# Report for Homework-1 (All answers)

* Answer to 1.1. attached as a different file with name answer\_1.1\_jeet
* Q(1.2) In LinearRegression\LinearRegression.java, fill in the missing lines in the function of estimating weights (𝛽) in the java code provided, using both (1) closed-form solution and (2) online updating (𝑠𝑡𝑒𝑝 𝑠𝑖𝑧𝑒 𝜂 = 0.1).
  + 1.2a – Report weigh beta and msein test data set for each version , are they same and why?
    - The beta values for both version can be found in the output file linear-regresssion-outpout.tzt and linear-regression-output-online.txt respectively. To get the values please read “Readme“ file.
    - The mse value for closed form and online update method without normalizing the data are 4.40 and 4.41 respectively. They are the same. The reason why they are same is because closed form solution is used on a linear function and online update method is used for non linear function. Since we are trying to find a straight line to fit among the given points we have a linear function. The gradient descent method on linear function does not give much difference in beta value. Also online update method works best with large data set. Online update method also produces a global minima on large set but in case of small data set as ours it produces a local minima which does not make a big difference in beta value of the two and hence we get the same values.
  + 1.2b – Implement z-score normalization and report if it affect beta and mse for both version of algorithm and why?
    - The Beta values are in the ouput files. The instruction are given in the read me file
    - The mes values after normalizing are 4.40 and 4.57 for closed form and online update respectively.
    - The mse values are almost same.
    - Normalizing helps in converging beta faster.
    - The reason they are almost same is because again online update works best for large data set and non linear function because for linear function online update gives local minima which does not affect the value of beta. Normalizing only helps in converging the beta faster.
* Q(2.1) – answer to this file is uploaded as a separate attachments name answer\_2.1\_jeet
* Q(2.2)-
  + 2.2a – calculate the average accuracy base on 5-fold cross validation.
    - The accuracy for information gain using 5 fold cross validation is 94.022%
    - The accuracy for gain ratio using 5 fold cross validation is approximately 94
  + 2.2b – Calculate average accuracy for both methods and which will you use to select the attributes and why
    - the average accuracy is given above.
    - The method to be chosen this problem doesn’t matter because
      * Id3 is used for discrete value and c4.5 can be used both for discrete and continuous values but since we have only discrete values c4.5 does not have any added advantage
      * C4.5 is also used to handle large data set but since our data is comparatively smaller using c4.5 over id3 does not have any added advantage
      * C4.5 also handles the issue of missing values. Since in out problem the missing values are handled as separate data points we don’t have to worry about them and using c4.5 does not give any advantage.
      * The Id3 method favours values that have more data points and hence their entropy is more but c4.5 averages that out but since the accuracy is approximately same there is no advantage in choosing one method over other.