Hacettepe University

Computer Engineering Department

BM409 Machine Learning Laboratory - 2023 Spring

Assignment 1

April 6, 2023

Student name:
Ataberk ASAR

Student Number:
b2210356135

Problem Definition

Linear regression model implemented and used for predicting car prices using real life dataset

Data Preprocessing

Data preprocessing is an important part of machine learning. Raw data has to be preprocessed in a way that machine learning model can work on it.

Feature Selection: “Id”, “Ilan Basligi”, “Renk”, “Il / Ilce” removed from dataset, since their effect is minimal to target (“Fiyat”).

Encoding Categorical Data: “Ilan Tarihi” column preprocessed to have only year information, and then encoded in binary form (1 for 2020, 0 for 2019). Values in “Seri” and “Model” colums are one-hot-encoded. But since there is too many categories (see Curse of Dimensionality), least occuring values in each column switched to “Rare” value and then one-hot-encoded.

Values in “Fiyat” column expressed as powers of 10 (e.g. 1000 -> 3) for normalization.

Solution

Linear regression model implemented using Python. Numpy, Pandas, Scikit-learn libraries used for data preprocessing. Matplotlib library used for evaluating experiment results.

Basic linear regression model does not have good performance on given training data. Thus different basis functions (linear basis function, polynomial basis function, gaussian basis function, and sigmoidal basis function) for feature space extension, and L2 regularization also implemented. Mean square error also implemented as loss function.

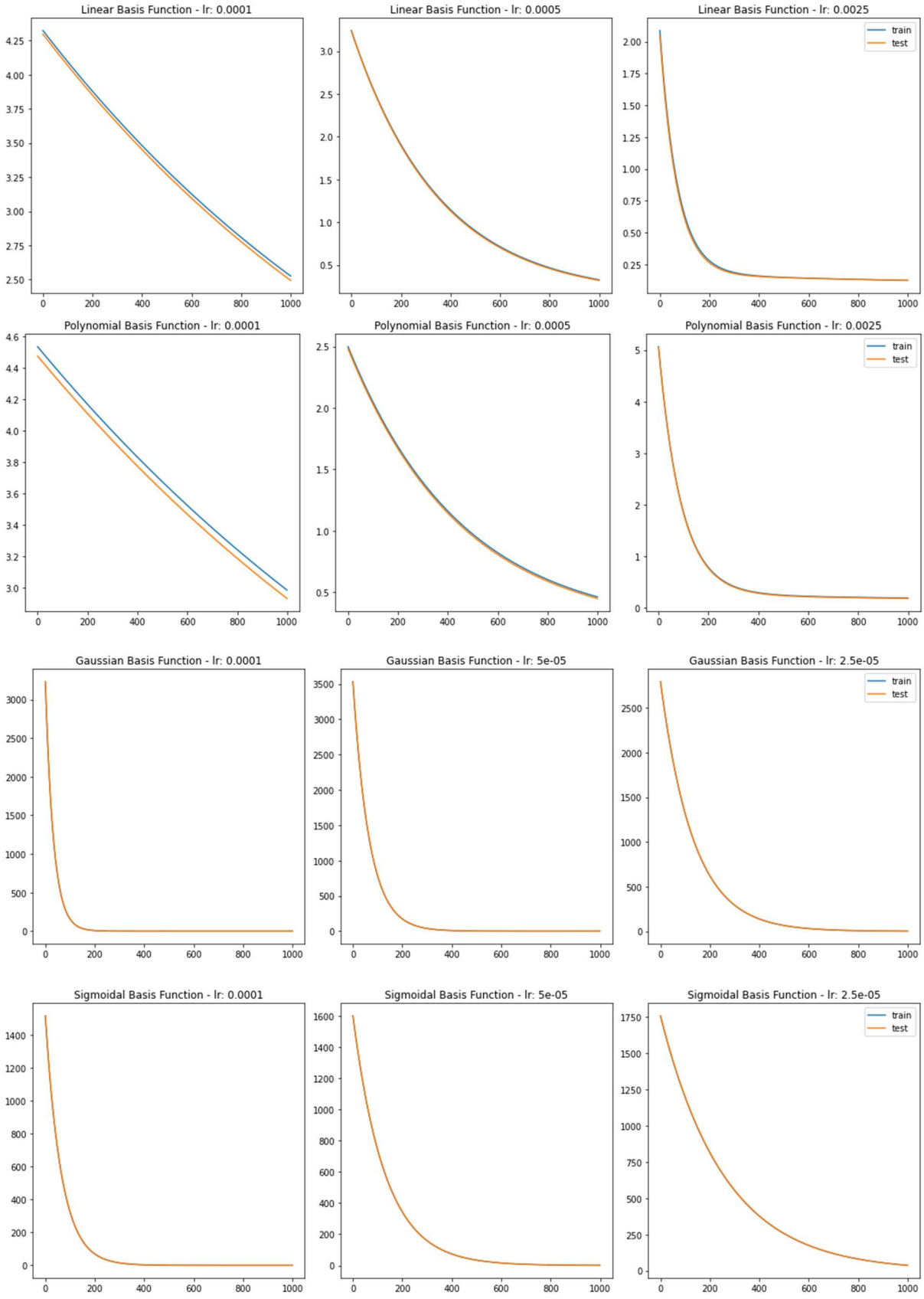
Experiment Results

Different experiment setups used for prediction. In Table-1, training losses and test losses over 1000 iteration for different models shown.

	Learning Rates				
Models	**25e-6**	**5e-5**	**1e-4**	**5e-4**	**25e-4**
Linear Basis Function	-	-	2.5257 2.4939	0.3262 0.3191	0.1243 0.1249
Polynomial Basis Function	-	-	2.9848 2.9311	0.4618 0.4491	0.1928 0.1853
Gaussian Basis Function	1.5848 1.5902	0.0653 0.0649	0.0828 0.0812	-	-
Sigmoidal Basis Function	37.5964 37.7398	0.7946 0.7977	0.0608 0.0580	-	-

Table-1

Loss graphs can be seen below over 1000 iterations



Regularization is technique used in case of overfitting, to give model more stability. As one can see from previous graphs, no experiment setup has overfitting. L2 regularization applied to the model with sigmoidal basis function and 1e-4 learning rate, for demonstration purposes. Results shown in Table-2.

	Lambda Values		
Model	**0**	**0.1**	**0.5**
Sigmoidal Basis Function	0.0656 0.0652	0.0659 0.0645	0.0811 0.0802

Table-2