13% of the girls are defined as boys according to the classifier. FAR=false_girls/total_girls and FRR=false_boys/total_boys.

10. In biometric applications, these metrics are referred to the cases of acceptance or denial. However, in this exercise the application is different as you are classifying between girls and boys. According to the results provided in the command window, is this classifier a detector of “girls” or of “boys” ? Or, in other words, which of the two classes (girls or boys) corresponds to the “acceptance” case ?

Answer: Girls corresponds to the ”acceptance” since FAR is 0.17, which is almost 3/17 (rounded downwards), and that it can be seen from the plot that 3 boys are on the wrong side of the threshold.

11. The example provided in *lms1.m* uses all data to train the linear classifier, and the resulting classifier is assessed in the same training dataset. Do you consider this is a valid scientific experiment ? Why ?

Answer: No, because we will have over-fitting and also if we evaluate with the same dataset it will give a wrong impression of the result of how good the classifier is.

12. The simulations contained in *lms2.m* and *lms3.m* have split the dataset in two parts: one for training and a second one for test. Run them both. Are their results equal ? Why ?

Answer: The difference between lms2.m and lms3.m is that lms2.m uses the first half of the dataset as training and the second half as test while lms3.m uses the second half of the dataset for training and the first half for test. This gives of course different results, so no.

13. [Optional-Medium] Create a new script *lms4.m* that, instead of splitting the dataset in two sequential halves, it select the samples in each partition randomly (with no repetitions). Hint: Matlab already provides some functions that may help you.

Answer: (See code lms4.m) The FAR and FRR changes for every run since it is random.

14. [Optional-Avanced] Create a new script *lms5.m* that will apply a cross validation approach where a random splits of the dataset is to be applied 5 times. The results obtained in each iteration will be finally averaged to obtain more confident FAR and FRR values.

15. [Optional-Advanced] Build your a new linear classifier named *orthogonal.m* by following these steps:
    - Estimate the average weight and heights for both boys and girls.
    - Plot this mean value and connect it with a dashed line.
    - Estimate the linear parameters a and b (y=ax + b) of the ortogonal line. Use it as a classifier.
    - Plot this new line in solid green.
    - Evaluate the new classifier, at least, with the data partitions used in *lms2.m* and *lms3.m*.

Answer: FAR=0.00 and FRR=0.17 for the partition of the dataset as in lms2.m and FAR=0.44 and FRR=0.00 for the partition of the dataset as in lms3.m. Which is an improvement compared to lms1.m, lms2.m and lms3.m.

16. [Optional-Medium] Build a new classifier *onenn.m* that will assign each validation sample to the same label of its closest training sample. This is the *1-Nearest Neighbour (1-NN)* classifier. Evaluate the new classifier, at least, with the data partitions used in *lms2.m* and *lms3.m*.

Answer: For lms2.m FAR=0.00 and FRR=0.29. For lms3.m FAR=0.50 and FRR=0.07. The conclusion from this is that depending on the partition of the dataset, there will be very different results. A good classifier should just have small deviations (or none) if the partition of the dataset changed.

17. [Optional-Advanced] Improve the 1-Nearest Neighbour Classifier by considering the K-Nearest Neighbours (k-NN). Save it into *knn.m*. Plot the FAR and FRR values for different K values.

Scoring Rubric

The final grading of this exercise will follow this chart:

A	Excellent	Two advanced questions or more
B	Very good	One advanced question or two medium ones
C	Good	One medium question
D	Pass	No optional questions at all
F	Fail	Basics not completed

Notice: Additional penalties may apply if the working attitude in lab or the delivery of submissions does not exactly meet the provided instrucions.