Support Vector Regression

Machine Learning for Data Science Homework #4

Vid Stropnik; 63200434

ABSTRACT

For the fourth homework in the Machine Learning for Data Science class, I worked with Support vector machines to achieve an analogue to Kernelized ridge regression, studied in the previous homework. Here, I present my own implementation of Support Vector regression and a quick exploration into it's differences and similarities to Kernelized ridge regression, while also evaluating some other properties and parameters of the method.

1 KERNELS AND PROCEDURE

In the homework, we implemented two kernels;

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

And used them to implement Support vector regression, with the prediction formula:

$$\hat{y} = ker(x, x')^T A_{opt},$$

where A_{opt} is the matrix of optimized interchanging parameters $\alpha_i, \alpha_i^*, \forall i \in [0, len(X)]$, with X denoting the independent variables in the observed training data. The optimization of these parameters was carried out using the `cvxopt.solvers.qp` library for Python, with the underlying optimization problems given in Eq. (10) in the provided article [1].

2 IMPLEMENTATION ON SINE DATASET

Given that the roles of parameters λ, M, σ were already explained in [2], they will not be reintroduced here. In a similar vein to what was done in the previous report, we show that both kernels find a good fit on the example sine dataset in Figure 1. Striving to find a good balance between the (low) number of support vectors and fit, the following parameters were considered for both observed kernels:

- $M = 10$
- $\sigma = 0.3$
- $\epsilon = 0.6$
- $\lambda_{poly} = 1, \lambda_{RBF} = 0.001$

We were able to control the aforementioned balance between support vector count and fit quality by changing the parameter ϵ, with higher values generally exhibiting slightly worse fits and lower vector-counts. Again, the plot shown in 1 is achieved by preprocessing the data using `sklearn`'s `StandardScaler`.

3 HOUSING DATASET & KRR COMPARISON

Applying SVR to a practical dataset, we were able to observe many interesting patterns of behaviour, especially when comparing it to

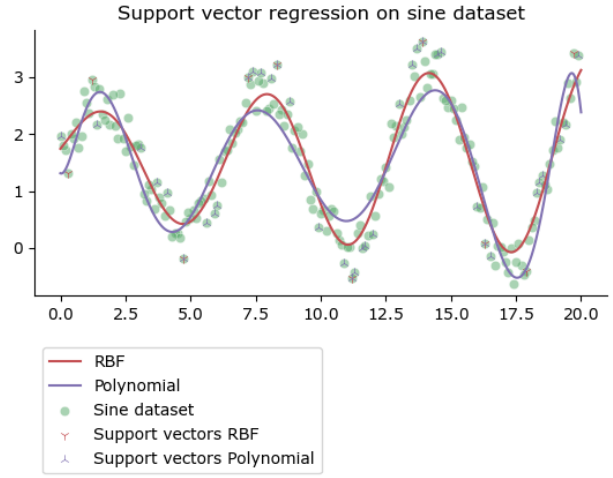

Figure 1: Plot of the two kernels' fits on a simple Sine dataset. The prediction data is generated on normalized data from the range $[0, 20]$.

the dynamics of the same plot on Kernelized ridge regression from [2][1].

All formalisms, such as the considered values of M, λ, σ, as well as the cross-validation procedure for selecting the best λ remain the same as in [2], with the parameter ϵ for solutions shown in Figure 2 equalling 8 for the Polynomial kernel and 10 in the case of RBF. Furhtermore, certain observations, such as the RMSE function's behaviour in cases with low values of σ and high values of M hold here as well.

In general, SVR seems slightly less sensitive to the selection of λ, when compared to Kernelized ridge regression, given that using an uniform value $\lambda = 1$ resulted in stable results in Figure 2. From this observation, we can hypothesise that SVR is indeed the superior model when considering less interpretable data, or experimenting with unfamiliar features.

Conversely, KRR seems to more clearly converge to the best performing hyper-parameter (σ, M), regardless of the used learning rate. This may imply that this model is a better fit when regressing for parameter selection or working with familiar data. This hypothesis is further supported by the observation that KRR does tend

[1] In [2], an error was made in the input of data on the housing dataset (wrong dependent and independent variables). While, when inputting the correct data, the major change is only the decrease in RMSE and the appearance of error spikes in certain edge cases, RMSE trends with **optimal parameter values** stay largely similar to what was originally reported, an updated (and correct) plot was considered for all comparisons done in this report. If desired, the author will provide these updated plots if contacted.

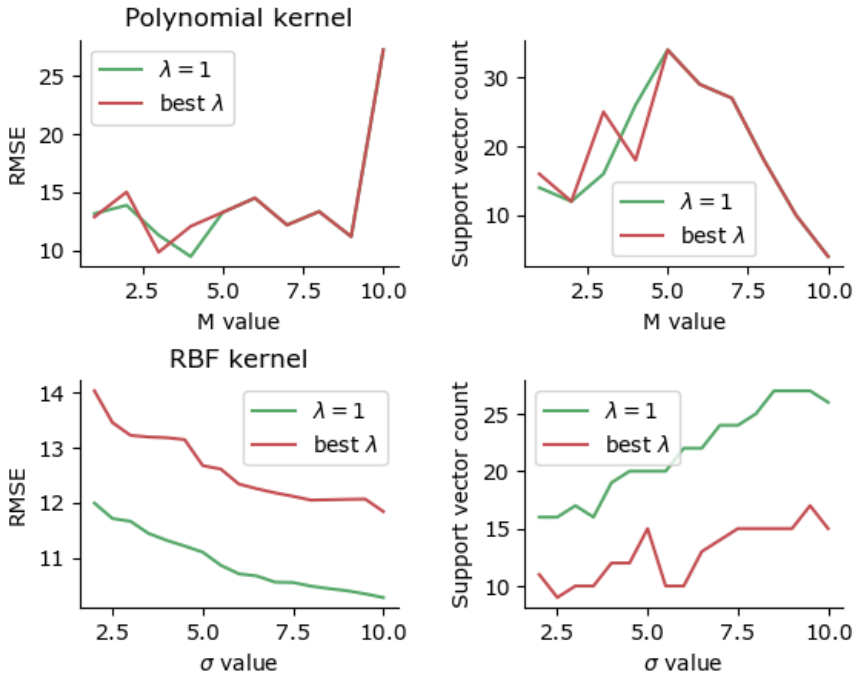

Figure 2: RMSE values and support-vector counts for predictions conducted with different values of M and σ using SVR with both kernels.

to yield slightly better (lower) RMSE scores when considering the optimal values of λ.

Finally, when observing the number of support vectors in both cases, we see that RBF generally uses a higher amount of support vectors than the polynomial kernel - this, of course, is not as clearly visible from Figure 2, but recall that the parameter ϵ is higher for the RBF Kernel. Also note, that the best λ plot is higher on the observed test data, but that is to be expected given the smaller amount of support vectors for that combination of outcomes.

4 CONCLUSION

In this homework, we showed another method of applying the kernel trick to model complex data using linear techniques. We compared support vector regression to kernelized ridge regression to observe the fact that, while SVR does tend to return slightly worse RMSE scores with optimal parameters, it is more robust and consistent across all hyperparameters and should, consequently, be considered as the preferred method. We also studied and explained the effect of the ϵ parameter and the number of considered support vectors on the method's consistency and accuracy, while providing coherent observations about the similarities of the two kernel-based methods, which, in the end, both succeed in what they set out to do quite well.

REFERENCES

[1] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.

[2] Vid Stropnik. Kernel ridge regression. *Machine Learning for Data Science Homework*, #3, 2021.