Non-Parametric Bases with Uneven Data

December 24, 2016

Assumptions for successful RBF interpolation

Gaussian RBF are particularly well suited for datasets that have:

1. constant oscillation period
2. evenly sampled training data across it's domain

Problem if violated assumptions

The problem if assumption 1 is violated is that there would be no good σ to approximate the data. Indeed the σ tells us what variance to take (i.e. what is the importance of the distance to a point). So if the data is periodic but first with a very small periodicity and then a very large one, there would be no good σ. This is exemplified by Figure 1 where we see that the best σ (i.e 0.25) is too small for the left part. Figure 2 shows us that if σ is large enough for the left part then it would be too large for the right part. In this precise case one way to go around the problem would be to visualize the data first (always a good idea) and to use 2 different sigmas depending on where you are.

The problem if assumption 2 is violated is that you could have most of your training data in one part of the graph and only few in the other one. In that case the prediction of data between training points far away would be very bad as in Figure 3. One way around that problem would be to regularize it (on order not to have big derivatives). An other way around it would be to visualize it take a bigger σ as in Figure 4. But the best way, if possible, is still (as in exercise 3.1) to re-sample the training data.

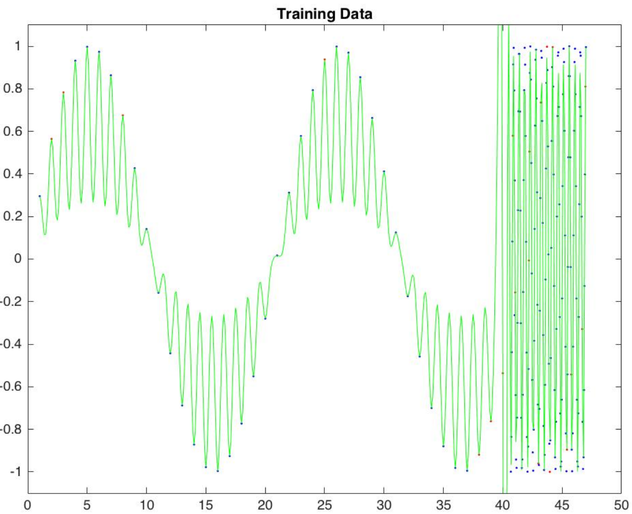

Figure 1: *RBF with data of different periodicity:* $\sigma = 0.25$ *(found by cross validation)*

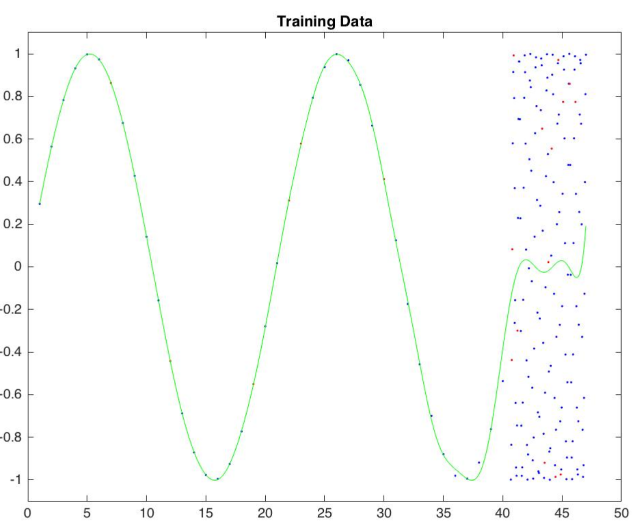

Figure 2: *RBF with data of different periodicity:* $\sigma = 3$

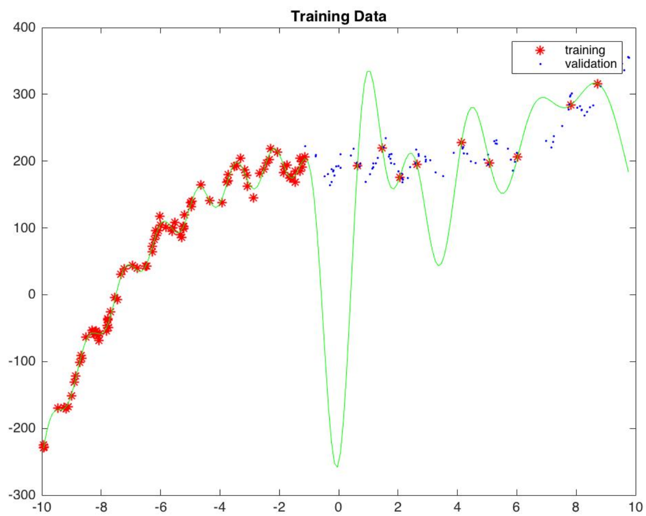

Figure 3: *RBF with training data not evenly sampled* $\sigma = 1$

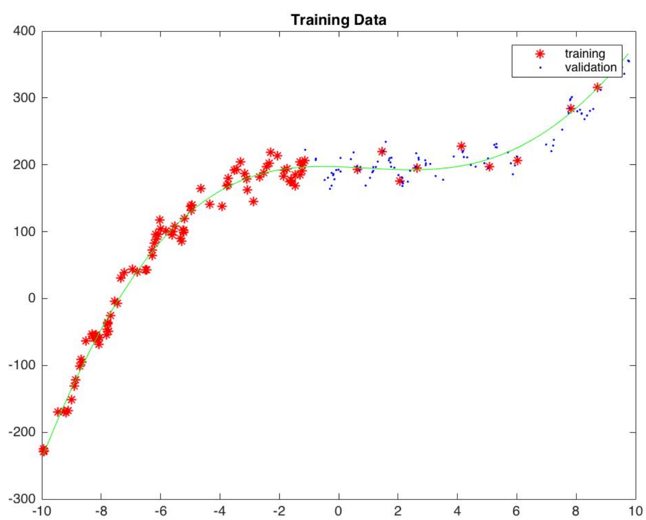

Figure 4: *RBF with training data not evenly sampled* $\sigma = 16$