ASSIGNMENT 2-REPORT
CS5691-PRML

Given csv files are converted to pandas dataframe and then dataframe is converted to 2 numpy arrays data points and labels
1)i)least squares solution W_{ML} to the regression problem Is obtained by using expression
$W_{ML} = (XX^{T})^{-1}XY$
X is data points of dataset
Y is labels of dataset

ii)here in this question ϵ(epsilon) was chosen and applied gradient descent algorithm to find appropriate w. This gradient descent algorithm was converged when the norm of difference of wt in this iteration and previous iteration is less than epsilon(||$w^{t+1} - w^{t}$||<ϵ)).ϵ=0.0001 was chosen.
Constant step size 0.000001(η^{t}) was chosen because this was converging the norm of difference of W_{ML} and w^{t} to zero as the algorithm converges implies that w^{t} was converging to W_{ML} and step sizes with greater order than order(0.000001) does not converge the norm of difference of W_{ML} and w^{t} to zero as the algorithm converges instead diverges it. For step sizes such as(0.000002 & 0.000003) the algorithm converges in a lesser number of iterations.as step size decreases the number of iterations for convergence of algorithm increases.
As the epsilon decreases, the algorithm converges in a larger number of iterations .

iii)in this question a bunch of 100 data points and corresponding labels were uniformly chosen from the dataset in each iteration of gradient descent and gradient descent is applied on the selected bunch of data points and labels. w_{avg}(average of w^{i}'s upto t iteration) is obtained in each iteration.and constant step size of 0.0001 was chosen. greater the step size faster the algorithm converges i.e norm of difference of W_{ML} and w_{avg} converges to zero in lesser iterations.when taken step size lesser than 0.0001 norm of difference of W_{ML} and w_{avg}does not converge to zero as algorithm converges.
Stochastic gradient descent algorithm takes large number of iterations to converge than gradient descent algorithm with relatively smaller ϵ and smaller step size η^{t}.

2)
In the source code 2(i) and 2(ii) were combined and marked as q2
In ridge regression loss function has extra term of λ||w^{2}|| is added in comparison to linear regression loss function.λ(lamda) is varied b/w 0 to 20.step size=0.000001 and epsilon=0.0001 was chosen.

K fold cross validation is applied and avg error is calculated for each λ.gradient descent is applied on the data set without the valid set and the valid set was chosen to calculate error. errors was averaged for each λ and error was λ was plotted. λ with the least error was chosen. W_R was calculated on the chosen λ with least error with the closed form expression(W_R= $(XX^T + \lambda I_d)^{-1}XY$).

Test error was computed using W_{ML} and W_R on the test dataset.turns out to be test error with W_R is lesser than the test error with W_{ML}.hence W_R is better than W_{ML}.

Whatever w we obtain we try to check how closer the predicted value(X^Tw) is to the actual value(Y).to know which w is better we calculate the mse(w)=E[$||$w^-w$||^2$] where w^ is the optimal w.we chose w with least mse(mean square error)
mse(W_{ML}) =(E[$||$w^-W_{ML} $||^2$]=σ^2trace($(XX^T)^{-1}$) if eigen values of XX^T are a1,a2,a3……ad then eigen values of $(XX^T)^{-1}$ are1/a1,1/a2,1/a3…….1/ad then trace(($(XX^T)^{-1}$) is sum(1/a1,1/a2,1/a3…….1/ad)

mse(W_R) =(E[$||$w^-W_R $||^2$]=σ^2trace($(XX^T + \lambda I_d)^{-1}$) eigen values of $XX^T + \lambda I_d$ are a1+λ,a2+λ ,a3+λ……ad+λ then eigen values of $(XX^T + \lambda I_d)^{-1}$ are 1/(a1+λ),1/(a2+λ),1/(a3+λ)……1/(ad+λ) then trace($(XX^T + \lambda I_d)^{-1}$) is sum(1/(a1+λ),1/(a2+λ),1/(a3+λ)……1/(ad+λ))
We can observe that mse in case of ridge regression is lesser than mse in case of linear regression.hence we chose W_R over W_{ML}.