Kernelized Ridge Regression

Urša Zrimšek

uz2273@student.uni-lj.si, 63200441

Introduction

This is the report of third homework for class Machine Learning for Data Science 1, in which we implemented kernelized ridge regression.

We used two kernels:

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

The models are applied to two datasets: one dimensional `sine.csv` dataset, that is composed of 200 rows with x and y values, and `housing2r.csv` dataset, that has 200 rows with 5 independent and one dependent column.

Results

Sine dataset

On sine dataset, our task was to fit the given datapoints, and plot the fit onto the points by a curve. To compare both kernels, we first needed to find the parameters that work well for each of them. We set λ - the regularization parameter - to 0.1 for both, and then compared different values of M - the degree of polynomial kernel, and σ - the width of RBF kernel. We can observe the comparison of different parameter values on Figures 1 and 2.

On Figure 3 we can see the comparison of both kernels with the "optimal" parameter values chosen above.

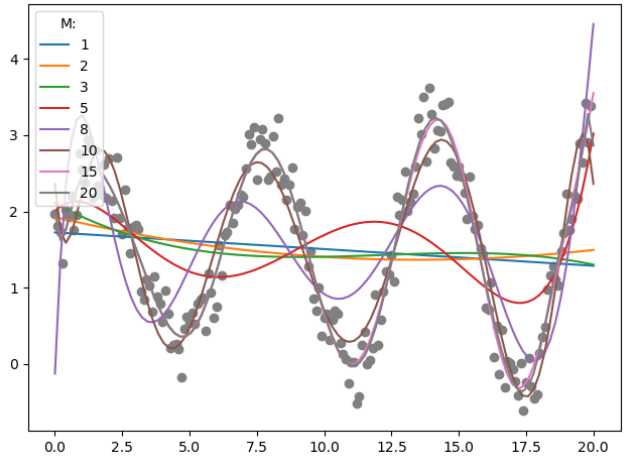

Figure 1. Fit of regression using polynomial kernels with different degrees. We can see that the degree of the kernel effects the fit a lot. We are not able to fit the data with small degrees, which is expected, since the data is too complicated for a simple function. For larger degrees, the fit around this points is similar to the fit of degree 15 and 20 - which we can see are almost identical. The problem with bigger degrees is overfitting, which we will also see later. We didn't show it on this graph, since on the edges the curve becomes so big that other can't be seen.

Housing dataset

On housing dataset we took the first 80% as the training set, and last 20% as test set. We compared RMSE for both kernels with different parameter values - M and σ. For each value of those two parameters, we did a 5- fold cross validation on the training data, to set the best value of regularization parameter λ. Then we plotted RMSE versus the parameter values, for models with $\lambda = 1$ and with λ chosen with cross validation, where we were choosing from $\lambda \in \{0.01, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}$. We calculated RMSE for $M \in \{1, \dots 10\}$ and for 2000 σ-s in between 0.01 and 20.

The results can be seen on Figures 4 and 5. On both we can see that sometimes $\lambda = 1$ can be better than the one chosen by cross validation. We could explain this by having

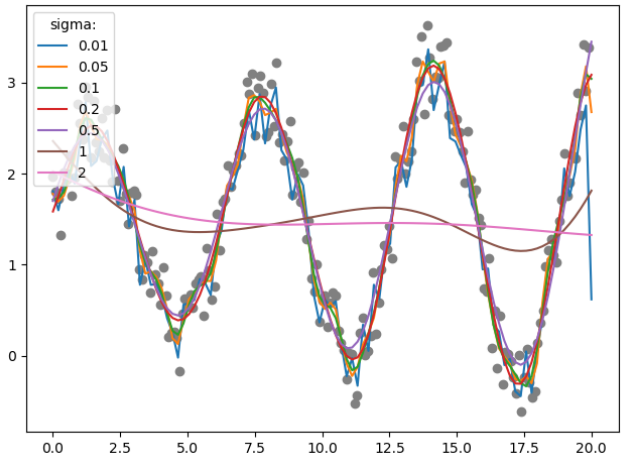

Figure 2. Fit of regression using RBF kernels with different bandwidths. We can observe that smaller σ leads to overfitting and large σ makes the fit to general. We could explain this by observing the definition of the kernel. Small σ is magnifying the effect of every datapoint, and big one is smoothing it. In this case, we chose 0.2 to be the best value.

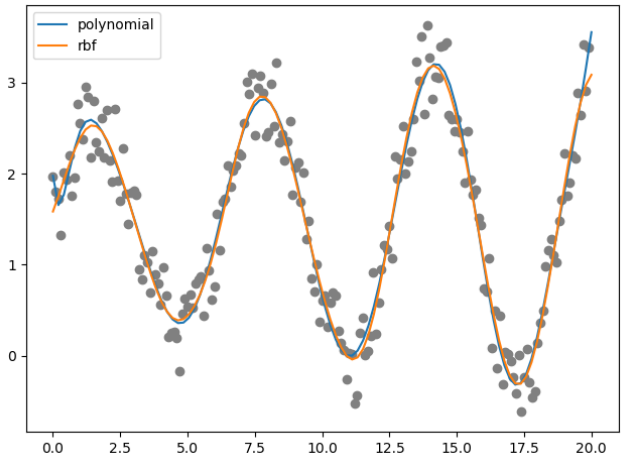

Figure 3. Fit of Polynomial and RBF kernel. Here we see the fit of polynomial kernel of degree 15 and rbf kernel with $\sigma = 0.2$. We can see that the fits are quite similar, but there is some difference in the edge points.

a small dataset, that's why the results on testing data vary from those at validation. Further we can look at the values of best λ-s. For polynomial kernel, they are generaly bigger for bigger M. This can be explained by overfitting. With bigger M, we overfit more to the training data, so we need to have larger regularization parameter. For RBF kernel, we have 2000 values of λ-s, so we only looked at each 100-th one, and didn't see such nice pattern.

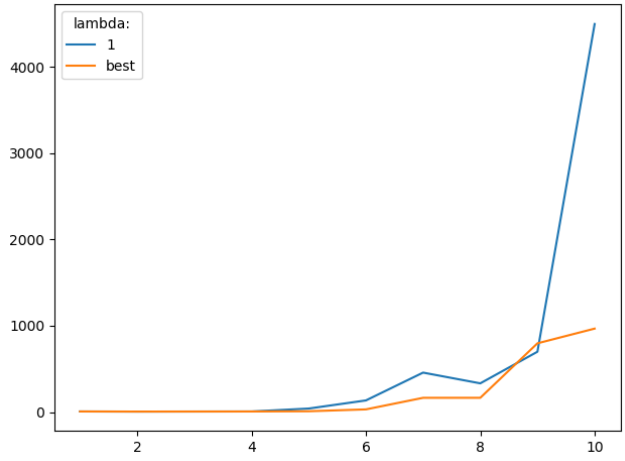

Figure 4. RMSE versus the degree of Polynomial kernel. We can observe that the error is large at bigger degree kernel, so we can conclude that this data is simple, and we overfit it quickly.

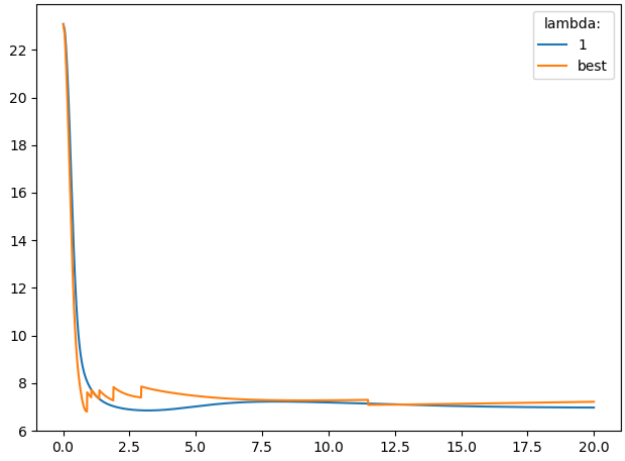

Figure 5. RMSE versus the width (σ) of RBF kernel. The error is large at the smallest σs, that is, as we saw before, the consequence of overfitting. We reach the smallest error at $\sigma = 0.9$ and $\lambda = 0.01$, and with $\lambda = 1$, the best value for σ is 3.2.