, Homework **1: Regression**

200 marks total

This assignment is to be done individually.

Important Note: You may not provide or use any solution, in whole or in part, to or by another student.

You are encouraged to discuss the concepts involved in the questions with other students. If you are in doubt as to what constitutes acceptable discussion, please ask! Further, please take advantage of the office hours offered by the instructor and the TA if you are having difficulties with this assignment.

DO NOT:

- Give/receive code or proofs to/from other students
- Use Google to find solutions for assignment

DO:

- Meet with other students to discuss assignments (it is best not to take any notes during such meetings, and to re-work assignments on your own)
- Use online resources (e.g. Wikipedia) to understand the concepts needed to solve the assignment

---

Question 1 (20 marks)

The weight and systolic blood pressure of 26 randomly selected males in the age group 25 to 30 are shown in the following table. Assume that weight and blood pressure are jointly normally distributed. Find a regression line relating systolic blood pressure to weight.

Subject	Weight	Systolic BP	Subject	Weight	Systolic BP
1	165	130	14	172	153
2	167	133	15	159	128
3	180	150	16	168	132
4	155	128	17	174	149
5	212	151	18	183	158
6	175	146	19	215	150
7	190	150	20	195	163
8	210	140	21	180	156
9	200	148	22	143	124
10	149	125	23	240	170
11	158	133	24	235	165
12	169	135	25	192	160
13	170	150	26	187	159

Question 2 (30 marks)

A study was performed on the wear of a bearing y and its relationship with x1 oil viscosity and x2 load. The following data were obtained.

(a) Fit a multivariate linear regression model to these data. (10 Marks)

(b) Use the model to predict wear when $x_1 = 25$ and $x_2 = 1000$. (5 Marks)

(c) Fit a multivariate linear regression model with an interaction term (x_1x_2) to these data. (15 Marks)

y	x_1	x_2
293	1.6	851
230	15.5	816
172	22.0	1058
91	43.0	1201
113	33.0	1357
125	40.0	1115

Question 3 (30 Marks)

In a regression problem, we predict O based on one n-dimensional vector (sample) X using the following relation:

$$o = w_0 + w_1 x_1 + w_1 x_1^3 + w_2 x_2 + w_2 x_2^3 + ... + w_n x_n + w_n x_n^3$$

(a) Suppose that we have m samples in our dataset. Write the Matrix form of O=ZW with complete displays of O, Z, and W matrices (Z is constructed based on X) (20 Marks)

(b) If we want to obtain W using the gradient-descent method, write the gradient descent relation for this problem such as below (10 marks)

$$w_i \leftarrow w_i + \ldots \text{ for } 1 \leq i \leq n.$$

The update relation is not required for the bias term (w_0)

Question 4 (70 marks) (implementation)

In this question you will implement linear basis function regression with polynomial and Gaussian bases.

Start by downloading the code and dataset from the website. The dataset is the AutoMPG dataset from the UCI repository (https://archive.ics.uci.edu/dataset/9/auto+mpg). The task is to predict fuel efficiency (miles per gallon) from 7 features describing a car.

Functions are provided for loading the data[1], and normalizing the features and targets to have 0 mean and unit variance.

```
[t,X] = loadData();
X_n = normalizeData(X);
t = normalizeData(t);
```

For the following, use these normalized features X n and targets.

You may also find the provided function designMatrix.m useful.

Polynomial basis functions

Implement linear basis function regression with polynomial basis functions. Perform the following experiments:

1. Using the first 100 points as training data, and the remainder as testing data, fit a polynomial basis function regression for degree 1 to degree 10 polynomials. Do not use any regularization. Plot training error and test error (in RMS error) versus polynomial degree. **Put this plot, along with a brief comment on what you see, in your report.** For the basis functions:
   (a) Include the bias function as always.
   (b) For each input variable x_i, $i = 1, \ldots, 7$, and for each power $k = 1, \ldots, degree$, there is a basis function $\Phi_{i,k}$ that returns x_i^k. So you should have $7 \cdot degree$ basis functions, plus the bias function, and hence that many weights. In words, treat each input variable as separate single variable and then follow the treatment in the book and use the single-variable powers up to the degree. The degree (maximum power) should run from 1 to 10.

2. It is difficult to visualize the results of high-dimensional regression. Instead, only use one of the features (use X n(:,3)) and again perform polynomial regression. Produce plots of the training data points, learned polynomial, and test data points. The code visualizes 1d.m may be useful. **Put 2 or 3 of these plots, for interesting (low-order, high-order) results, in your report. Include brief comments.**

[1] Note that loadData reorders the datapoints using a fixed permutation. Use this fixed permutation for the questions in this assignment. If you are interested in what happens in "reality", try using a random permutation afterwards. Results will not always be as clean as you will get with the fixed permutation provided.

3. Implement L_2-regularized regression. Again, use the first 100 points, and only use the 3rd feature. Fit a degree 8 polynomial using $\lambda = \{0, 0.01, 0.1, 1, 10, 100, 1000\}$. Produce a plot of train and test set error versus regularizer value. Use a semilogx plot, putting regularizer value on a log scale. **Put this plot in your report**.

Gaussian basis functions

Implement linear basis function regression with Gaussian basis functions. You may use the supplied dist2.m function. For the centers μ_j use randomly chosen training data points (use randperm in MATLAB). Set $s = 2$. Perform the following experiments:

1. Using the first 100 points as training data, and the remainder as testing data, fit a Gaussian basis function regression using 5, 15, 25, . . . , 95 basis functions. Do not use any regularization. Plot training error and test error (in RMS error) versus the number of basis functions. **Put this plot, along with a brief comment on what you see, in your report.**
2. Implement L_2-regularized regression. Again, use the first 100 points (do **not** only use the 3rd feature but also use them all). Fit a regression model with 90 basis functions using $\lambda = \{0, 0.01, 0.1, 1, 10, 100\}, 1000$. Produce a plot of train and test set error versus regularizer value. Use a semilogx plot, putting regularizer value on a log scale. **Put this plot in your report**.

Question 5 (20 marks) (implementation)

You have been given a one-dimensional dataset **(q4_dataset.csv)**. Consider column X as input features and column Y as labels and visualize the data given in a two-dimensional space.

a) If you want to fit a Linear Regression model on this dataset, there would be a basis (kernel) function that maps the features to a better space. Can you guess which one?

b) Divide this dataset into two **train** and **test** subsets by random sampling such as the test set includes 20% of the total dataset. Train a Linear Regression model using the kernel function asked in part (a) and evaluate it on the test set by **computing MSE (Minimum Squared Error)**. **Notice That in this task you are not allowed to use any ML libraries.**

Question 6 (30 marks) (implementation)

Abalone is one of the famous datasets available on the UCI repository. Given this dataset, the task is to predict the number of lines on abalones which indicates their age. **Features given on this dataset are all continuous except one which defines the gender of abalone**. more information about the dataset and features can be found at https://archive.ics.uci.edu/dataset/1/abalone/.

After preparing the dataset by removing the **gender feature** and splitting the dataset into two **train** and **test** sets (similar to question 5), fit a Linear Regression model and report **MSE** on the train and test set using the following basis (kernel) functions:

a) Linear

b) Polynomial (degree=2)

c) Polynomial (degree=3)

d) RBF kernel functions.

Submitting Your Homework

You must create your assignment in electronic form in **PDF format**. Submit your assignment using the *new* online assignment submission server **https:// vc.kntu.ac.ir . Check early that you are set up for this, especially that you can log in to the server.**

You should create a report with the answers to the questions and figures described above. Make sure it is clear what is shown in each figure. **I INCLUDE** your source **6**de**.**