Data analysis and data Mining
Report 1

Mauro Angelini
Alessio Gilardi

October 26, 2018

Contents

First exercise **2**

Second exercise **4**

First exercise

In this first exercise we implemented the regression of a function in one-dimensional case, for example: the quadratic function

$$y = x^2, \ \forall x \in [0, 1]$$

Vengono definite:

- Function to be rebuilt
- The number of samples
- The samples disturbed by the Gaussian additive noise
- The variance of the noise
- The degree of the polynomial regressor

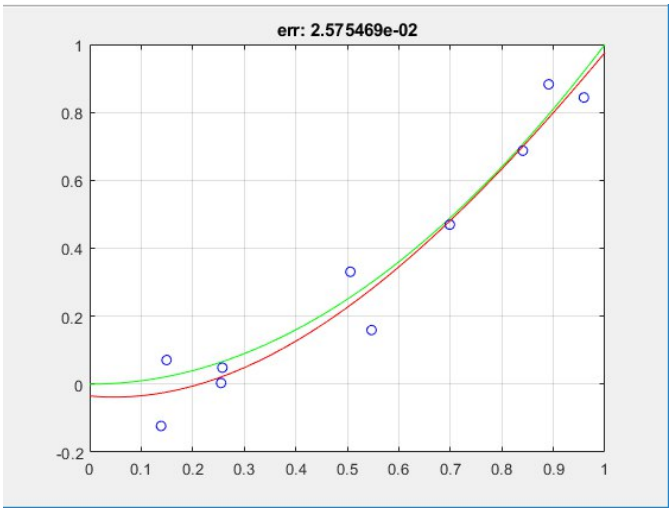
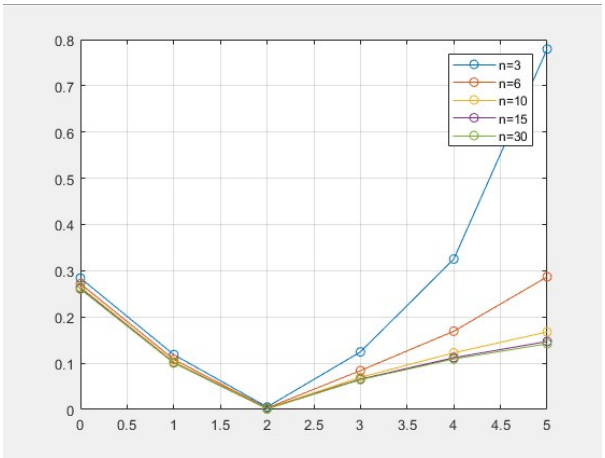


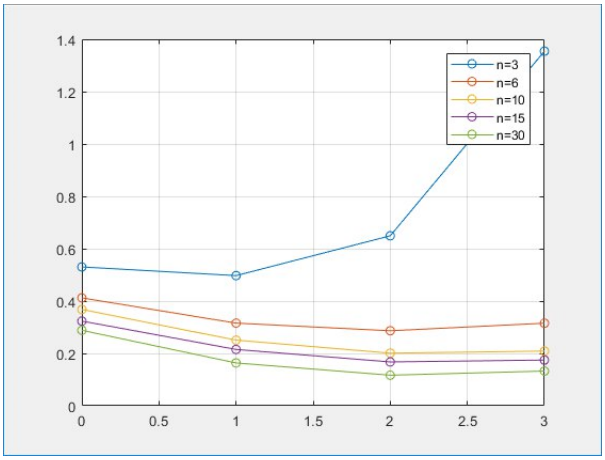
Figure 1: Regression Function

Figure 1 Regression function of a quadratic function

Starting from this first test, we notice that increasing the complexity of the polynomial too much produce a great loss of precision in the reconstruction. It's therefore necessary to evaluate the variation of reconstruction, by changing n (Number of samples), p (Degree of the polynomial regressor) and σ (Noise variance). So the error is calculated between YT (True function) and YP (Prediction function) and plotted in the graph.
Now we notice that in order to make the measurement of error more precise, it's better to repeat the experiment a sufficiently high number of times, for example, we make 30 repetitions of the prediction.



(a) Error by varying n and p (from 0 to 5)



(b) Error by varying n and p (from 0 to 3)

Figure 2: Comparing error with different values of p

	p = 0	p = 1	p = 2	p = 3	p = 4	p = 5
n = 3	2.842e-01	1.186e-01	5.349e-03	1.245e-01	3.253e-01	7.796e-01
n = 6	2.719e-01	1.084e-01	3.093e-03	8.410e-02	1.693e-01	2.869e-01
n = 10	2.647e-01	1.030e-01	2.056e-03	6.957e-02	1.225e-01	1.677e-01
n = 15	2.622e-01	1.012e-01	1.514e-03	6.572e-02	1.120e-01	1.471e-01
n = 30	2.603e-01	1.004e-01	1.095e-03	6.449e-02	1.092e-01	1.421e-01

(a) Error by varying n and p (from 0 to 5)

	p = 0	p = 1	p = 2	p = 3
n = 3	5.303e-01	4.972e-01	6.493e-01	1.354e+00
n = 6	4.121e-01	3.157e-01	2.865e-01	3.154e-01
n = 10	3.678e-01	2.507e-01	2.010e-01	2.089e-01
n = 15	3.235e-01	2.152e-01	1.674e-01	1.746e-01
n = 30	2.882e-01	1.636e-01	1.168e-01	1.325e-01

(b) Error by varying n and p (from 0 to 3)

Figure 3: Comparing error with different values of p

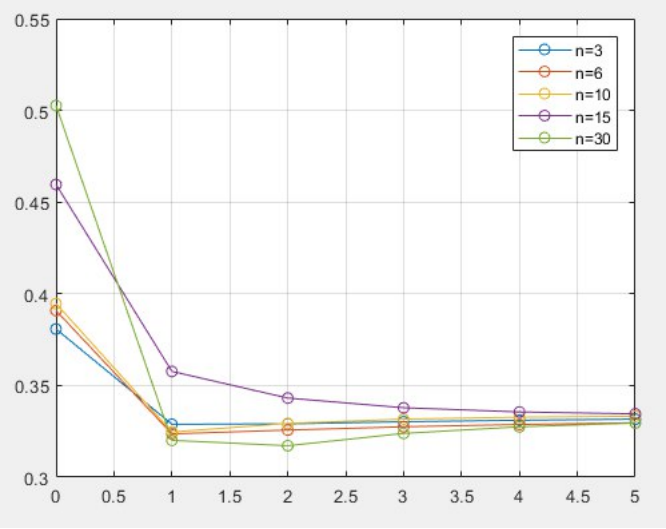
We performed tests with some values of n and p, from which it becomes clear how the error always maintains the same trend. In particular, observing the error plots, it's evident that the maximum degree of precision is obtained for p values around the degree of the original function. Obviously, using a second degree polynomial the error is minimal since the function to be reconstructed is a parabola. Utilizing a large number of samples n, in general, also increases accuracy, while an increase of p results in a loss of precision. It's evident that the variation of the error between one execution of the exercise and the other is minimal for the values greater than n. While using a small number of samples n, it's better to use low-grade polynomials, since an excessive number of degrees of freedom results in an attempt of the regression function to pass for each sample, making the reconstruction less precise, in particular in presence of a lot of noise, as it will try to pass for the samples even if noisy.

Second exercise

Ripeto la regressione polinomiale dell'esercizio precedente con l'aggiunta di regolarizzatore (bias).

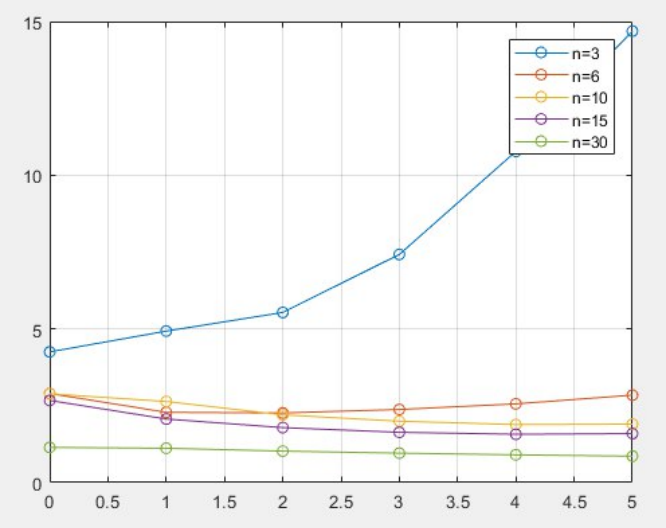
Con l'uso del *bias* λ cerco di regolarizzare il mio polinomio: in sostanza aumetando lambda semplifico il mio risultato mentre dimunendolo utilizzo una funzione più complessa per la regressione, quindi modificare λ è concettualmente simile a cambiare il grado del polinomio, offrendo, però, una maggior granularità nella scelta del valore ottimale. λ rappresenta il livello di fiducia nella qualità dei dati in ingresso: più l'ingresso è rumoroso più conviene utilizzare λ grandi e quindi adottare soluzioni semplici in modo to not fit the noise, meno è rumoroso e più possiamo diminuire λ ed utilizzare soluzioni più comlplesse in order to fit data.

In this exercise it's repeated the polynomial regression of the previous one with the addition of a regularizer *bias*. With the use of *bias* λ we try to regularize our polynomial: so by stretching λ we simplify my result while using it to use a more complex function for regression, so modifying λ is conceptually similar to changing the degree of polynomial, offering, however, a greater granularity in the choice of the optimal value. λ represents the level of confidence in the quality of input data: the more noisy the input, the better it's to use λ large and then adopt simple solutions to not fit the noise, the less it is noisy and the more we can decrease λ and use more complex solutions in order to fit data.



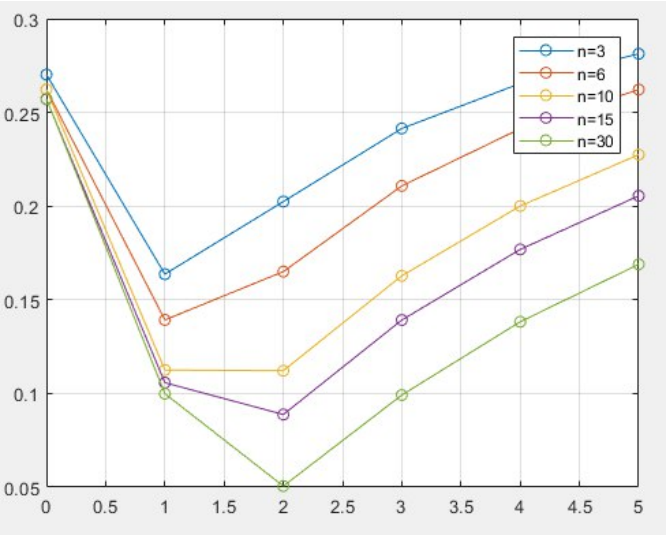
	p = 0	p = 1	p = 2	p = 3	p = 4	p = 5
n = 3	3.808e-01	3.289e-01	3.291e-01	3.302e-01	3.311e-01	3.316e-01
n = 6	3.908e-01	3.236e-01	3.257e-01	3.275e-01	3.287e-01	3.297e-01
n = 10	3.946e-01	3.247e-01	3.293e-01	3.317e-01	3.327e-01	3.332e-01
n = 15	4.596e-01	3.577e-01	3.432e-01	3.379e-01	3.356e-01	3.345e-01
n = 30	5.026e-01	3.201e-01	3.172e-01	3.239e-01	3.275e-01	3.295e-01

Figure 4: $\lambda = 100$ $\sigma = 10$



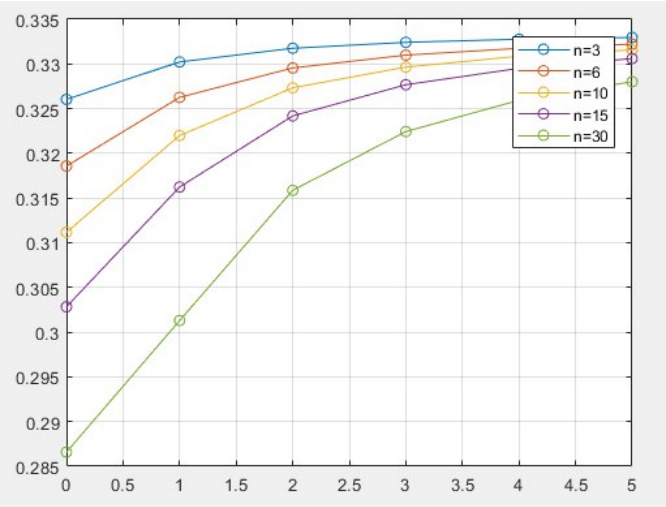
	p = 0	p = 1	p = 2	p = 3	p = 4	p = 5
n = 3	4.252e+00	4.928e+00	5.535e+00	7.421e+00	1.077e+01	1.469e+01
n = 6	2.899e+00	2.286e+00	2.266e+00	2.379e+00	2.562e+00	2.845e+00
n = 10	2.889e+00	2.638e+00	2.204e+00	1.997e+00	1.888e+00	1.902e+00
n = 15	2.672e+00	2.068e+00	1.789e+00	1.636e+00	1.572e+00	1.593e+00
n = 30	1.145e+00	1.117e+00	1.024e+00	9.576e-01	9.037e-01	8.603e-01

Figure 5: $\lambda = 0{,}001$ $\sigma = 10$



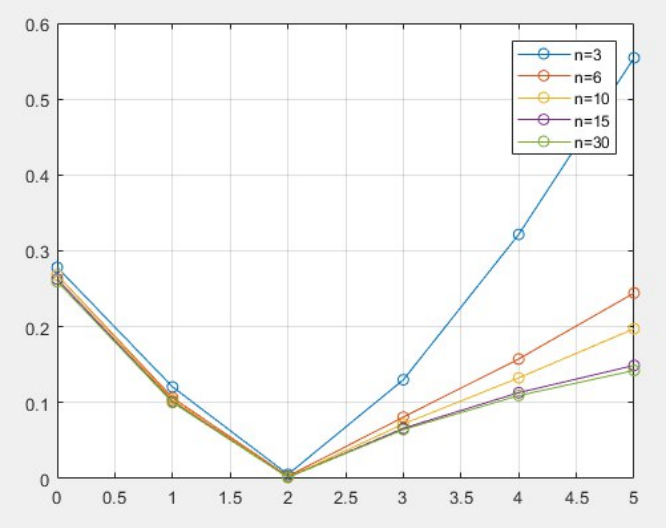
	p = 0	p = 1	p = 2	p = 3	p = 4	p = 5
n = 3	2.702e-01	1.637e-01	2.024e-01	2.414e-01	2.655e-01	2.813e-01
n = 6	2.623e-01	1.393e-01	1.651e-01	2.108e-01	2.413e-01	2.622e-01
n = 10	2.623e-01	1.125e-01	1.122e-01	1.628e-01	2.001e-01	2.275e-01
n = 15	2.571e-01	1.058e-01	8.881e-02	1.392e-01	1.769e-01	2.055e-01
n = 30	2.573e-01	9.990e-02	5.058e-02	9.918e-02	1.383e-01	1.689e-01

Figure 6: $\lambda = 1$ $\sigma = 0{,}01$



	p = 0	p = 1	p = 2	p = 3	p = 4	p = 5
n = 3	3.260e-01	3.302e-01	3.318e-01	3.324e-01	3.328e-01	3.330e-01
n = 6	3.186e-01	3.263e-01	3.296e-01	3.310e-01	3.317e-01	3.322e-01
n = 10	3.112e-01	3.220e-01	3.273e-01	3.296e-01	3.309e-01	3.316e-01
n = 15	3.028e-01	3.162e-01	3.242e-01	3.277e-01	3.295e-01	3.306e-01
n = 30	2.866e-01	3.013e-01	3.159e-01	3.224e-01	3.259e-01	3.280e-01

Figure 7: $\lambda = 100$ $\sigma = 0{,}01$



	p = 0	p = 1	p = 2	p = 3	p = 4	p = 5
n = 3	2.783e-01	1.207e-01	5.476e-03	1.301e-01	3.217e-01	5.545e-01
n = 6	2.682e-01	1.064e-01	3.047e-03	8.095e-02	1.576e-01	2.444e-01
n = 10	2.681e-01	1.036e-01	1.560e-03	7.190e-02	1.326e-01	1.974e-01
n = 15	2.628e-01	1.013e-01	1.481e-03	6.620e-02	1.131e-01	1.491e-01
n = 30	2.596e-01	1.002e-01	1.418e-03	6.451e-02	1.093e-01	1.424e-01

Figure 8: $\lambda = 0{,}00001$ $\sigma = 0{,}01$

List of Figures

1 Regression Function 2

2 Comparing error with different values of p 3

3 Comparing error with different values of p 3

4 $\lambda = 100$ $\sigma = 10$ 4

5 $\lambda = 0{,}001$ $\sigma = 10$ 5

6 $\lambda = 1$ $\sigma = 0{,}01$ 5

7 $\lambda = 100$ $\sigma = 0{,}01$ 6

8 $\lambda = 0{,}00001$ $\sigma = 0{,}01$ 6