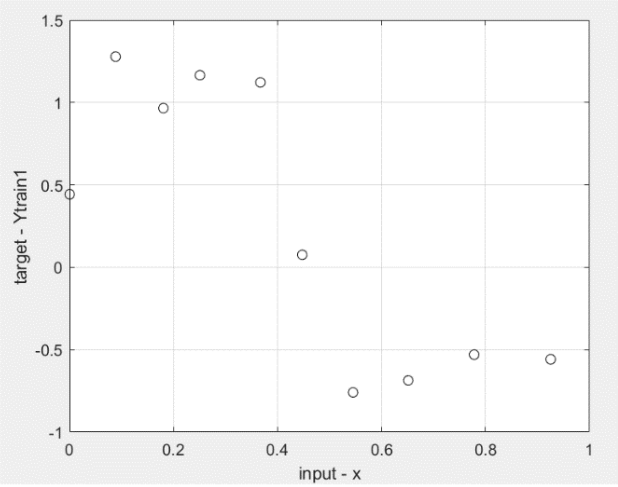
Module-Lab 5: Probabilistic Machine Learning

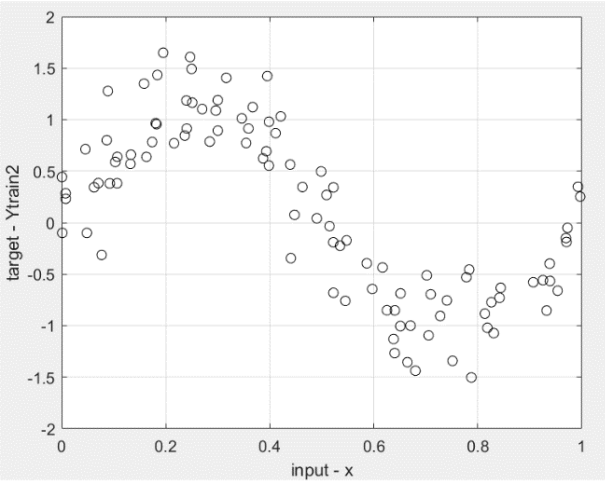
Exercise 1 This Exercise is part of the Lab/Practical session of the ACP module. Please, access the materials through: UCStudent Aprendizagem Computacional Probabilística -> Práticas e Laboratórios -> "Materiais -> Material Apoio", where you will find the files: ***Lab5.pdf, DatasetLab5Training1.txt, DatasetLab5Training2.txt, DatasetLab5Test.txt***

This exercise is related to "non-linear" regression, in particular polynomial fitting, and the goal is to examine the behaviour of simple models – in terms of overfitting, generalization, and capacity - as the size (N) of the training set increases. The models are trained for increasing order (given by M).

Hint: students are advised to see/revisit the Lab4´s exercises and the module#4 slides (*ie, Lec4_Regression.pdf*).

a) Open/load the files DatasetLab5**Training1**.txt and DatasetLab5**Training2**.txt using Matlab or Python IDE. Then, generate a graph of both training points (eg, **plot(x,y)** in Matlab), as shown in the figure below





b) Calculate the number of points ie, the size of, the **training** and the **Test** sets. The latter is comprised in "*DatasetLab5**Test**.txt*". The datasets were generated from the function $\mathrm{y(x) = sin(2*pi.*x) + 0.1*cos(10*pi.*x)}$ with additive noise.

Plot the function $\mathrm{y(x)}$ in the same graph of the **Training1** set.

c) Using `polyfit` (in MATLAB), or something equivalent, obtain the parameters **w** (also called coefficients) of the polynomial models having orders M= **0, 1, 2, …8, 9** for the two training sets: **Training1** and **Training2** sets ie, we will obtain 10 models per training set.

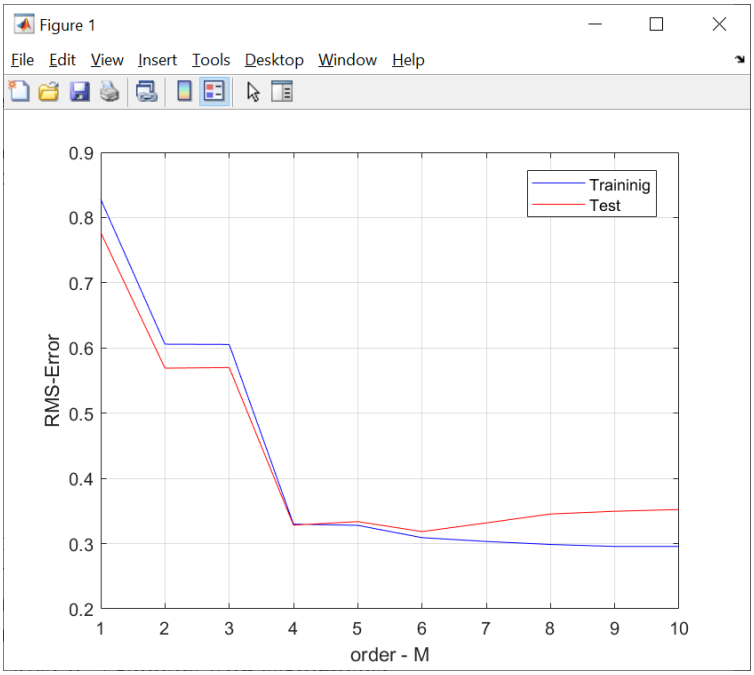
d) Now, make a code to compute the RMS error on the first training set and, using the trained models obtained from **Training1**, compute the RMS error on the Test set. Repeat the work for the **Training2** set and compute the corresponding RMS error on the test set as well. Models' performance, measured in terms of $\mathrm{RMS_{error}}$, can then be compared.

e) Finally, generate two graphs of the root-mean-square error:

1) the first plot is the RMS_{error} evaluated on the training set **Training1** and also on the test set as function of the models' order (ie, M= **0, 1, …9**)

2) the 2nd plot is the RMS_{error} evaluated on the training set **Training2** and on the test set.

Figure below shows an example of one of the two plots we wish



These experiments allow us to see that, for a given model **complexity** (ie, depending on the order M), the over-fitting problem become less severe as N (the size of the training set) increases.

Exercise 2 Visit the UCI repository (https://archive.ics.uci.edu/ml/datasets.php) and choose one or more dataset related to **Regression** and then explore such dataset/s in order to – hopefully - obtain a solution to the related task using simple regression techniques like the ones we have studied here.

Hint: give preference to "small" datasets (not having many instances/examples), comprising numeric variables/attributes.