Exercise 2: Our first classifier.

I. Goal of the exercise

In this exercise you will practice the basic pipeline of the supervised learning task. Implement a simple classifier. And will try to solve several hinderances found in the process.

II. Deliverables

As you progress in this exercise, you will find several questions you are expected to answer them properly with adequate figures when required and deliver a document with all these evidences in due time. A file or files with the working code used for generating and discussing the results must be also delivered.

III. Our fist classifier.

We are given the data in `diabetes.mat` and our goal is to predict the whether a person suffers from diabetes or not given her medical record. Our first model to try is linear regression as explained in "A gentle introduction to supervised learning".

A. Understanding and preprocessing our problem.

The first step in the learning pipeline is to have a general picture of your dataset particularities.

B. Data set analysis

Load the dataset and describe the basic properties of the data,

Question block 1:

1) Which is the cardinality (number of examples) of the training set?
2) Which is the dimensionality of the training set?
3) Which is the mean value of the training set?

As you can see there are some missing values with value `NaN` and some categorical data.

Question block 2:

1) Create a new dataset $\mathcal{D}_1$, replacing the `NaN` values with the mean value of the corresponding attribute without considering the missing values.
2) Create a new dataset $\mathcal{D}_2$, replacing the `NaN` values with the mean value of the corresponding attribute without considering the missing values conditioned to the class they belong, i.e. replace the missing attribute values of class $+1$ with the mean of that attribute of the examples of class $+1$, and the same for the other class.
3) **[Optional :]** Explain another method to deal with missing values and apply it to preprocess the training data. Include the reference of the method used. Consider this new dataset as $\mathcal{D}_3$.
4) Which are the new mean values of each dataset?

C. A simple classifier

Our first classifier is a thresholded regressor. Use and/or modify any of the methods you implemented for regression and apply it to find a linear classifier.

Question block 3:

1) In this model you have to learn the threshold value. Explain how you can accommodate this parameter.
2) Report the normal vector of the separating hyperplane for each data set $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$.
3) Compute the error rates achieved on the training data. Are there significant differences? Report the method used and their parameters.

Training error is a poor estimation of the generalization error. Let us test what happens in a test set created by holding-out a certain percentage of the original dataset.

Question block 4:

1) Repeat the learning process in block 3 using just $\mathcal{D}_2$ but **holding-out** the last fifth of the data set for testing purposes, i.e. use the first $4/5$-th for training and the last $1/5$-th for testing. Follow *exactly* the following steps in your process:
   a) Clear your workspace and close all figures: `clear all, close all, clc`
   b) Preprocess the data replacing the `NaN` using the method for creating $\mathcal{D}_2$.
   c) Split your data in two sets: the first $4/5$-th is to be used for training and the last $1/5$-th will be used for testing purposes.
   d) Train your model on the training set.
   e) Answer the following questions: Which is the error rate on your training data? Which is the error rate on your test data? Are they similar? Did you expect that behavior? Why?

Question block 5:

1) Repeat the process in block 4 changing the order of some of the steps. Follow *exactly* the following steps in your process:
   a) Clear your workspace and close all figures: `clear all, close all, clc`
   b) Split your data in two sets: the first $4/5$-th is to be used for training and the last $1/5$-th will be used for testing purposes.
   c) Preprocess the data replacing the `NaN` using the method for creating $\mathcal{D}_2$. *But this time use only the data corresponding to the training set*.
   d) Train your model on the training set.
   e) *Replace the* `NaN` *values using the means computed on the training data*
   f) Answer the following questions: Which is the error rate on your training data? Which is the error rate on your test data? Are they similar? Did you expect that behavior? Why?
   g) *Compare these results with the ones in block 4. Do we achieve better or worse results? Why?*

Question block 6:

1) Repeat the process in block 5 changing the percentage of the data for training and testing. Plot a graph with the training and test error rates for each splitting percentage point. Comment the results.
2) Add to the plot the upper bound on the generalization error using the equation of the slides for VC dimension equal to $d + 1$. Discuss the result.
3) How many samples does the bound predict in order to have 1% error deviation with a confidence of 95%? And with confidence 50%? What about 5% and 10% error deviation with 95% confidence? Comment the behavior according to your observations.