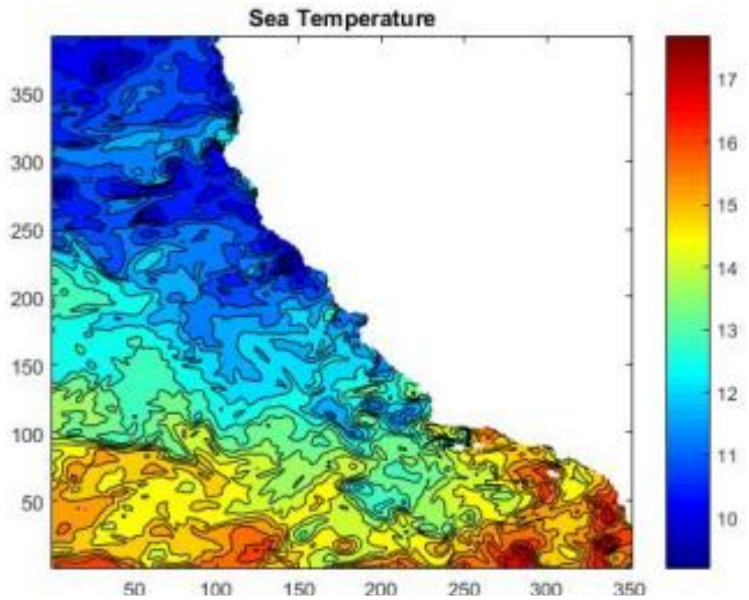
UNIVERSITÄT ZU LÜBECK
INSTITUTE FOR ROBOTICS
AND COGNITIVE SYSTEMS

rakesh.kondeti@student.uni-luebeck.de

Introduction

In this project, we use probabilistic linear regression to predict unkown temperature values at the eastcoast of the states, based on a real world data set, 'the California Coastal Regional Temp Field Dataset'.



Introducing the linear regression function:

$$y = x^T \omega \qquad (1)$$

where, y is a dependent variable and represents a temperature value. x is independent variable and stands for our coordinates(latitude and longitudes) and ω denotes a vector of weights.

Solution Method

1. Ridge-Regression

To solve the above described problem we utilize Ridge-Regression:

$$\omega = (A^T A + \lambda I A^T) A^T y \qquad (2)$$

with

$$A = [\Phi(x_1)^T, ..., \Phi(x_i)^T] \qquad (3)$$

and

$$\Phi(x) \tag{4}$$

being our basis function, applied to every input vector x of our dataset.

Ridge regression is a method for analyzing multiple regression data with collinear nature, which in essence is a solution to the least squares problem. Furthermore we used two different basis functions, applied to the input-Data (in this case input-data means positions)

2. Basis function

As introduced in eq. (3), a basis function is applied to every input vector (or in this context: position at the coast). The general goal of using a basis function is to be able to model nonlinear characteristics of our data, while still using linear regression. This trick enables us to increase the level of complexity that our regressor is able to handle.

(a) Identity basis-function

The identity basis-fuction simply returns the vector itself, so no function is applied:

$$\Phi(x) = x \tag{5}$$

(b) Gaussian basis-function

The Gaussian basis function applies the following function to every input vector x:

$$\Phi_i(x) = exp\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\} \tag{6}$$

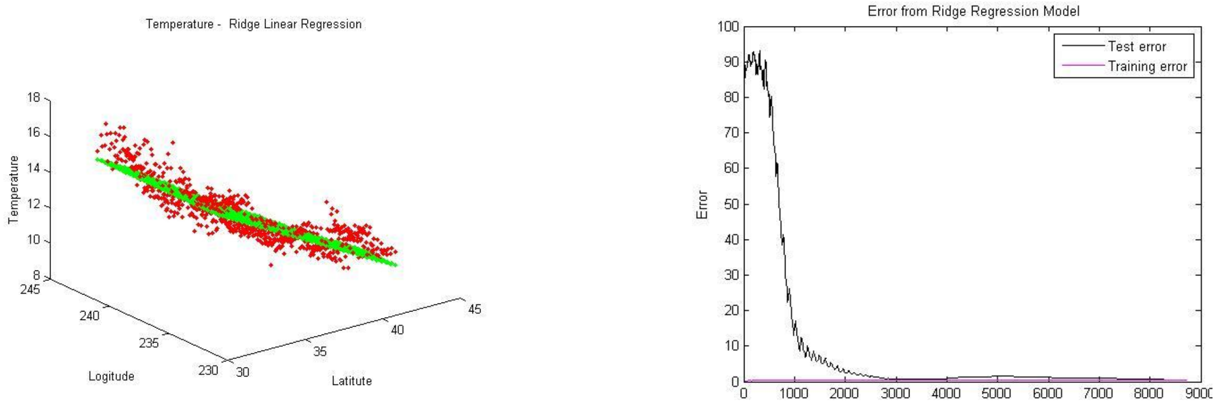
Which evectively maps our input position to a scalar. But due to the fact that we apply this function not only once, but i times, each time with a different μ we map from a two-dimensional input (our coordiates) to a i-dimensional input, with i beeing the nuber of mean-centers we spread across our dataset.

Results

3. Results with identity basis

The identity basis is the most simple basis, it does not apply any kind of transformation to the input data:

$$\Phi(x) = x \tag{7}$$



(a) Ridge Regressionresultona test data and $\lambda = 1$

(b) Mean squared error on training and testing set, plotted over increasing train-set size

Figure 2

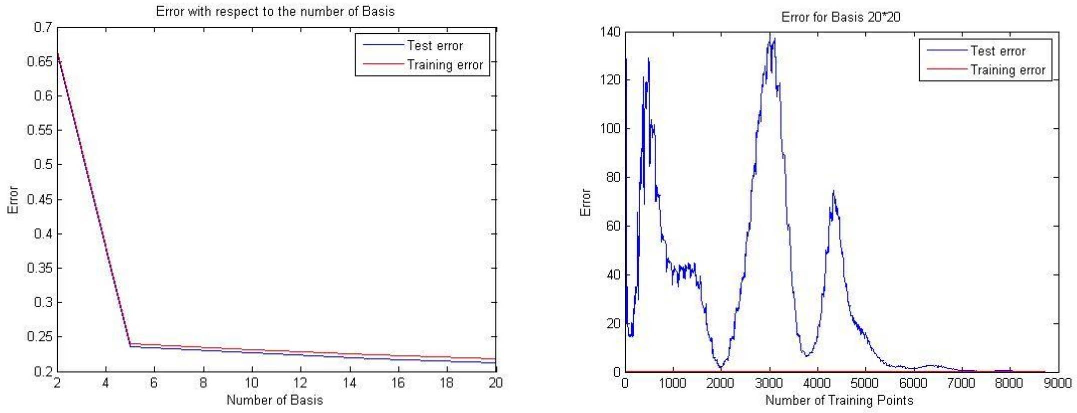
As expected we observe a smooth, linear transition in the predicted temperature. If we look into the Regression graph, we can see that the temperature gradient - predicted by our regressor with identity basis - matches the general temperature gradient of our dataset. With this in mind, we clearly observe underfitting behavior, due to the fact that a identity basis won't be able to model fine-grained temperature and hence bias.

If we plot the mean squared error (MSE) in dependency of the size of our training set, we observe a very stable behavior until we reach a training size so small that it does not allow our regressor to generalize our training set. (training size < 8k points in total = underfitting)

Using the majority of our Data to train leads to a bigger error on the testing set, but this is mainly because of the size of the dataset, not due to overfitting.

4. Results with gaussian basis

To solve the problem of underfitting we will use the gaussian-basis function introduced in eq. 6 to enable our regressor to model the complexity of our dataset.



a)Gaussian basis vs Error on training and test data

b)Error vs Number of Training points in 20*20 basis

The data is projected into Gaussian Basis by using equation 6. It can be observed that the error of train data and test data decreases as the number of Gaussian basis increases. Though Error vs number of Training samples graph does not seem to be right, it can be generalized that Error minimizes as the number of training samples increases.

Conclusion

5. thought on ridge regression

In our project, ridge regression delivered good results on estimating the temperature values, but errors might occur if one chooses λ too high in comparison to the desired output range in y, effectively adding a bias on the output which might offset the output to undesirable values. Also on higher dimensional input vectors ridge regression might fail on 'cutting out' irrelevant dimensions because ridge regression shrinks it's coefficients towards 0 but never exactly to 0.

6. thoughts on the identity-basis regression

Generally we are satisfied with de predictions of our model, even the underfitting behavior of our identity-basis regression might be reasonable depending on the desired accuracy. Especially for small datasets we think the identity-basis regression might perform quite well.