ML: Assignment:2
Arushi Jain, 2013023

Programming Questions:

1. The training data is used in quantums of 10 from 50-100 for the files lin.txt and sph.txt.

   The main file in the code, inputs the file number (0/1) and calls the **extracting_data(filename,phi,alpha,max_iter,delta,fileValues)** and **calc_plot_values(filename,phi,alpha,max_iter,delta,fileValues,threshold) functions**. This extract the data from the file and runs the gradient descent algorithm on sets of 50-90% data one by one via the **lin_reg(X,phi,max_iter,delta,alpha,threshold)** function. The whole data is then predicted using the model formed. We have phi = 0 for linear regression.
   In linear regression, we take the equation to be : h(x) = theta0 + theta1 * x
   We estimate theta0 and theta1 by gradient descent to get good prediction values.

   MSE against the percentage of data points (50-90) is plotted.
   We see that the mean square error decreases as the testing data size increase because the prediction of data improves with increase in testing data size.
   Hence the plot for both lin.txt and sph.txt for MSE vs amount of data plot is a decreasing curve.

2. The training data is used in quantums of 10 from 50-100 for the files lin.txt and sph.txt.

   The main file in the code, inputs the file number (0/1) and calls the **extracting_data(filename,phi,alpha,max_iter,delta,fileValues)** and **calc_plot_values(filename,phi,alpha,max_iter,delta,fileValues,threshold) functions**. This extract the data from the file and runs the gradient descent algorithm on sets of 50-90% data one by one via the **lin_reg(X,phi,max_iter,delta,alpha,threshold)** function. The whole data is then predicted using the model formed. We have phi = 0 for linear regression. The whole data is then predicted using the model formed.

   This is same as the above part, except that we also implement polynomial and gaussian regressions. phi is set to 1 for polynomial regression and phi is set to 2 for gaussian regression. The data set is altered accordingly before running the gradient descent for each case. For polynomial we add an additional factor of x^2 in the dataset, while in the case of gaussian we calculate w(i) = exp (− (x (i) − x)^2 / (2 * τ^2)) for each value of x.

   The combined plots can be found in the folders of sph.txt and lin.txt in report_plots folders (attached)

The **lin_reg(X,phi,max_iter,delta,alpha,threshold) function** is then run on the obtained data in a similar manner as part a. Combining this part with part a, we note down the MSE vs amount of data plots for phi = 0 (blue), phi = 1 (green), phi = 2 (red) and put them in a single graph in different colors. This is done for sph.txt and lin.txt

In the case of lin.txt, the plot has higher MSE values for gaussian regression curve as compared to the corresponding values for polynomial and linear curve. Polynomial regression curve has a smaller decrease in the MSE values with increase in amount of data as compared to linear regression. Hence, linear regression is best here.

In the case of sph.txt, the plot has higher MSE values for gaussian regression curve as compared to the corresponding values for polynomial and linear curve. Polynomial regression curve has a smaller values of MSE values with increase in amount of data as compared to linear regression. Hence, polynomial regression is best here.

3.

We have plotted scatter plots for lin.txt and sph.txt. The scatter plots contain multiple graphs for amount of data set as 50,60,70,80,90 percent, each combined with phi = 0, phi = 1, phi = 2
For each model and each data set, the predicted values are plotted along with the actual values (y) against the corresponding x values on the x axis.
In the case of linear regression, a straight line plot is obtained. In the case of lin.txt, the slope of the predicted line for the equation decreases as predicted data moves closer to the actual data (50-90%). In sph.txt the slope of the line increases. The original data in the sph.txt is more tending towards a polynomial shape and is not well defined by linear regression. lin.txt has more scattered data and is comparatively better defined by the equation predicted for it.
Similarly, we can extend the argument to for polynomial and gaussian plots. The plot of predicted data moves towards the higher density of number of points as the size of dataset increases (more accuracy).
The plots can be found in the folders of sph.txt and lin.txt in report_plots folders (attached) for all values of phi.

4.

We have added an extra parameter to the original equation of theta calculation for ridge regression:
$\theta j := \theta j + \text{sum} (\alpha(y (i) - h\theta(x (i))) xj(i)/m - \lambda\alpha\theta j/m)$
The ridge regression is implemented for the files seeds_dataset, iris, and AirQuality.

Various average values of MSE are calculated for dataset in quantums of 10 with different values of delta. The value of delta which gives the minimum average MSE after multiple runs is chosen as the appropriate average value.
The MSE vs amount of data plots are made for the cases of phi = 0, phi = 1, phi = 2 for each of the three files.
All plots show decreasing graphs for all the plots which indicates convergence.

The plots can be found in the ridge_regression folder in report_plots folders (attached) for all values of phi.

5.

We have taken the data sets of iris, seeds and airquality. We perform 10 fold cross validation and obtain the mean and standard deviation of the error.
For seeds:

phi = 0:

mean: 0.753020824915

std_dev: 0.00653126846128

phi = 1:

mean: 0.939573455283

std_dev: 0.0185771980128

phi = 2:

mean: 0.800530719668

std_dev: 0.0122514616798

For iris:

phi = 0:

mean: 0.308962208266

std_dev: 0.0200125560943

phi = 1:

mean: 0.257041661994

std_dev: 0.00650358546386

phi = 2:

mean: 0.730553802643

std_dev: 0.0272931283405

For Air Quality:

phi= 0

mean: 1.33336040812

std_dev: 0.120552761298

phi= 1

mean: 0.999139594487

std_dev: 3.11904715494 e-05

phi= 2

mean: 0.996316711307

std_dev: 0.000286031626474