AMATH 482/582: HOMEWORK 2

SATHVIK CHINTA

Abstract.

1. Introduction and Overview

This time we are given data from a winery in Portugal. We are given a series of chemical measurements and the corresponding quality of the wine in our data.

2. Theoretical Background

We will firstly be using a simple linear regression model to predict the quality of the wine. This is very similar to our last problem, where we used linear regression to classify handwritten digits.

However, we will be using a more complex models as well since we will find that linear regression's error is very high on both training and test data. We will use Kernel Ridge Regression in order to do so. Ridge Regression can be defined as

$$minimize_{f \in H_k} ||f(x) - Y||^2 + \lambda ||f||^2_{H_k}$$

Where H_k is the kernel function induced by features that map from a high dimensional space to $\mathbb{R}$. In our case, we will use Gaussian (RBF) and Laplacian Kernel functions in order to model our data. The Gaussian Kernel function can be defined as

$$exp(-\frac{||x - x'||_2^2}{2\sigma^2})$$

And Laplacian can be similarly defined as

$$exp(-\frac{||x - x'||_1}{\sigma})$$

Notice that we no longer have Y in our equations, and isntead we have sigma. This is simply because of notation, we defined $\sigma = \sqrt{2Y}$. As such, we need to find the ideal values of both σ and λ. We will use the following procedure to find the ideal values:

Take a 10-fold cross validation set. In cross validation, we split the training data and use some pieces for training and other for testing while we find our ideal values for hyperparameters.

In our case for each fold, we will check the negative mean squared error of the model on the training data, until we find the parameters that output the best results. We will then use these parameters for our model. If either one of the parameters is equal to the endpoints of either one of our parameter's ranges, we move our range in order to encapsulate the best results. We will also use 10 values within the search range for each parameter, effectively giving us a 10 x 10 search grid.

With those parameters figured out, we will use them in order to train our model on the training data. We will then check how well they perform against the test dataset.

Date: February 25, 2022.

3. Algorithm Implementation and Development

4. Computational Results

After doing the normalization of the training data along with the linear regression, I got the following results

LINEAR REGRESSION TRAINING MSE	LINEAR REGRESSION TEST MSE
0.6278484956554882	0.747169690518721

Table 1. Mean squared error of Linear Regression

We can see that the error is pretty high in both cases.
Looking at the RBF Kernel, I got the following hyperparameters as optimal

α	γ	λ	σ
0.19842513149602492	0.03645403248675365	-2.33333	1.888889

Table 2. Optimal hyperparameters for RBF Regression

This resulted in the following MSE

GAUSSIAN KERNEL TRAINING MSE	GAUSSIAN KERNEL TEST MSE
0.4443644835200732	0.6707146529560462

Table 3. Mean squared error of Gaussian Kernel

Looking at the error values, it's clear that we are performing better than the Linear Regression model. Both training and test error values are lower than before, with a higher percentage decrease for the training error. Let's see if we can make these results even better with the Laplacian Kernel
I got the following hyperparameteres as optimal for the Laplacian Kernel

α	γ	λ	σ
0.2143109957132682	0.21431099571326834	-2.2222222	0.6111111

Table 4. Optimal hyperparameters for Laplacian Regression

With these parameters, I got the following MSE

LAPLACIAN KERNEL TRAINING MSE	LAPLACIAN KERNEL TEST MSE
0.058153392810602604	0.6095835762585938

Table 5. Mean squared error of Gaussian Kernel

Both errors decreased, but the training error decreased by a lot! Even though we are only about 39% accurate on the test data, we are about 95% accurate on the training! This is a clear sign of overfitting, but it still looks like this model performs the best on the test data out of all the models.
So, putting them all together, here are the results:

	TRAINING	TESTING
LINEAR REGRESSION	0.6278484956554882	0.747169690518721
GAUSSIAN	0.4443644835200732	0.6707146529560462
LAPLACIAN	0.058153392810602604	0.6095835762585938

TABLE 6. Mean squared error of all methods

On the new wine data, the results are as follows:

	PREDICTION
LINEAR PREDICTION	[6.00469789 5.28767761 5.56363072 6.067022 5.94248207]
RBF PREDICTION	[6.02506681 5.45131357 5.42488008 6.18782517 6.1157892]
LAPLACIAN PREDICTION	[6.06872311 5.45686107 5.63144745 6.00113834 6.03145372]

TABLE 7. Raw predictions on new wine data for all models

We are not getting integers in our predictions since we still have some uncertainty no matter how good our model is. This is becasue of the confidence of our model. If we were to round these results to the nearest integer, they would look like so

	PREDICTION
LINEAR PREDICTION	[6 5 6 6 6]
RBF PREDICTION	[6 5 5 6 6]
LAPLACIAN PREDICTION	[6 5 6 6 6]

TABLE 8. Raw predictions on new wine data for all models

An interesting observation is that the predictions for Linear and Laplacian are the exact same.

5. Summary and Conclusions

Acknowledgements