PRML Assignment 2

October 27, 2022

Ravi Prakash Singh | CS22M069
M.Tech CS | Indian Institute of Technology, Madras

1. You are given a data-set with 400 data points $\{0,1\}^{50}$ generated from a mixture of some distribution in the file A2Q1.csv (Hint: Each datapoint is a flattened version of a $\{0,1\}^{10*5}$ matrix)

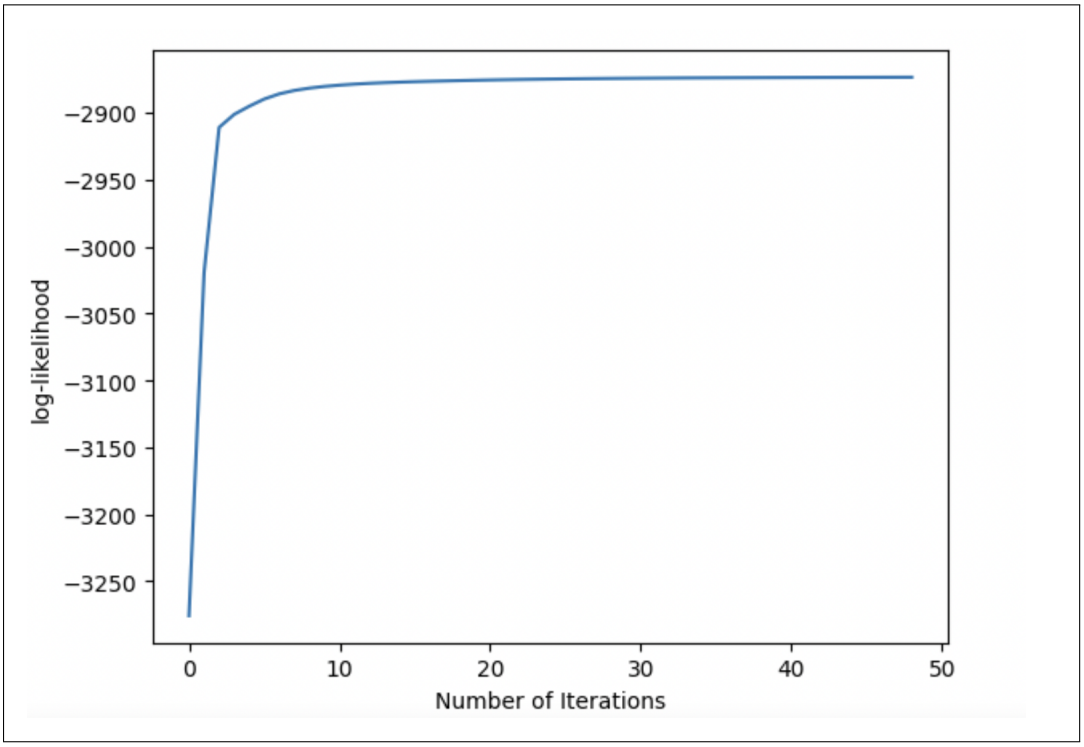
   (a) Determine which probabilistic mixture could have generated this data.(It is not a Gaussian mixture).Derive the EM algorithm for your choice of mixture and show your calculations. Write a piece of code to implement the algorithm you derived by setting the number of mixtures K = 4. Plot the log-likelihood (averaged over 100 random initialization) as a function of iterations.

Ans: I have used multivariate Bernoulli's distribution because data given to us is 50 dimensional and data points can take values either 0 or 1. Since data is in discrete nature so multivariate Bernoulli's distribution can best explain the data.
The derivation for EM algorithm is as below :

$$p_k^d = \frac{\sum_{i=1}^{n} \lambda_k^i f_d^i}{\sum_{i=1}^{n} \lambda_k^i}$$

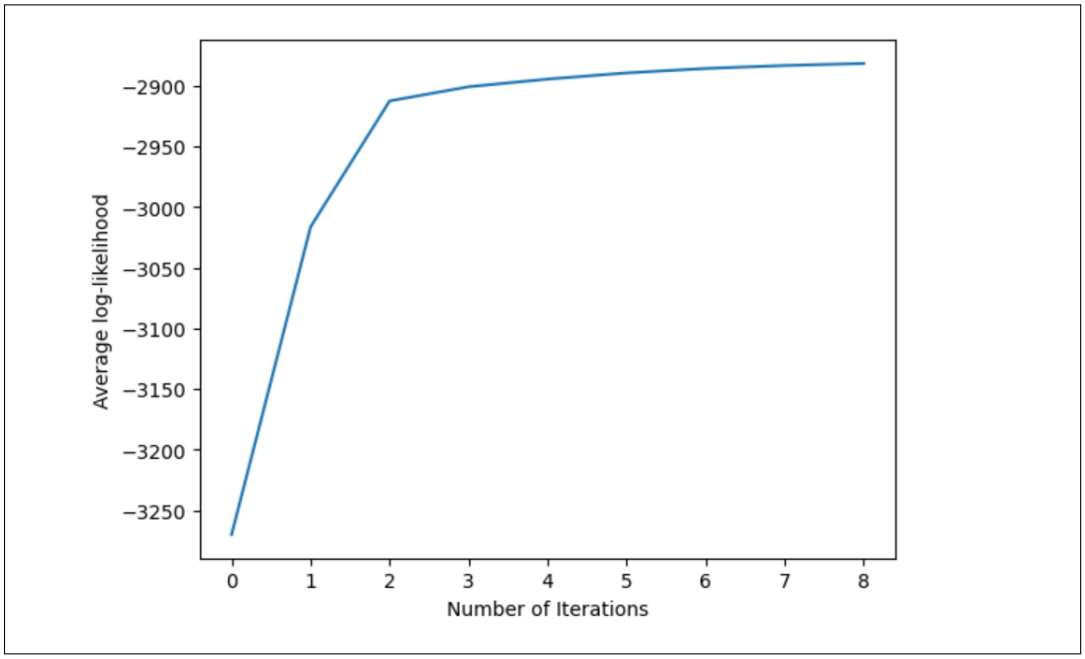
$$\pi_k = \frac{\sum_{i=1}^{n} \lambda_k^i}{n}$$

$$\lambda_k^i = \frac{\pi_k \prod_{d=1}^{50} (p_k^d)^{f_d^i} (1-p_k^d)^{1-f_d^i}}{\sum_{l=1}^{k} \pi_l \prod_{d=1}^{50} (p_k^d)^{f_d^i} (1-p_k^d)^{1-f_d^i}}$$



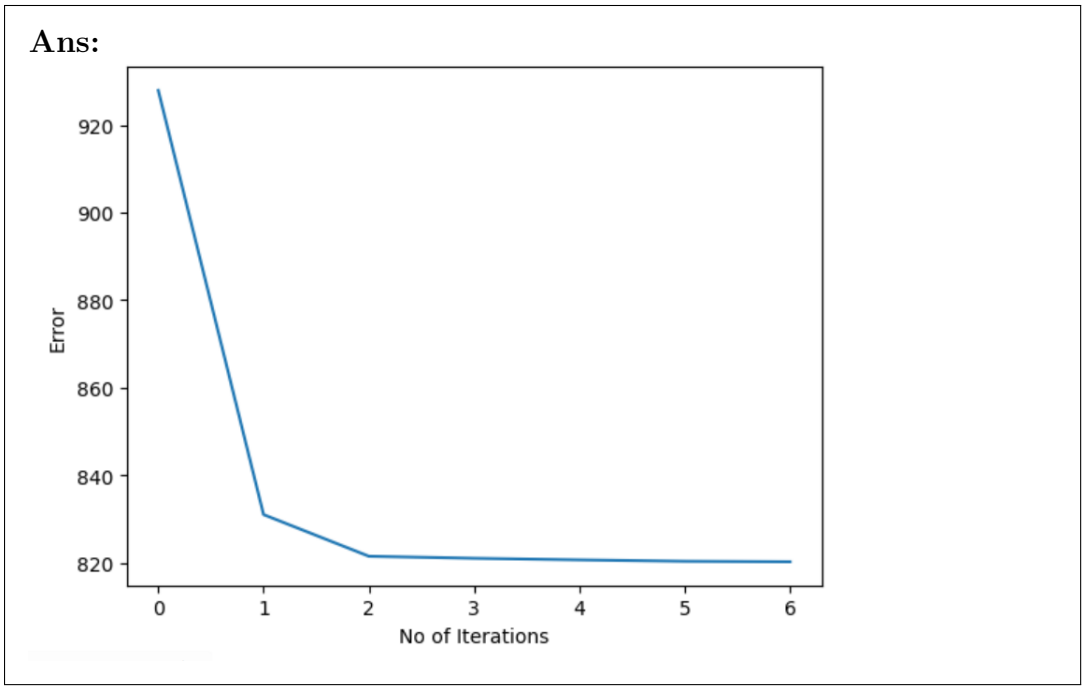
(b) Assume that the same data was in fact generated from a mixture of Gaussians with 4 mixtures. Implement the EM algorithm and plot the log-likelihood (averaged over 100 random initialization of the parameters) as a functions of iterations. How does the plot compare with the plot from part(i)?Provide insights that you draw from this experiment.

Ans: I found out that log-likelihood is increasing smooth in the case of multivariate Bernoulli's distribution with respect to no. of iterations whereas the the increase is not so smooth in multivariate Gaussian distribution with respect to number of iterations. Also, the no.of iterations needed for convergence in multivariate Bernoulli's distribution is around 50-60 whereas no.of iterations needed for convergence in multivariate Gaussian distribution is around 10.



(c) Run the K-means algorithm with K = 4 on the same data. Plot the objective of K - Means as a function of iterations.

Ans:



(d) Among the three different algorithms implemented above, Which do you think you would choose to fro this dataset and why?

Ans: We can compare the performance of EM algorithms by Convergence time and π obtained after convergence .The total time taken for convergence in Multivariate Bernoulli EM algorithm is less than Multivariate Gaussian EM algorithm because of time computing determinant and inverses in case of Multivariate Gaussian EM algorithm though the no of iterations in Multivariate Bernoulli EM algorithm is higher than Multivariate Gaussian EM algorithm.K means algorithm is trying to hard cluster the datapoints based on the features.

2. You are given a data-set in the file A2Q2Datatrain.csv with 10000 points in $(\mathbb{R}^{100}, \mathbb{R})$ (Each row corresponds to a datapoint where the first 100 components are features and the last component is the associated y value).

   (a) Obtain the least squares solution Wml to the regression problem using the analytical solution.

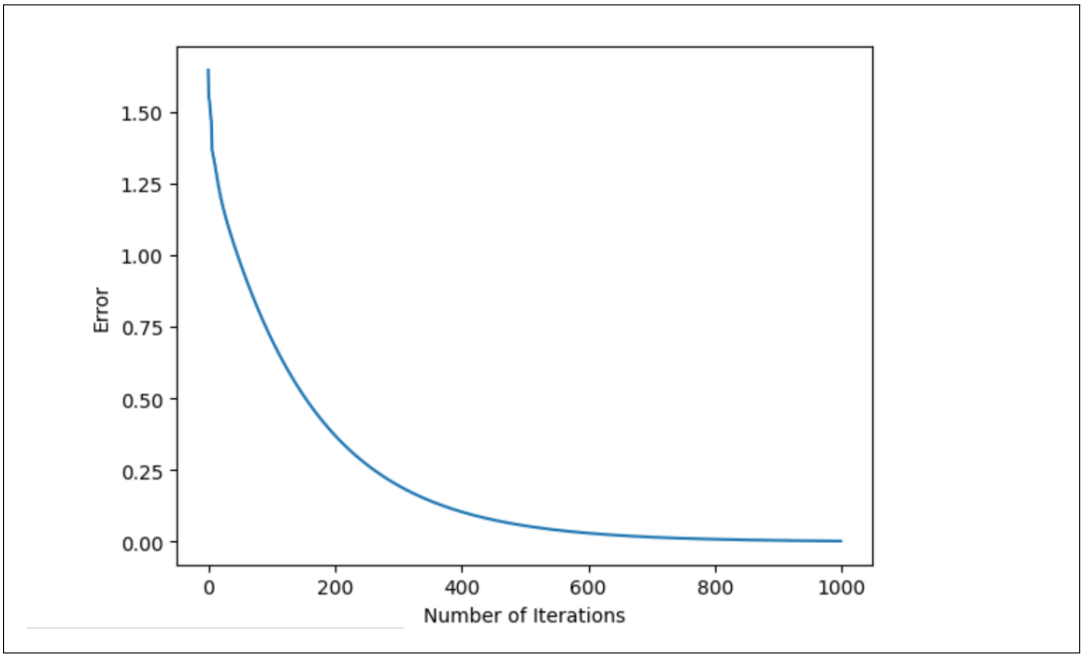
   Ans:

   - least Square error on training Data= 396.86441862
   - least Square error on testing Data= 185.363655584

   (b) Code the gradient descent algorithm with suitable step size to solve the least squares algorithm and plot $\|\mathbf{w}^t - \mathbf{w}_{ML}\|_2$ as a function of t. What do you observe?

   Ans: I observed from the plot that when W is initialized with 0,then it is taking around 1000 iterations to converge whereas when it is initialized randomly, then it is taking around 200000 iterations to converge.

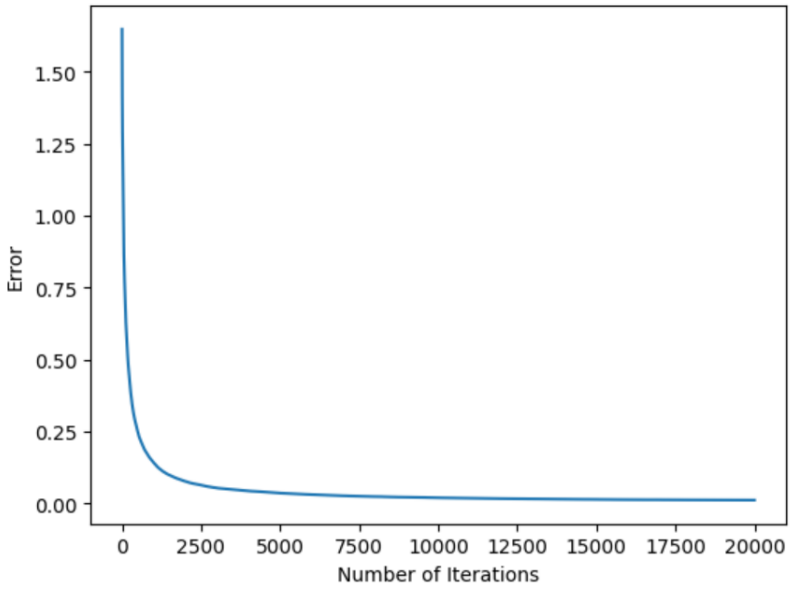
   - Least square error on training data=396.86441862
   - Least square error on training data=185.363655584



(c) Code the stochastic gradient descent algorithm using batch size of 100 and plot $\|\mathbf{w}^t - \mathbf{w}_{ML}\|_2$ as a function of t. What are you observations?

Ans: I observed that stochastic gradient descent is taking very less iterations to find optimal $\mathbf{w}_{stochastic}$ than gradient descent.Error on training data is higher in Stochastic gradient descent than Normal Gradient descent whereas Error on testing data is less in case of Stochastic gradient descent compared to normal gradient descent

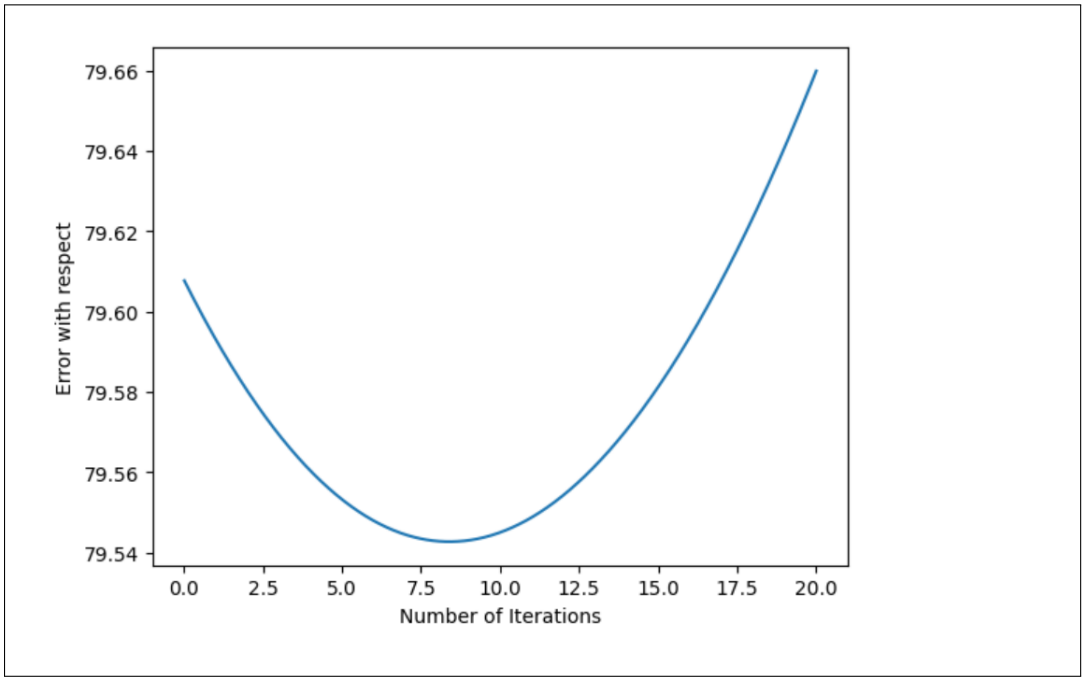
- Least square error on training data=398.54327613485856
- Least square error on testing data=180.43049667171624



(d) Code the gradient descent algorithm for ridge regression. Cross-Validate from various choices of λ and plot the error in the validation set as a function of λ. For the best λ chosen, obtain $\mathbf{w}_R$. Compare the test error (for the test data in the file A2Q2Datetest.csv) of $\mathbf{w}_R$ with $\mathbf{w}_{ML}$. Which is better and why?

Ans: I observed that least square error on training data is higher on $\mathbf{w}_R$ compared to $\mathbf{w}_{ML}$ whereas the least square error on test data is higher on $\mathbf{w}_{ML}$ compared to $\mathbf{w}_R$.
In my opinion, Ridge regression is better than maximum likelihood estimator because for unseen data,it is performing better.

- Minimum lambda = 8.39999999999991
- Least square error on training data=398.33408144455166
- Least square error on testing data=184.1186630577116


79.66
79.64
79.62
79.60
79.58
79.56
79.54
Error with respect
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Number of Iterations