Zoe Hsieh
Comp135, Fall 2020
12/2/20

Problem 0

0a)

To be sure to overfit to a typical training set and achieve zero training error, the hyperparameter C would need to be small and gamma would need to be large. A large gamma would allow data points to be close to each other, which would result in a tendency to overfit.

0b)

True -- because SVM would predict a better model with a sparse set of data, so the penalty wouldn't affect the optimization.

Problem 1

1a)

True, because it can minimize projection error and maximize variance, which leads to a higher quality model and helps with model complexity.

1b)

It would be the same, because the other components would not be dependent on the height, thus, the change in units would not affect the first K PCA components.

1c)

Yes, this is correct, because to project the test dataset, you have to do it the same way you did to the train data, which does include computing the mean to center the data, and then finally projecting down to K-dimensions.

Problem 3

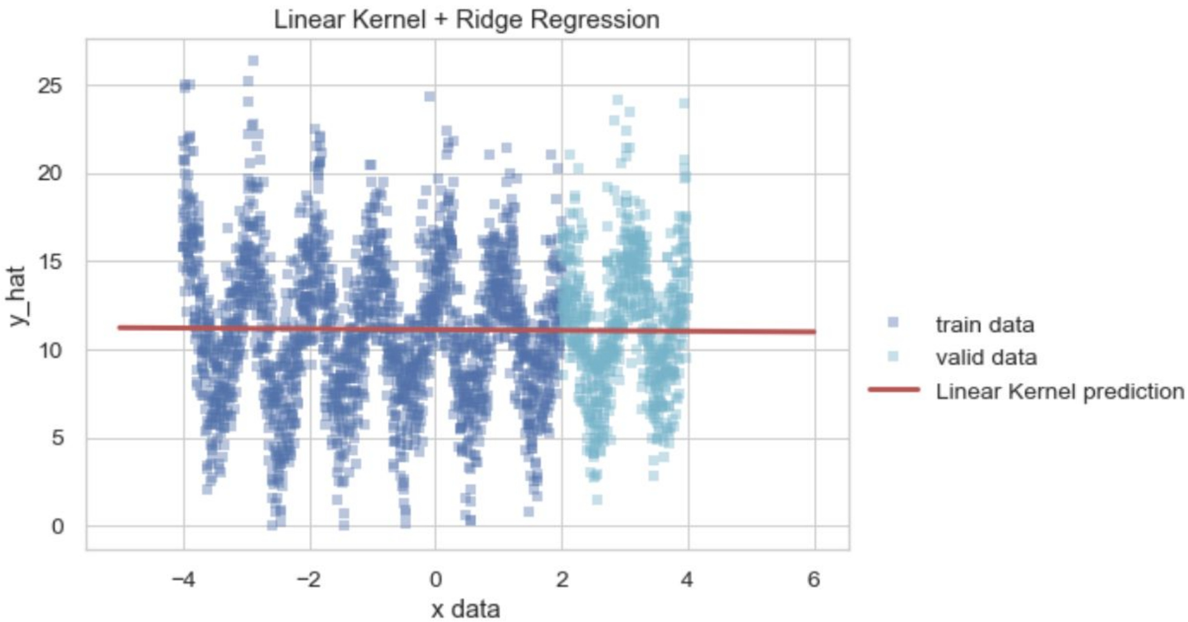

Figure 3. Predictions of the best Linear Kernel model, overlaid on the training and validation sets.

The predictions make sense because since we are running a linear regression, the regression line must be linear, and it is predicting in the middle of the data for the most accuracy. The model would be just as good at interpolating as to extrapolating since the prediction model will always predict the same y_hat value.

Problem 4

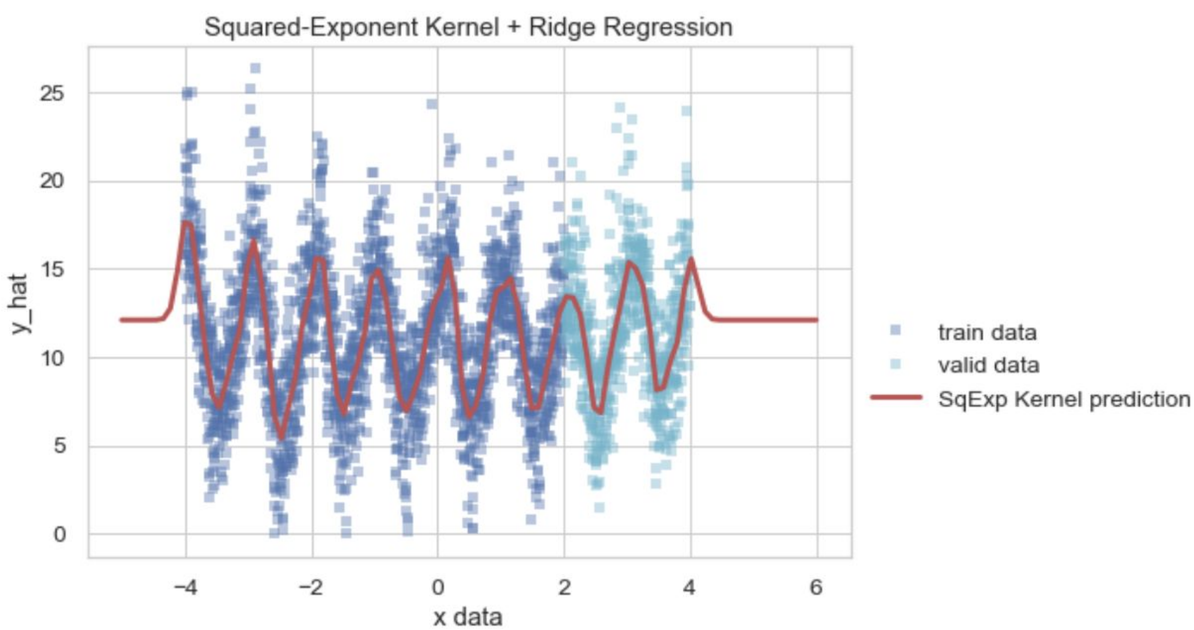

Figure 4. Predictions of the best LinSqExp Kernel model, overlaid on the training and validation sets.

The predictions for this model also makes sense because it fits the data. However, it would be good at interpolating since it follows the data shape well within the range of data given, but for extrapolating, it would be worse since it will predict the same y_hat value for data outside of the range.

Problem 5

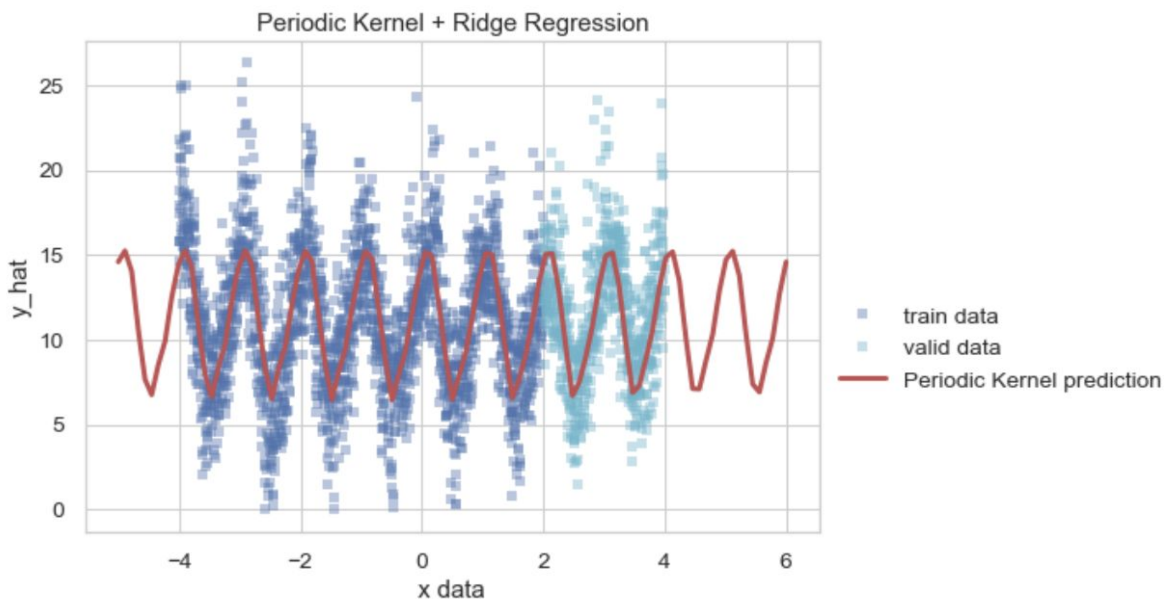

Figure 5. Predictions of the best Periodic Kernel model, overlaid on the training and validation sets.

The predictions for this model makes sense, since it fits the data very well and continues the shape outside the data. This model would be good at both interpolating and extrapolating since the predictions match the shape of the data within and out of the range of input data.

Problem 6

Split Method	**Train+Valid**	**Test**
Linear Kernel	3.251	3.402
Sqexp Kernel	2.115	3.295
Periodic Kernel	2.213	2.004
Periodic Nearest Neighbor	11.106	11.446

Table 6. Performance on Train+Valid and Test sets -- Mean Absolute Error

6a)

Based on my results, the Periodic Kernel regression had the best performance. In a sense, this makes sense because the shape of the predicted model fits well on the dataset, and continues the same shape for ranges out of the data range.

6b)

The margin error of the linear model seems meaningful in the real world, because temperature is usually predicted based on how the current season/time of the year has been consistently over the past few years, and tends to stay the same.