Assignment 01

To be solved INDIVIDUALLY or in GROUP of at most two elements

Submit by October 18, 2019, 23h59 by email to jaime.cardoso@fe.up.pt

1. Write a Python/Matlab function to compute the predictions according to the mean Euclidean distance to the sample points of each class normalized by the corresponding standard deviation.
The function should have the following interface function `[prediction] = meanNormalizedPrediction(dataClass1, dataClass2, dataUnknownClass)` where `dataClass1` is an array N_1xd; `dataClass2` is an array N_2xd; `dataUnknownClass` is an array $N_t$xd; and `prediction` is an array N_tx1. *d* is the dimension of the features.

Consider the data in the 'heightWeightData.txt' file. The first column is the class label (1=male, 2=female), the second column is height, the third weight.

a) Determine the training error on your sample using only the height feature value. Make use of the function `meanNormalizedPrediction` you wrote.
b) Repeat a) but now use the two features, height and weight.
c) Is it ever possible for a finite set of data that the training error increases when we add more and more attributes?

2. Consider the probabilistic model for linear regression as discussed in the class (see slide 11). Assume the model without the bias.

For a generic $\mathbf{x}_{test}$ observation, what's the variance in the output variable $\hat{y}$?

What's the variance of $\hat{y}$ in the origin, $\mathbf{x}_{test}=0$?

What's the variance of $\hat{y}$ when the norm of $\mathbf{x}_{test}$ is very large?