Kernelized ridge regression

Maruša Oražem[1]

[1] *mo3757@student.uni-lj.si, 63200439*

In this homework, we focused on kernelized ridge regression.

Part 1 - Kernels

We had to implement two different kernels:

1. $\kappa(x,x') = (1+xx')^M$
2. $\kappa(x,x') = e^{-\frac{||x-x'||^2}{2\sigma^2}}$

We have implemented vectorizable versions of kernels, without any looping. After that, we have applied kernelized ridge regression to the *sine.csv* data set. To get kernelized ridge regression, we applied kernel trick and what we get is the following:

$$\hat{y}(x') = \kappa(x')(K+\lambda I_n)^{-1}y,$$

where

$$\kappa(x') = \begin{bmatrix} \kappa(x',x_1) \\ \vdots \\ \kappa(x',x_n) \end{bmatrix}^T, \quad K = \begin{bmatrix} \kappa(x_1,x_1) & \dots & \kappa(x_1,x_n) \\ \vdots & \ddots & \vdots \\ \kappa(x_n,x_1) & \dots & \kappa(x_n,x_n) \end{bmatrix}.$$

Using a kernel, we can perform ridge regression in the space whose inner product is represented by the kernel. After applying this to the *sine.csv* data set, we get the results shown on figure 1. Results of course vary for different set of parameters. If we choose different parameters we get flattened curve (smaller M and bigger σ, seen on figure 2) and overfit (bigger M, smaller σ, seen on figure 3).

Part 2 - Housing data set

For this part of homework, we had to apply two kernels from previous section to the *housing2r.csv* data set. We have split the data to train set - first 80%, and test set - last 20% of the data. For each kernel we have ploted the RMSE on the testing set versus a kernel parameter value. For polynomial kernel we have choose $M \in [1,10]$ and for RBF we have choose σ from 0.1 to 20 with a step of 0.1. For each kernel we have ploted two versions, one with regularization parameter $\lambda = 1$ and the other one with λ that was set with internal cross-validation, for each value separately. We can see results for polynomial on figure 4 and for RBF on figure 5.

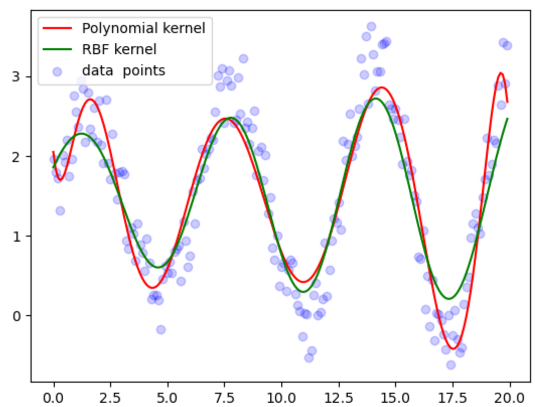

Figure 1. Figure shows us How the polynomial and RBF kernels perform on the *sine* data set. Blue points represents data points and with green (RBF) and red (polynomial) line are represented the curves we obtain from regression.

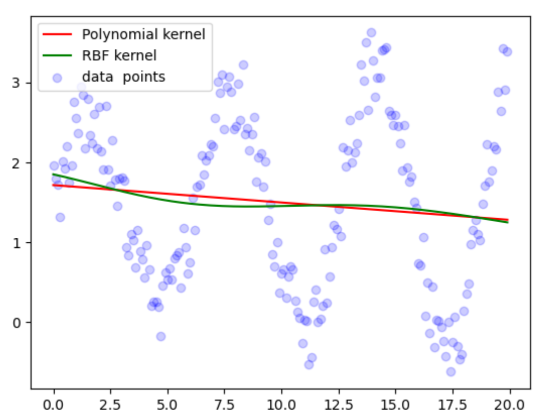

Figure 2. On this figure we can see the flattened curve we get from setting wrong parameters, that is too small M and to big σ.

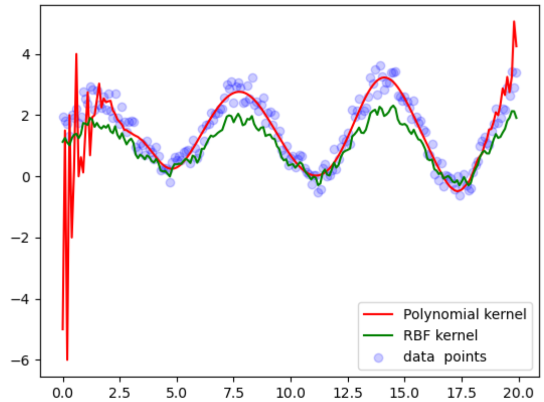

Figure 3. On this figure we can see the overfitted curve we get from setting wrong parameters, that is too big M and too small σ.

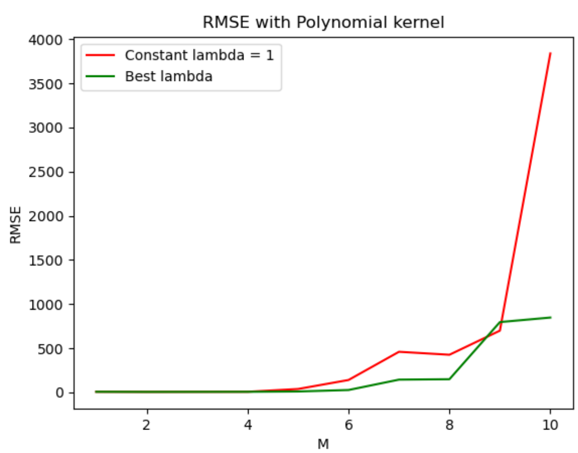

Figure 4. On this figure we can see the RMSE for kernelized ridge regression if we choose polynomial kernel as described before. Red line represents the results if for every M we choose the same regularization parameter λ and the green line represents RMSE if we choose the best λ, that we got fro internal cross validation for every $m \in M$. Lambdas that were considered are from 0.1 to 100 with a step 0.2.

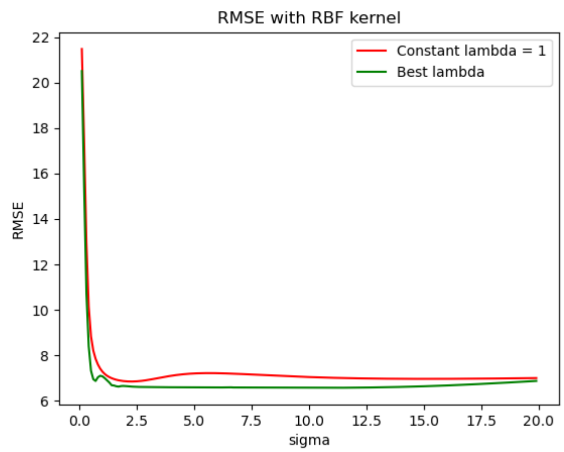

Figure 5. On this figure we can see the RMSE for kernelized ridge regression if we choose RBF kernel as described before. Red line represents the results if for every σ we choose the same regularization parameter λ and the green line represents RMSE if we choose the best λ, that we got fro internal cross validation for every $m \in M$. Lambdas that were considered are from 0.001 to 1 with a step 0.01.