CS 5691 : Pattern Recognition and Machine Learning

Assignment 2

Report

Santosh S R
ME20B157

Question 1

1) You are given a data-set with 400 data points in {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.)

(i) 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 initializations) as a function of iterations.

Answer:

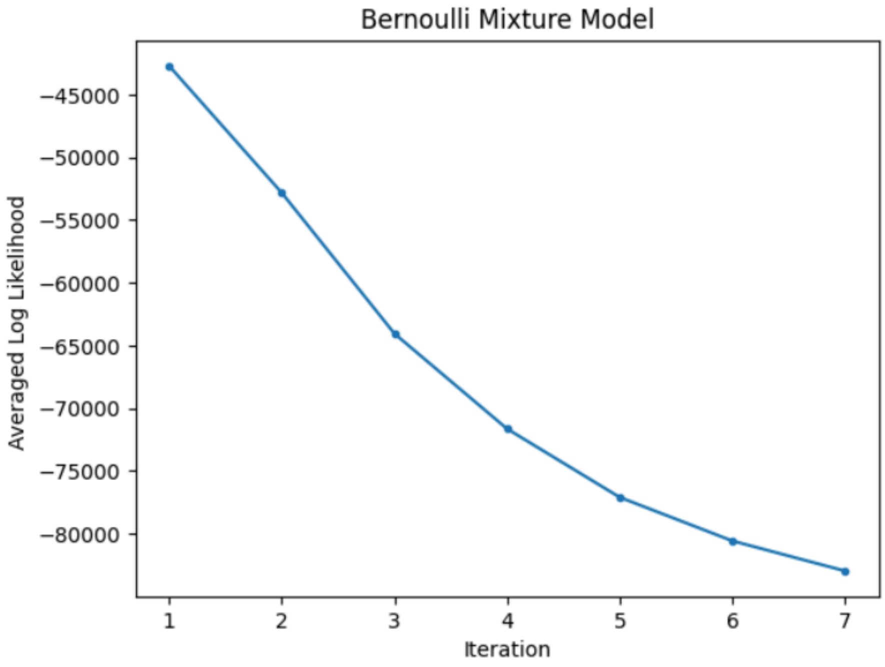
The probable mixture model that generated the data would be Bernoulli mixture model.

This is answer is backed by it being:

- Firstly all the data points or sub-data points are binary {0,1}
- The Convergence of the Bernoulli MM was the smoothest and greatest in comparison to the Gaussian MM and K-Means Methods.
- The Gaussian showed random convergence and Divergences.
- The K-Means algorithm showed very slow convergence.
- I can observe that each Cluster of the Bernoulli MM has a different probability distribution for the sub data points in the 10*5 binary

values, resulting in such a combination, so increasing the k to 50 may lead to an overfit of 50 independent Bernoulli distributions.

Plot: The Convergence of 100 Averaged EM Log Likelihood for Bernoulli Mixture Model.



Where the respective Log Likelihood was calculated by:

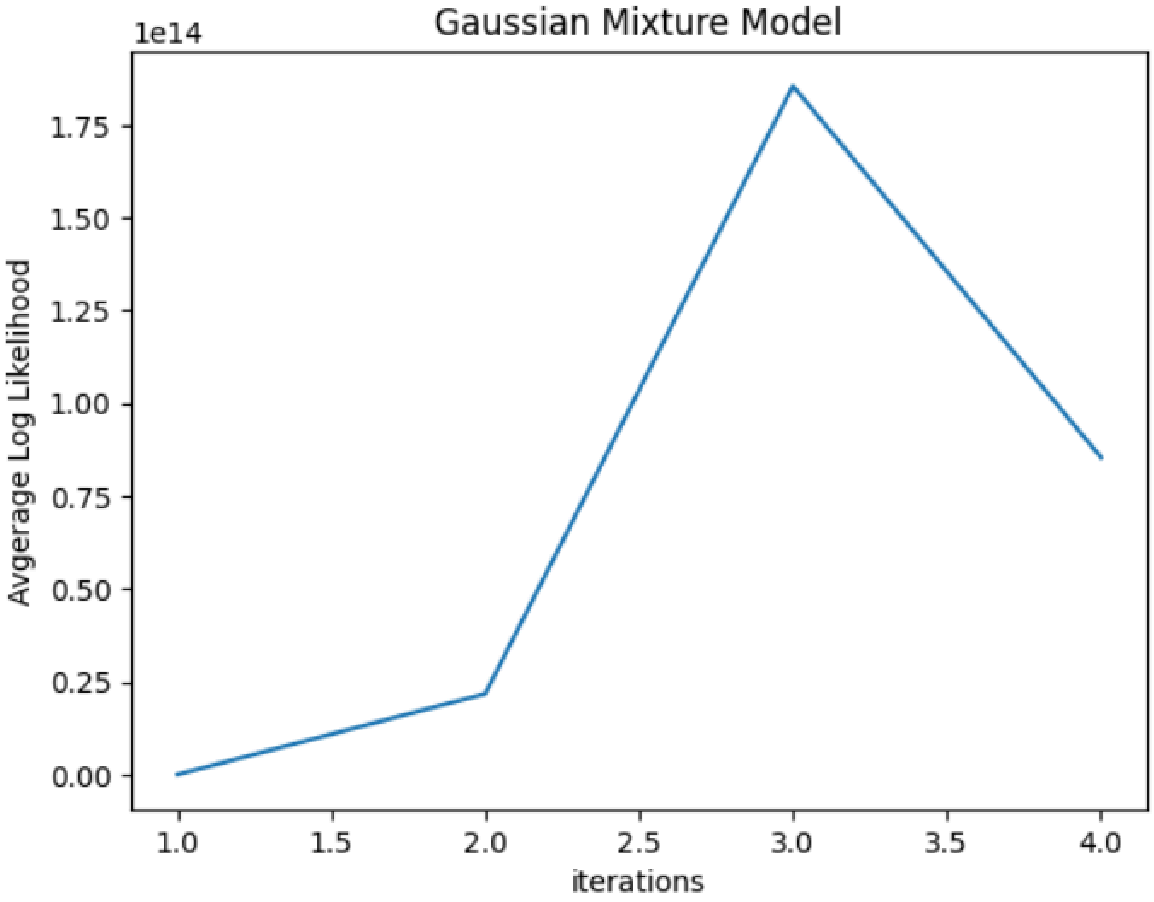
$$\begin{aligned}\mathcal{L}(\theta) &= \ln \Pr(\mathbf{X}, \mathbf{Z} | \mu, \pi) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{n,k} \left(\ln \pi_k + \sum_{i=1}^{D} x_{n,i} \ln \mu_{k,i} + (1 - x_{n,i}) \ln(1 - \mu_{k,i}) \right).\end{aligned}$$

(ii) 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 initializations of the parameters) as a function of iterations. How does the plot compare with the plot from part (i)? Provide insights that you draw from this experiment.

Answer:

Plot: The Convergence of 100 Averaged EM Log Likelihood for Gaussian Mixture Model.



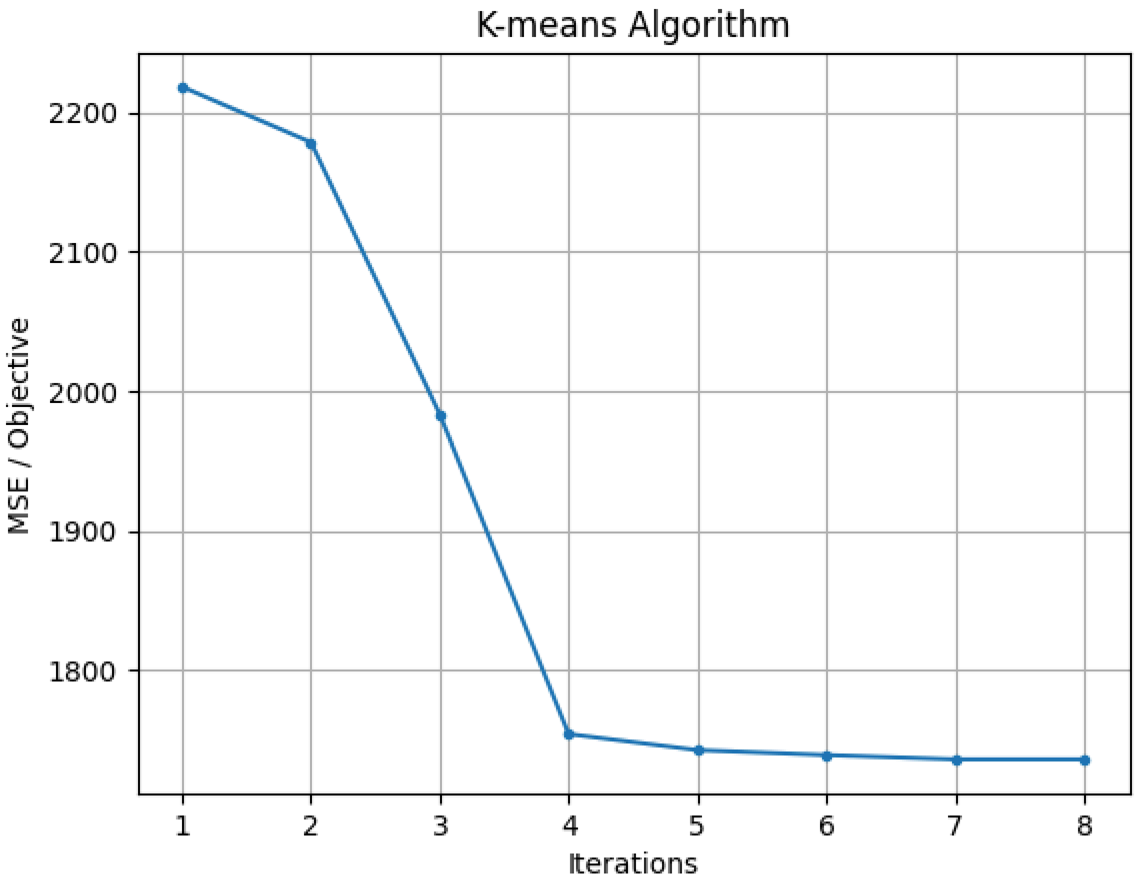
We can clearly see the inconsistent Convergence, also the time taken was exceedingly high so the iteration time was stopped at a manageable value.

The Mixture models are greatly influencing the convergence and time taken to Iterate the EM algorithm

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

Answer:

Plot: The Mean Squared Error / Objective Convergence with respect to the number of iterations

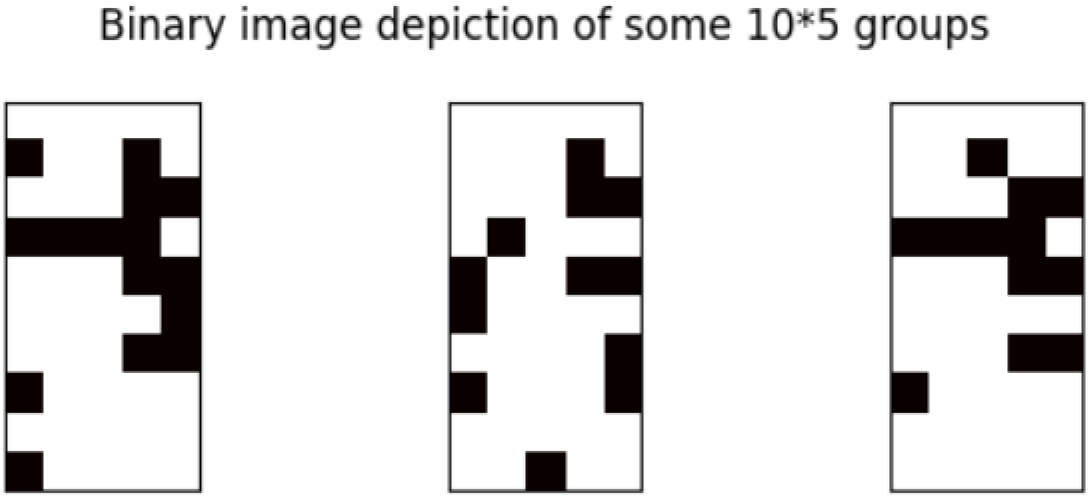


Clearly we can see an Convergence, but the convergence rate is very slow once it gets past the 4th iteration, that is once it gets past the part where random initialisation error is done with.

(iv) Among the three different algorithms implemented above, which do you think you would choose to for this dataset and why?

Answer:
I would go for the Bernoulli Mixture models as I am convinced [mentioned why in part answer (i)] of it being the actual best Generator for this data set given that the actual number of models mixed [K] is given.

I am unable to find if the data point is actually a simple image like:



or something more complex.
So with this information I would opt to use EM Algorithm with Bernoulli mixture model and k = 4 Assumption for this dataset.

Question 2

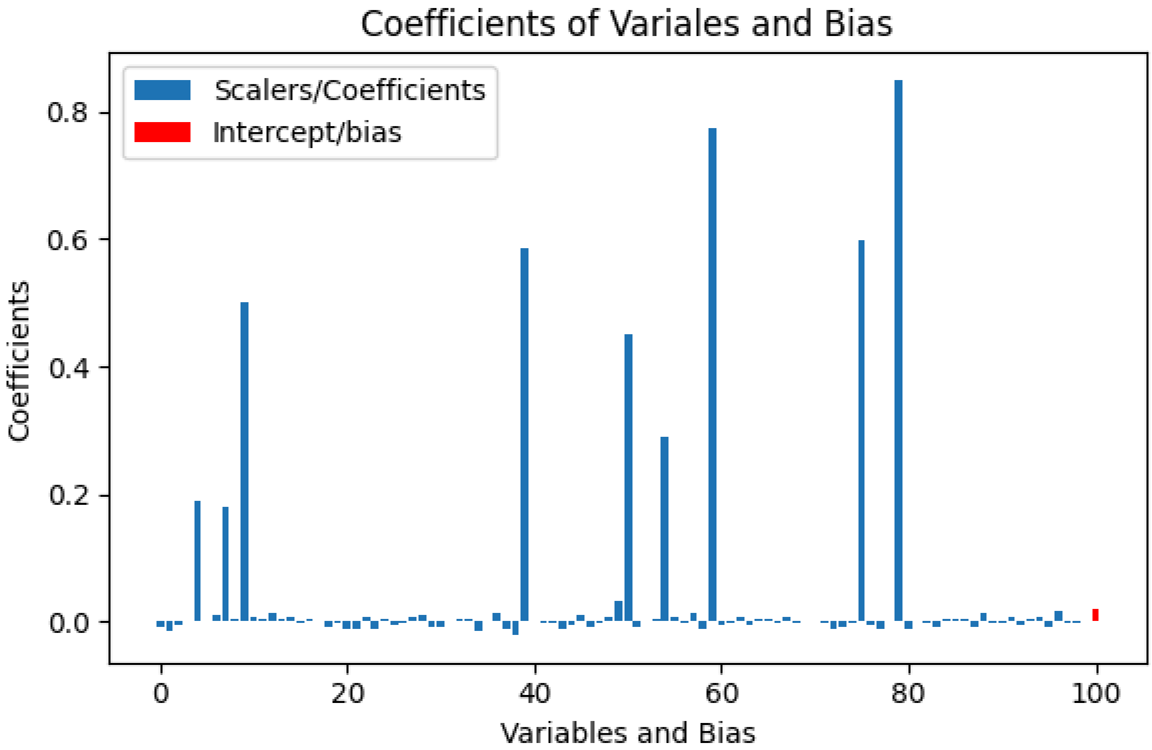
You are given a data-set in the file A2Q2 Data train.csv with 10000 points in (R 100 , R) (Each row corresponds to a datapoint where the first 100 components are features and the last component is the associated y value).

(i) Obtain the least squares solution w_{ML} to the regression problem using the analytical solution.

Answer:

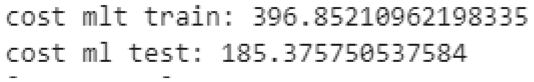
Analytical solution: $w_{ML} = (X^T X)^{-1} X^T y$

Plot: variable index wise Coefficient bar graph and final being bias



We can see that the bias is not relatively scalable to the important variables' Coefficients, so the line approximately is centered.

The custom error data of the analytical solution:



```
cost mlt train: 396.85210962198335
cost ml test: 185.375750537584
```

(ii) Code the gradient descent algorithm with suitable step size / optimal to solve the least squares algorithms and plot $||w^t - w_{ML|}||2$ as a function of t.What do you observe?

Answer:

Gradient Descent algorithm:

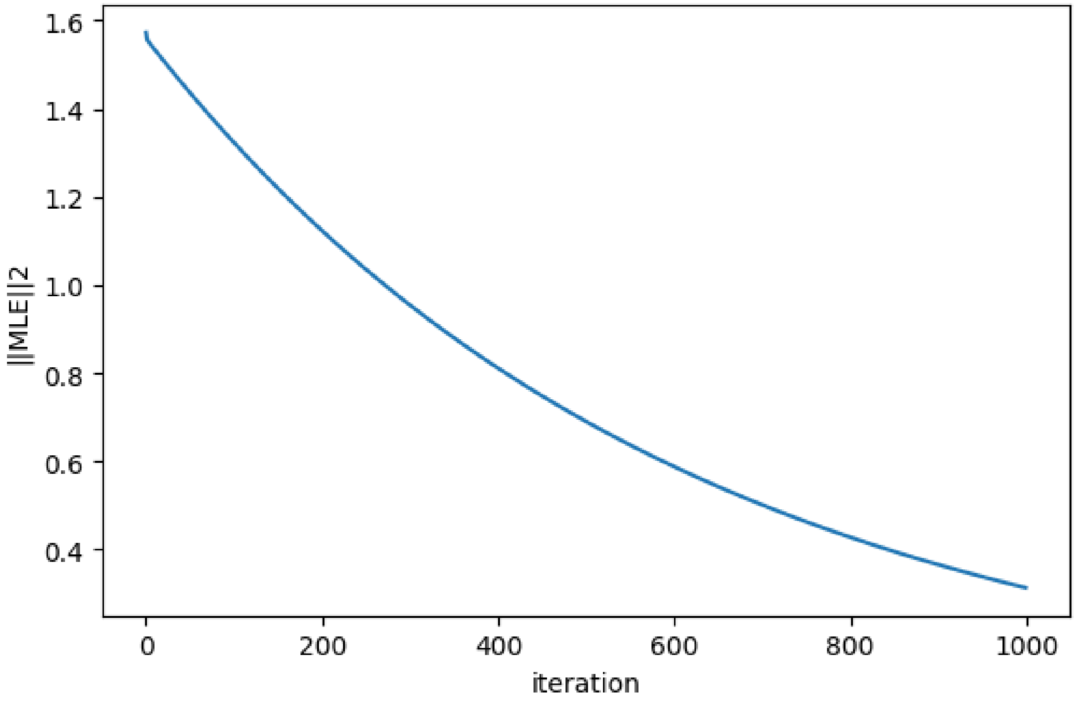
Plots: Gradient Descent

Plot1 - MLE

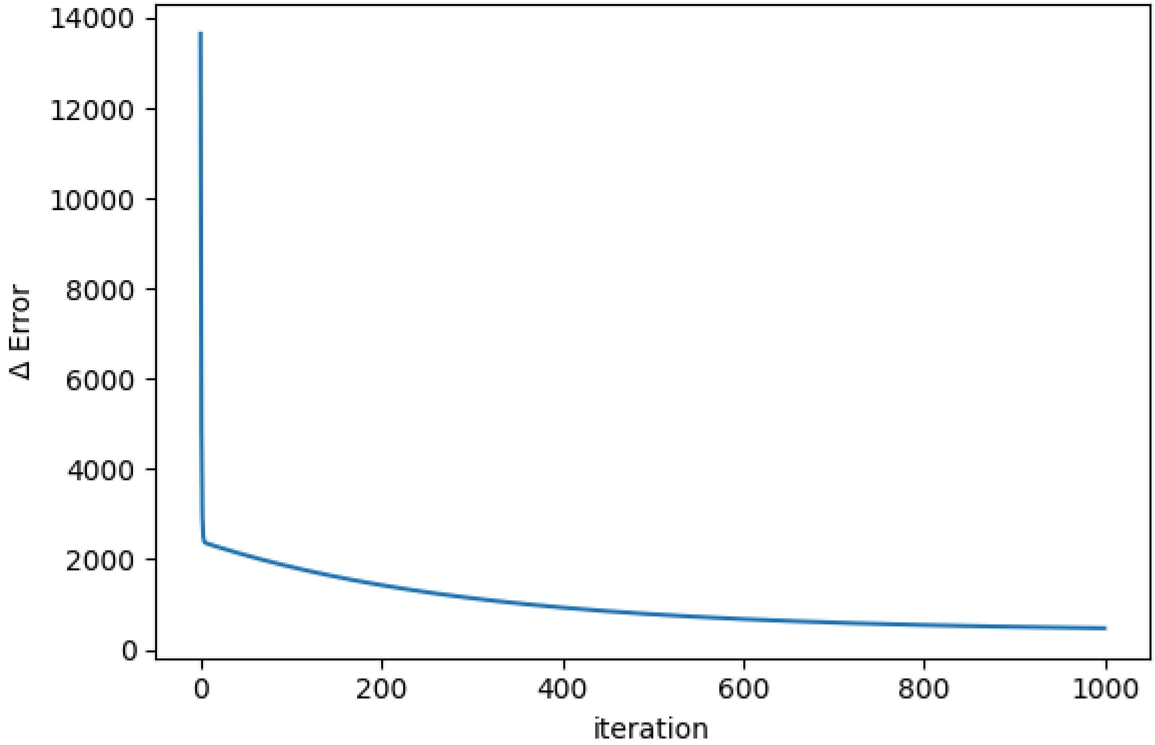
Plot2 - ∆ Error With respect to Iteration counter.

Plot 1 :

:



Plot 2 :



Iterator: $w^{t+1} = w^t - \frac{2\alpha}{n}((X^TX)w^t - X^Ty)$ \\ 2α/n → Step size

(iii) Code the stochastic gradient descent algorithm using batch size of 100 and plot $||w^t - w_{ML|}||2$ as a function of t. What are your observations?

Answer:

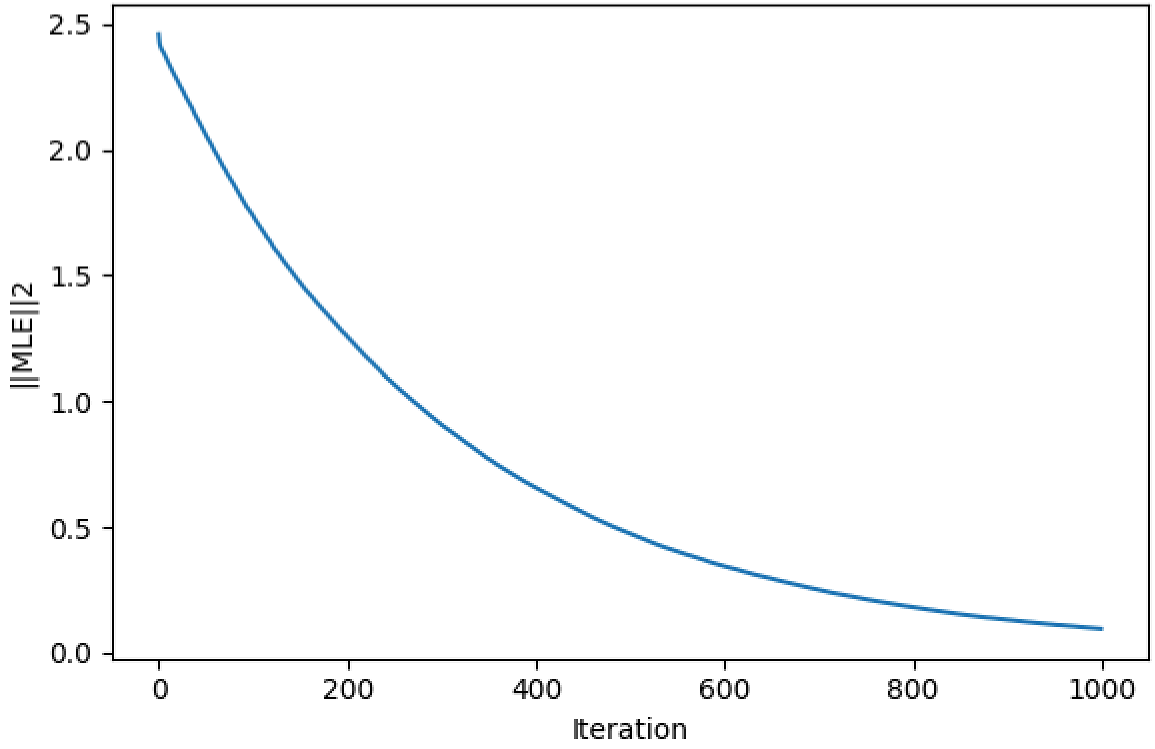
New Iterator: $w^{t+1} = w^t - \frac{2\alpha}{n}((\tilde{X}^T\tilde{X})w^t - \tilde{X}^Ty$

Plots: Stochastic Gradient Descent

Plot1 - MLE

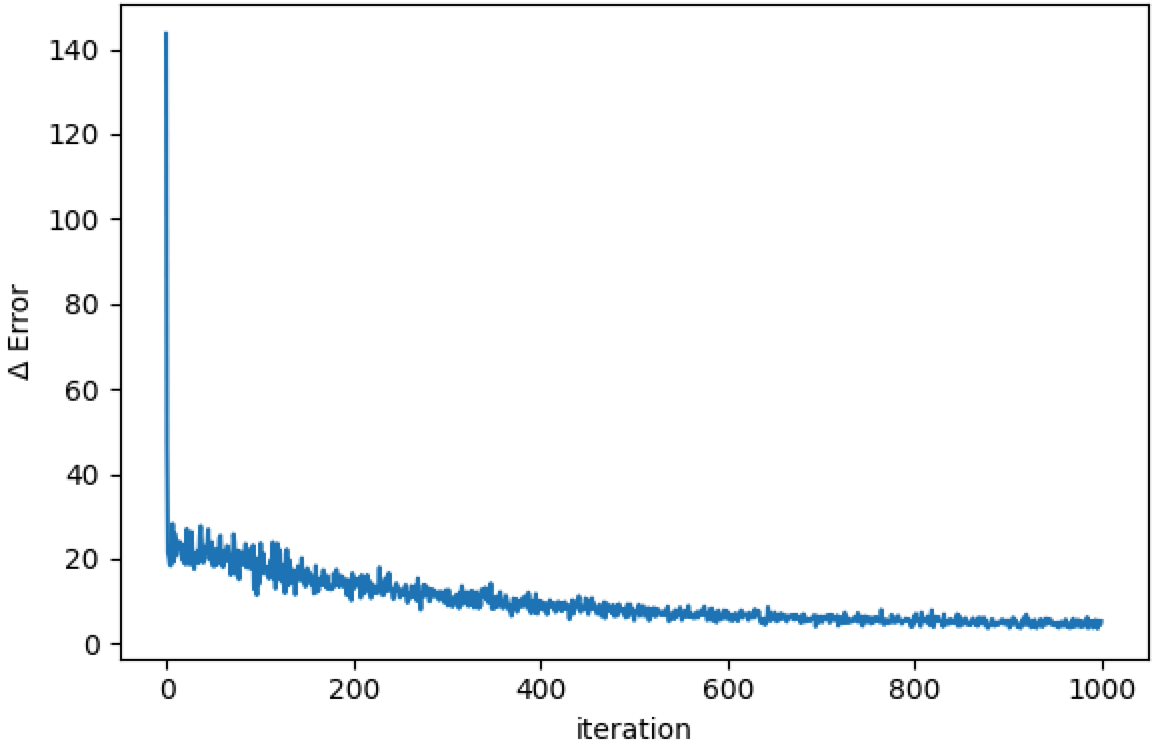
Plot2 - Δ Error With respect to Iteration counter.

Plot 1:



2.5
2.0
1.5
1.0
0.5
0.0
||MLE||2
0
200
400
600
800
1000
Iteration

Plot 2:



140
120
100
80
60
40
20
0
Δ Error
0
200
400
600
800
1000
iteration

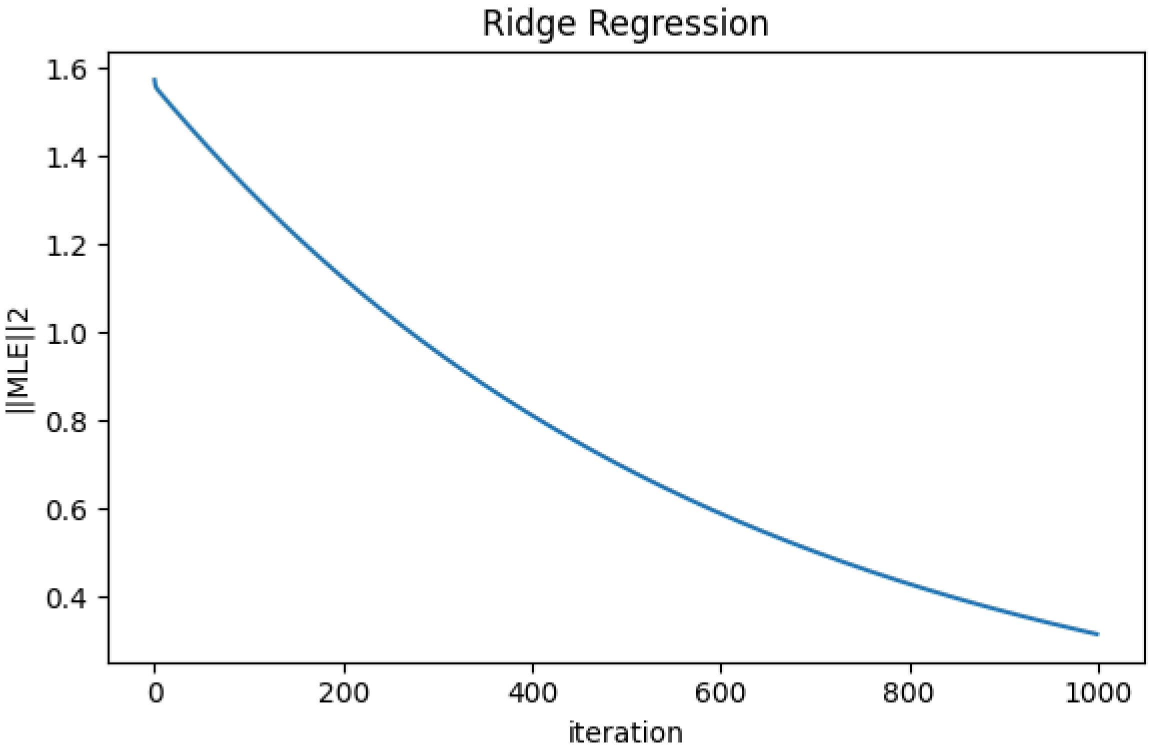
We can see the ripple-like deviation in the second graph, that is brought to us by the stochastic choice of steps.

(iv) Code the gradient descent algorithm for ridge regression. Cross-validate for various choices of λ and plot the error in the validation set as a function of λ. For the best λ chosen, obtain w_R. Compare the test error (for the test data in the file A2Q2Data test.csv) of w_R with w_{ML}. Which is better and why?

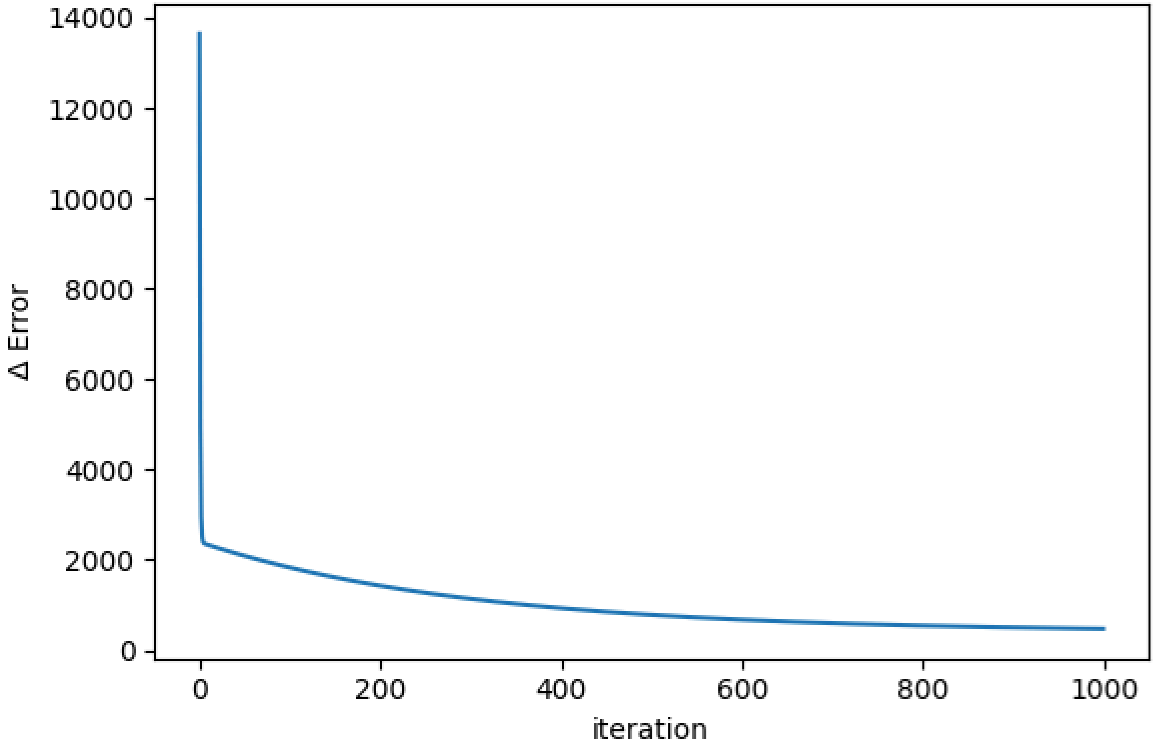
New Iterator: $w^{t+1} = w^t - \frac{2\alpha}{n}\left(\left(X^TX + \lambda I\right)w^t - X^Ty\right)$

Plots: Ridge Regression - Gradient Descent
Plot 1 - MLE
Plot 2 - Δ Error With respect to Iteration counter.

Plot 1:



Plot 2:

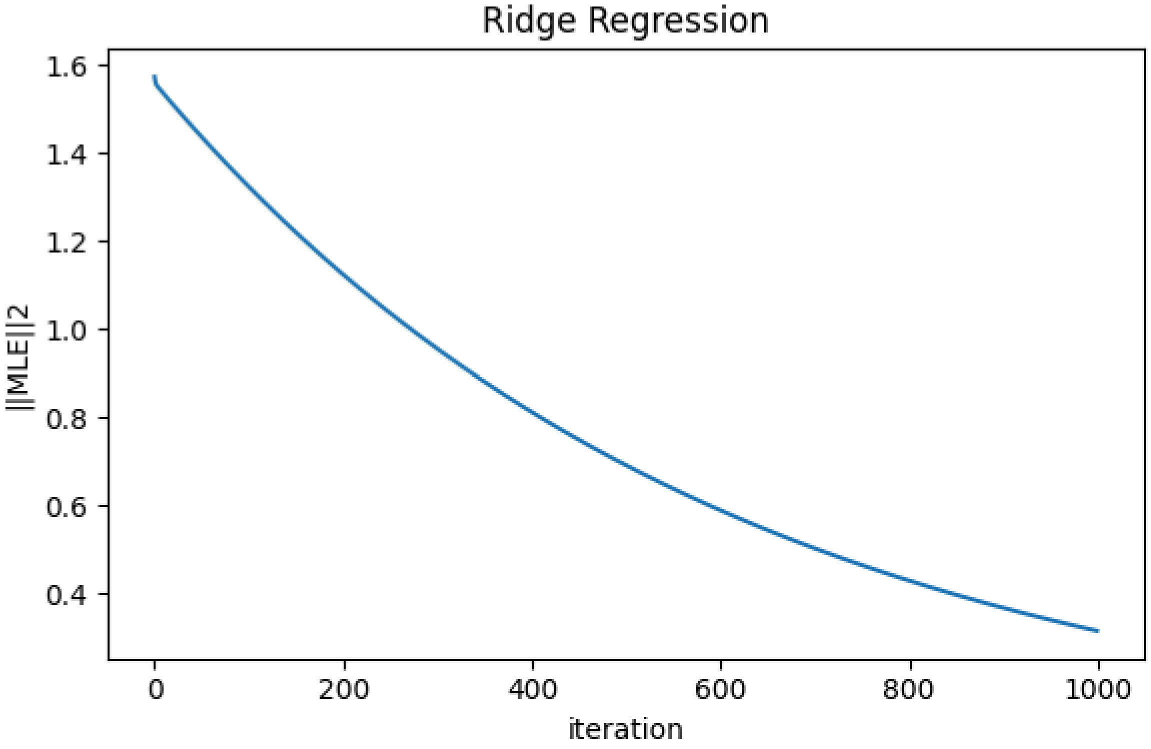


Cross Validation using K-Fold Method:

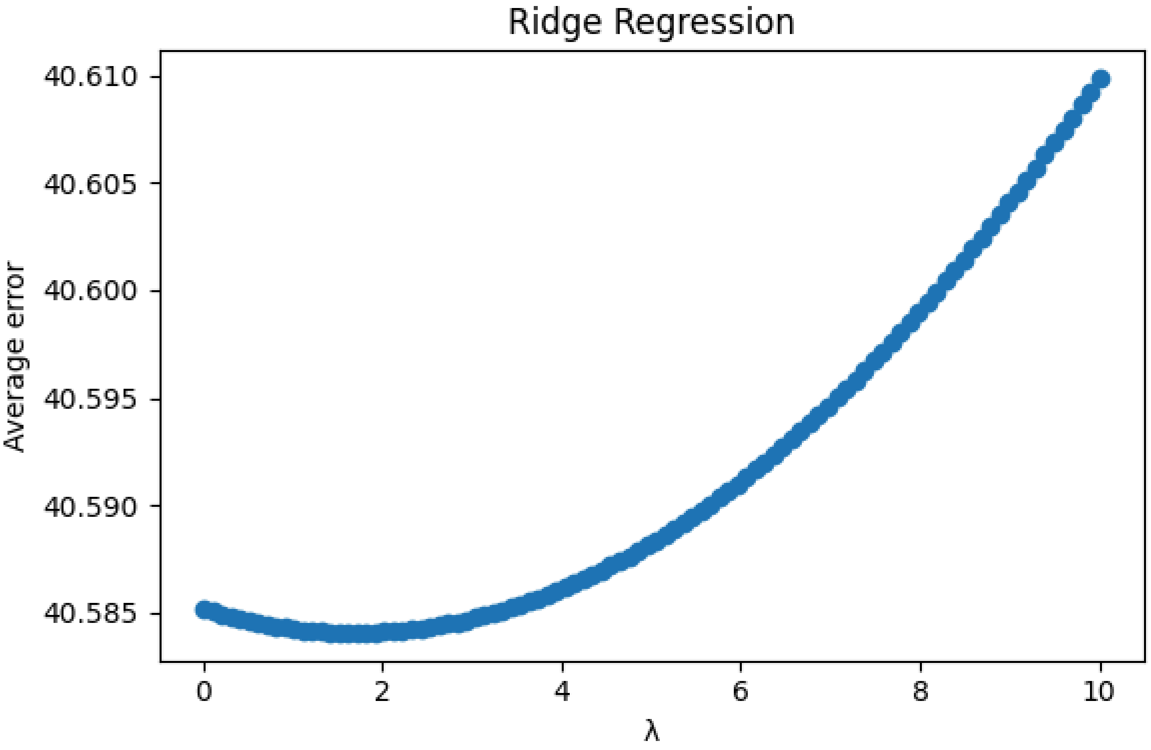
New Iterator: $w = \left(X^T X + \lambda I\right)^{-1} X^T y$

Plots: K - Fold Cross Validation.

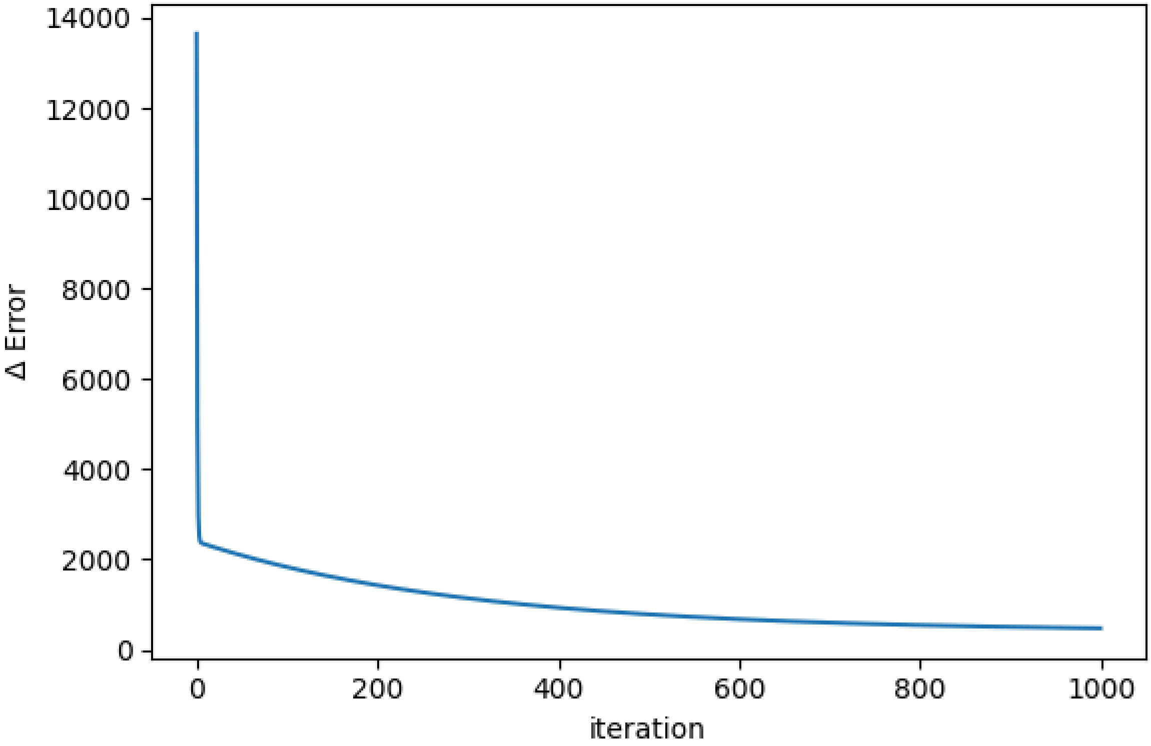
Plot 1 - Average Error of 100 groups



Plot 2 - MLE



Plot 3 - ∆ Error With respect to Iteration counter.



10-fold cross-validation to find the best lambda (λ) for a linear regression model. The data was split into 10 folds, with 9 for training and 1 for validation in each fold. We compared validation errors across different lambdas and picked the one with the lowest average error. This approach helps prevent overfitting and choose the best model complexity.

λ_{Best} was found to be : 1.725454

and W_R test was = 155.4135420

Observations:

Method	**Train Error**
Analytical	396.858
Steepest Gradient Descent	472.512
Stochastic Gradient descent	477.402
Ridge Regression	473.436
Ridge Regression - Cross Validated	473.298

Method	**Test Error**
Analytical	185.375
Steepest Gradient Descent	155.553
Stochastic Gradient descent	155.540
Ridge Regression	155.489
Ridge Regression - Cross Validated	155.413

We can clearly see that Ridge Regression gives best results consistently.