Exercise 1: A basic supervised problem

Introduction to Machine Learning

University of Barcelona

November 16, 2015

Authors:
Camps, Julià
Serra, Xavier

1 Question Block 1

1. *Plot the training samples and their corresponding label.*

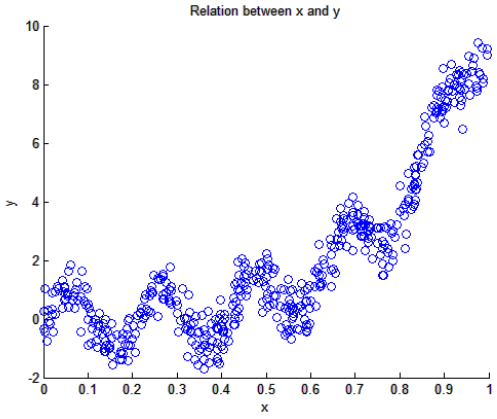

Figure 1: Relation between X and Y

2 Question Block 2

1. *Which is the optimal value of the linear regression weights?*

   First we expand the data for the regression algorithm introducing the ones extra columns to the X attributes matrix: a ones column for the offset value, in order to get a weight w_0 for x^0. So the approximation that we want to find will be a polynomial function of degree 1, like the following one: $f(x) = w_1 x^1 + w_0 x^0$. Using the analytical regression algorithm:

   $$W = (XX^T)^{-1}XY$$

   We obtained that the optimal values for the weights using analytical linear regression are:

   - $w_0 = -2.0474$
   - $w_1 = 7.9948$

2. *Plot the data set and the line learnt by the model. Does it looks like a good linear approximation?*

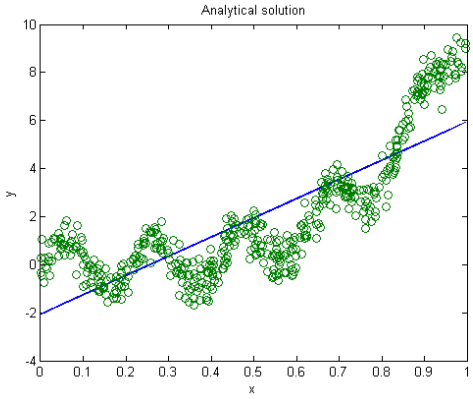

Figure 2: Figure containing a plot of the data distribution and the analytic solution for linear regression.

In figure 2 we can appreciate that the linear regression is explaining the main direction of the data distribution, so it's a good linear approximation. However, lots of information are not being explained, as we can appreciate that the data distribution requires clearly a high level polynomial. The linear regression simply insinuates the behaviour.

3 Question Block 3

1. *Which is the optimal value of the linear regression weights using the descent method?*

   - $w_0 = -2.0474$
   - $w_1 = 7.9948$

2. *Which are the parameters of the descent method used to obtain the optimal value?*

   - Learning rate = 0.1
   - Maximum number of iterations = 1000
   - Error threshold to stop iterations = 10^{-5}

3. *Plot the convergence curve of the method.*

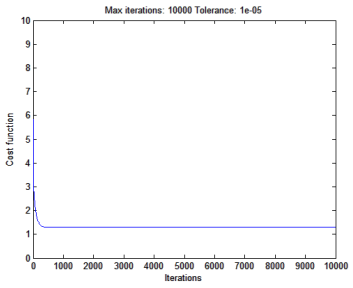

Figure 3: Fitting of the data

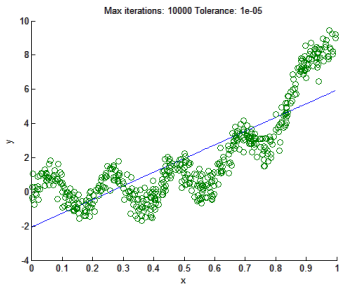

Figure 4: Convergence curve

4. *Zoom in the flat convergence part. Does it oscillate? Why?*

   No, it does not. Actually, it is a result quite surprising for us. The reason is that, with a learning rate quite high, it is surprising to find an strange local optima. However, by some coincidence, it may occur, so we accept the result.

5. *Change the learning rate to 0.1. Plot the convergence curve of the method.*

   Instead of doing what was asked, as we had already used the 0.1 learning rate as the best one, we have decided to plot here a learning rate of 0.001:

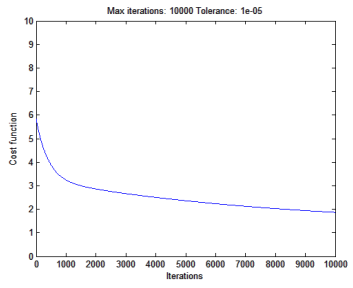

Figure 5: Fitting of the data

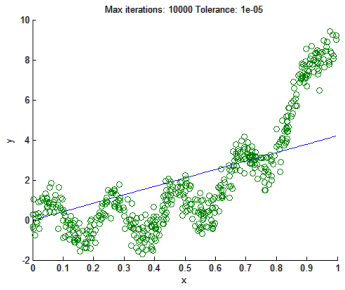

Figure 6: Convergence curve

4 Question Block 4

Using a new descent algorithm:

$$\triangle x = -\frac{\triangledown f(x)}{\| \triangledown f(x)\|_2}$$

1. *Which is the optimal value of the linear regression weights using the modified descent method?*

   - $w_0 = -2.0472$
   - $w_1 = 7.9949$

2. *Which are the parameters of the descent method used to obtain the optimal value?*

   - Learning rate = 0.01
   - Maximum number of iterations = 1000
   - Error threshold to stop iterations = 10^{-5}

3. *Plot the convergence curve of the method.*

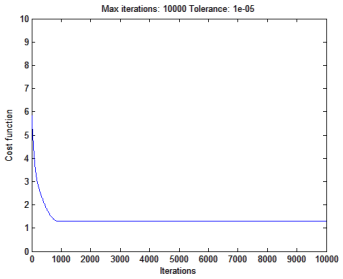

Figure 7: Fitting of the data

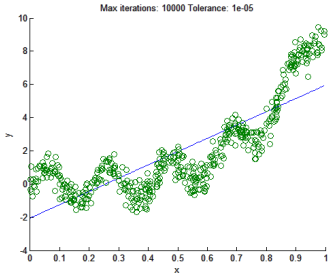

Figure 8: Convergence curve

4. *Zoom in the flat convergence part. Does it oscillate? Why?*

   Yes, it does. Here, we met once again with a surprising result. The new gradient is normalized, so we could expect that the speed of descending would regulate itself the closer it got to the local optima. However, according to the results, it seems that it kept oscillating without actually reaching it. Further research should be done in order to determine the exact reason.

5 Question Block 5

1. *Transform the training set into the set with examples described by z considering $p = 3$. Apply, the analytic solution code (if properly coded it should work without modifications). Which is the optimal value of the weights?*

   First we expand the data for the regression algorithm introducing extra columns to the X attributes matrix:

   - The previously added ones column to simulate the offset values, so that we have a weight w_0 for x^0.
   - Two extra columns containing x^2 and x^3 to simulate two new attributes

   Therefore, the approximation that we want to find will be a polynomial function of order 3, like the following one: $f(x) = w_3x^3 + w_2x^2 + w_1x^1 + w_0x^0$. Using the analytical regression algorithm:

   $$W = (XX^T)^{-1}XY$$

   We obtained that the optimal values for the weights using analytical linear regression are:

   - $w_0 = 0.3901$
   - $w_1 = -2.1077$
   - $w_2 = -2.1699$
   - $w_3 = 13.6737$

2. *Plot the data set and the curve just found. Does it fit better the data? Why?*

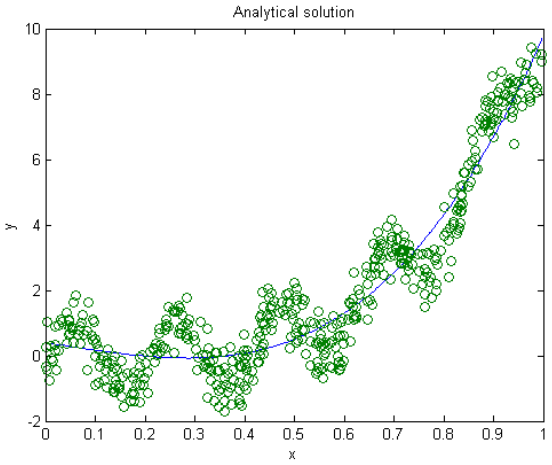

Figure 9: Figure containing a plot of the data distribution and the analytic solution for linear regression.

In figure 9 we can appreciate that using two extra attributes produce a much better approximation than before. But although it fits much more accurately the main direction flow of the data distribution, it's not explaining the detailed distribution of the data.

6 Question Block 7

The best fit is given by the polynomial displaying smaller RMS (root mean square) value.

$$RMS = \sqrt{\sum_{i=1}^{n}(f(x_i; w) - y_i)^2}$$

1. *Use the first half of the data set for training and the second half for validation.*

2. *Optimise the models (you can use any of the methods implemented before) and plot the validation set and the 6 models plots.*

3. *Plot the training set and the 6 models plots.*

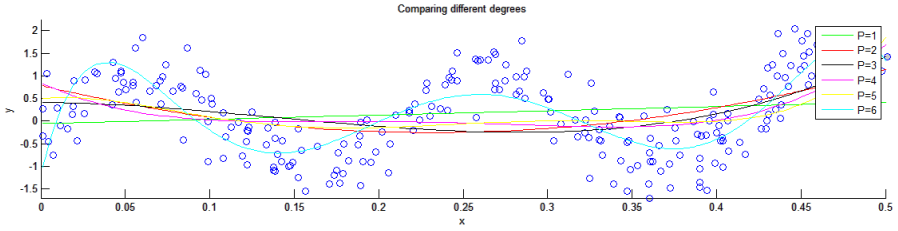

Figure 10: Fitting of the data of the 6 models

4. *Compute the RMS error on the training set and on the validation set. Plot both error curves and describe their behaviour.*

   Those exponents without RMS of the test have actually a too high value to appear in the graph. In order to better show the tendency, we have provided two graphs at a different scale, but even by doing so, some of the values are still out of range

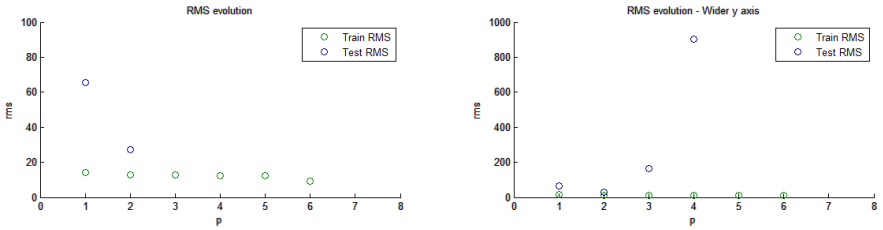

Figure 11: RMS of the 6 models

5. *Does the selected model agrees with the model that performs the best on the training set? Which one do you think is the optimal choice? Why? (You are not supposed to give an exhaustive answer to this question but your impressions and ideas. We will see the reasons why this effect happens shortly.)*

   No, it does not. The best option is thought to be a polynomial of degree 2, as it has the better validation RMS value. Even though there are other degrees with a better training RMS value, their difference with the value of degree 2 is negligible, and even more if we compare it with the difference between test RMS values.

   The reason for this is thought to be that the higher degrees overfit the data, as they try to obtain the best possible model according to the training data, and when faced with new data they do not work so well.