Indian Institute of Technology, Madras
CS5691: Pattern recognition and Machine Learning
PRML Assignment-II Report

Ashutosh Kakadiya - CS18S013,
Rajan Kumar Soni - CS18S038

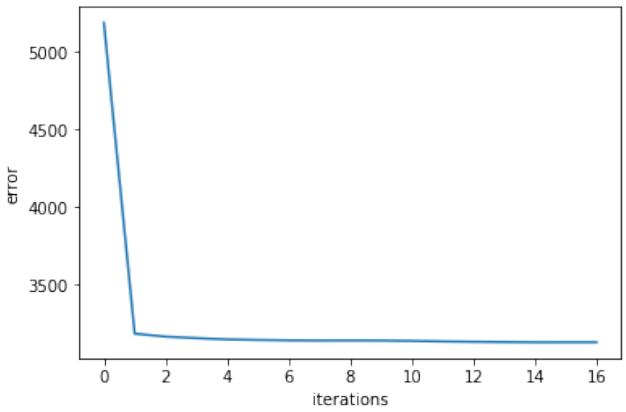
October 2019

Contents

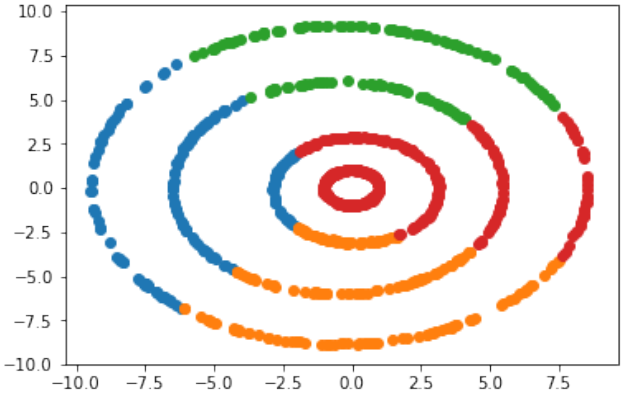
Chapter 1

K Means

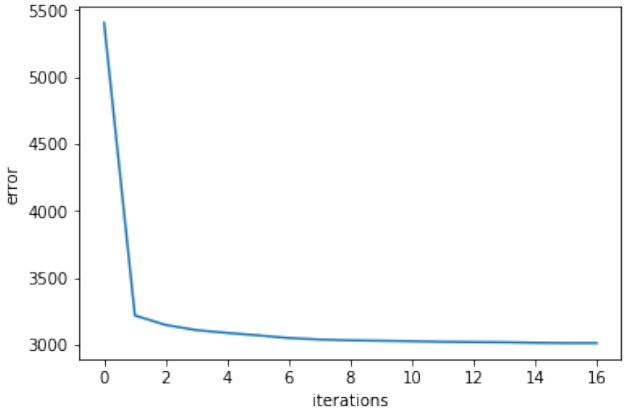
1. **Error plot and Cluster Plot for different Initializations:**



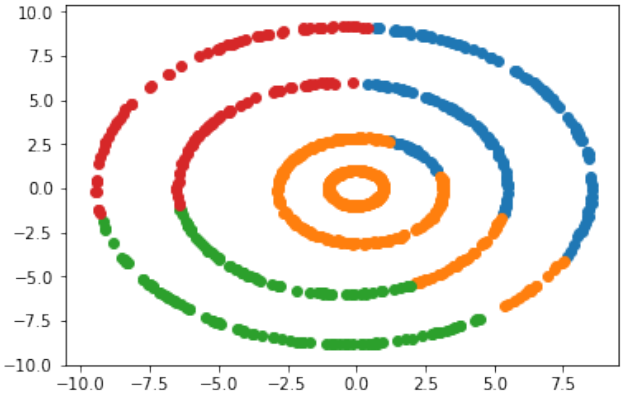
(a) error plot



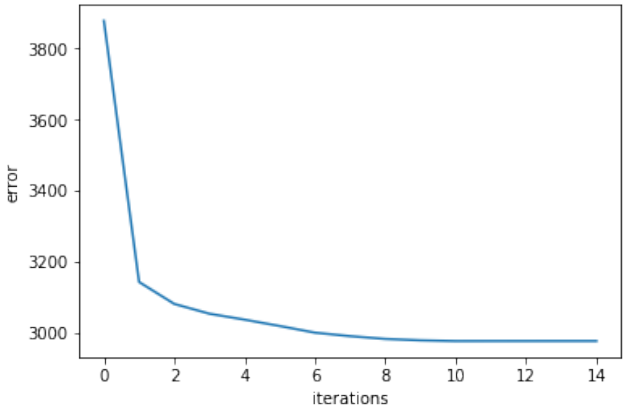
(b) cluster plot



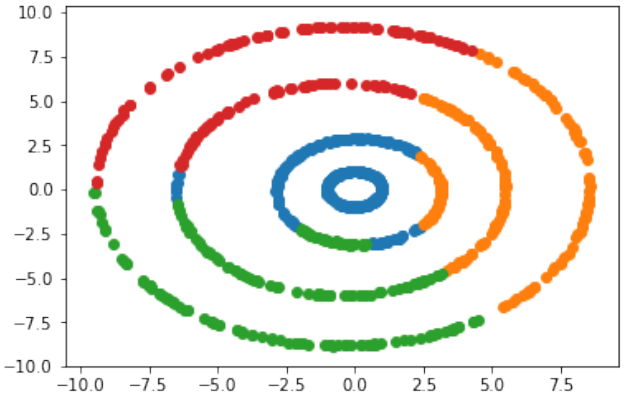
(a) error plot



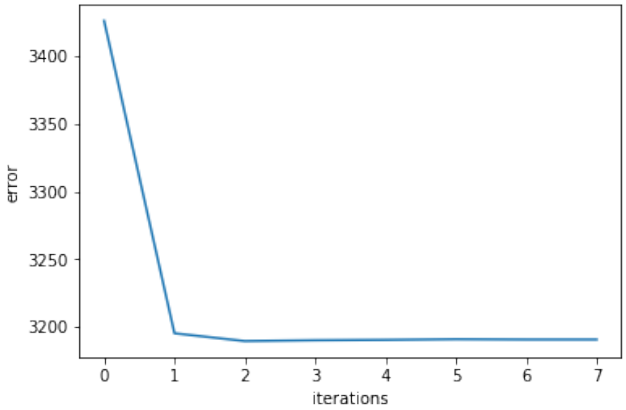
(b) cluster plot



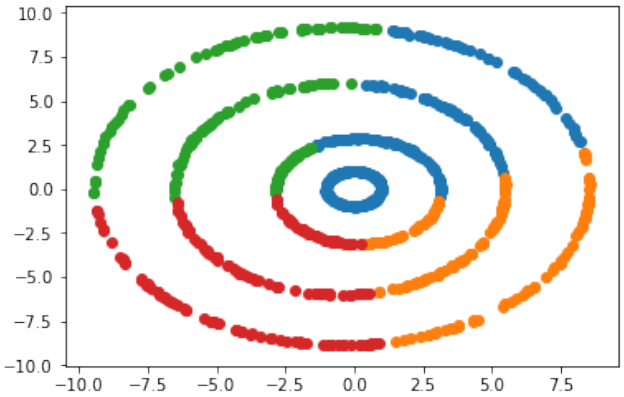
(a) error plot



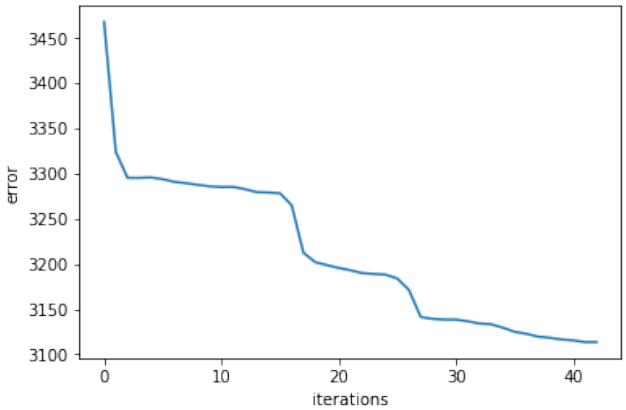
(b) cluster plot



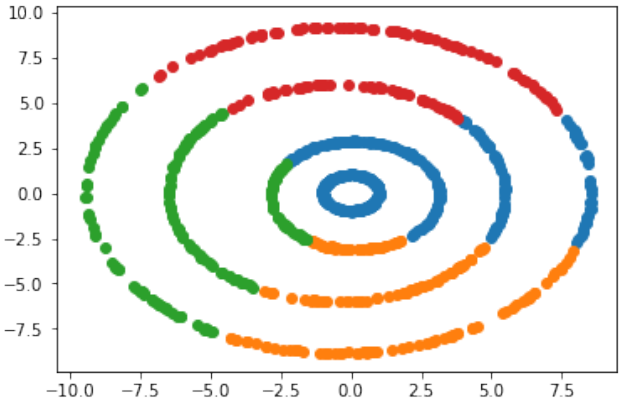
(a) error plot



(b) cluster plot



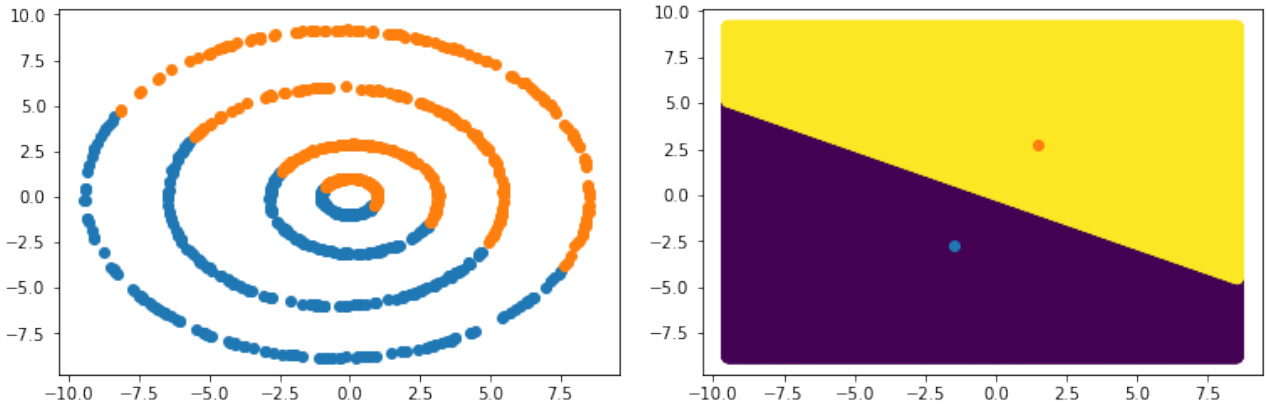
(a) error plot



(b) cluster plot

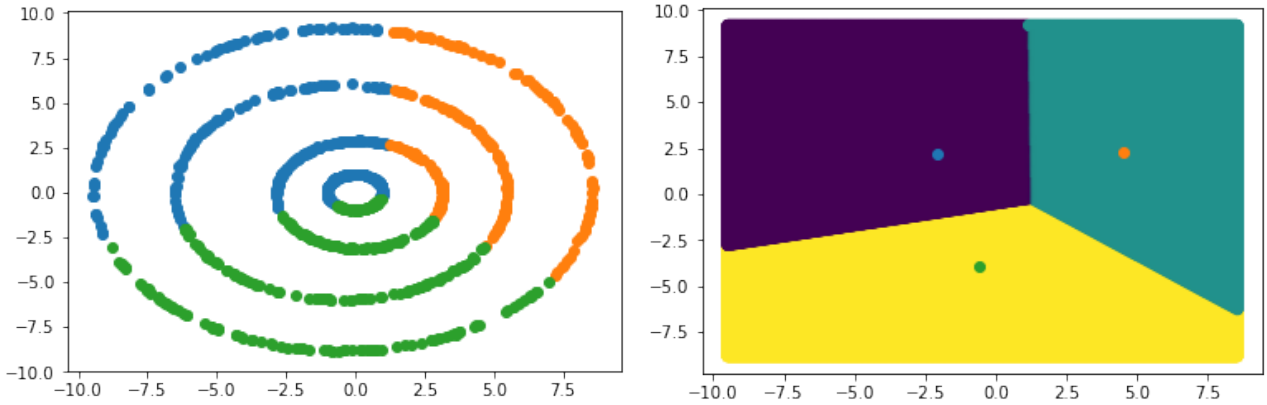
As we can observe different initialization leads to different convergence point and having different convergence rate. it shows that there exist more than one local minima in error surface

2. **Cluster plot and Voronoi plot:**



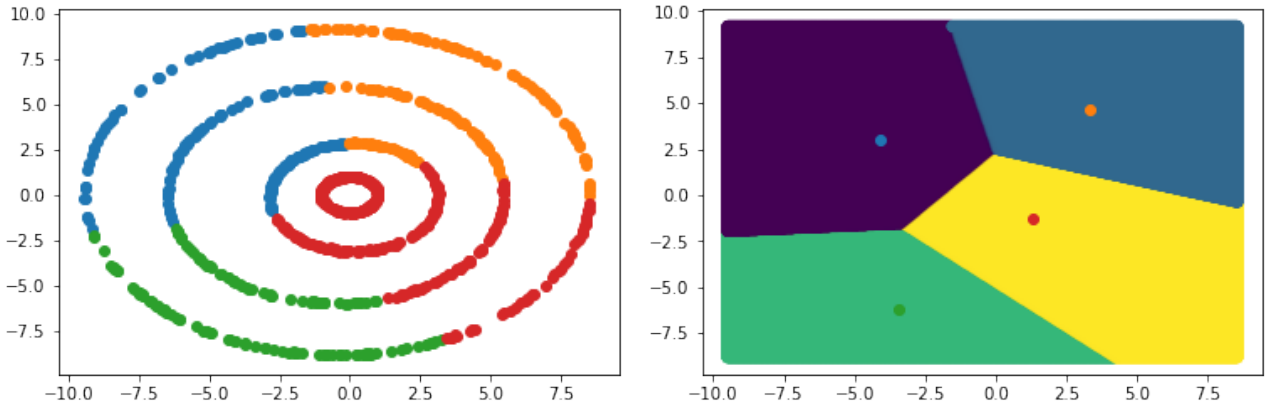
(a) cluster plot (b) voronoi plot

Figure 1.6: plot for K=2



(a) cluster plot (b) voronoi plot

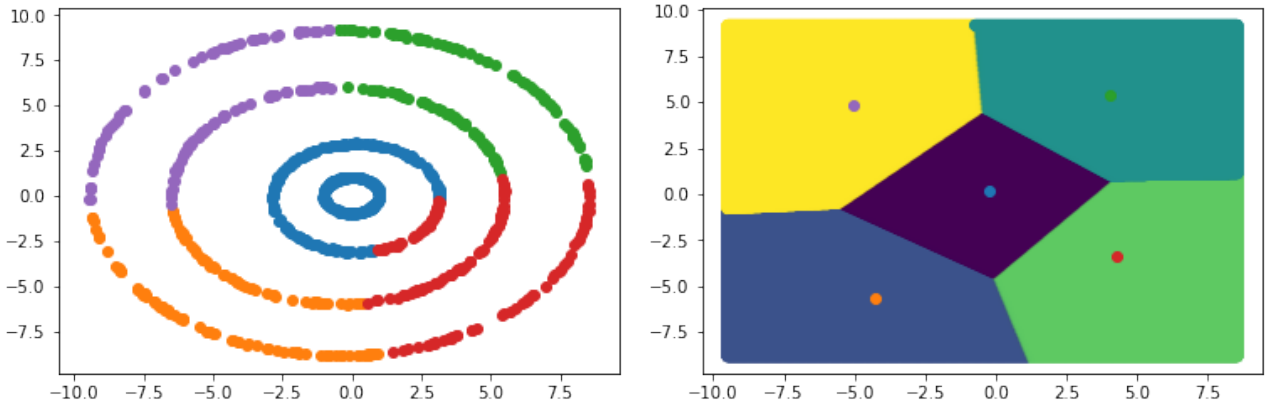
Figure 1.7: plot for K=3



(a) cluster plot

(b) voronoi plot

Figure 1.8: plot for K=4



(a) cluster plot

(b) voronoi plot

Figure 1.9: plot for K=5

3. **Spectral Clustering :**

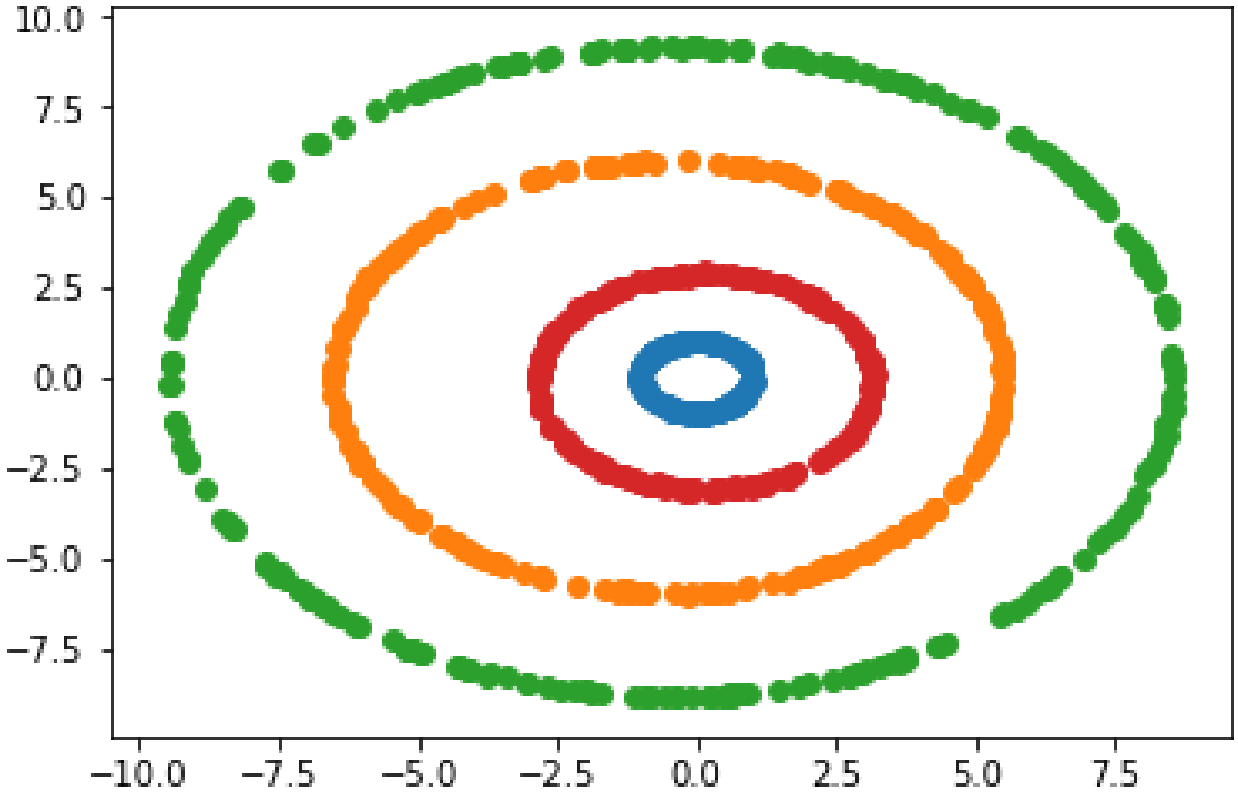


Figure 1.10: cluster plot

We have implemented **RBF** kernel. we can see data points can be clustered in 4 clusters. As rbf is distance based kernal. It measures the distance between two points. So points which are closer to each other comes in one cluster. Giving required value of K we can find distinguished clusters.Because rbf projects the vector into infinite dimension feature space where data points are separable. Tuning the parameter gamma we can find good cluster. For this gamma = .4 gives good result.

Chapter 2

Gradient Descent

1. **Closed Form Solution W_{ML} :**
   Closed form solution is obtained by : $(X^T X)^{-1} X^T Y$
2. **Gradient Descent plot $||\mathbf{w^t} - \mathbf{w_{ML}}||$:**

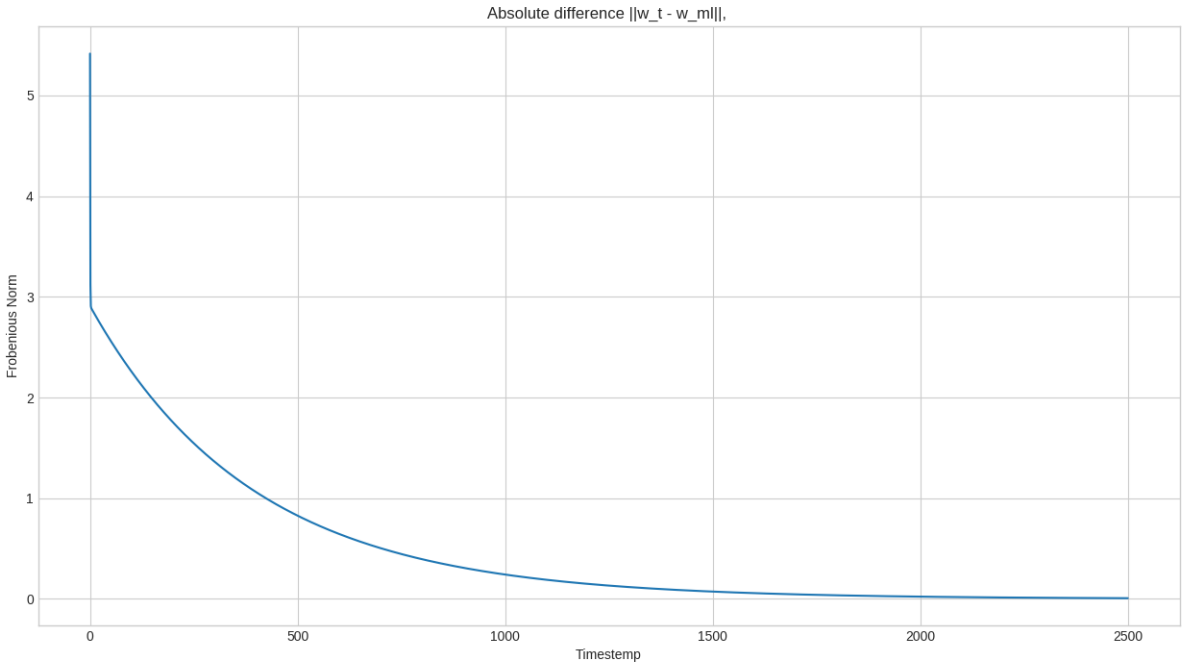


Figure 2.1: Log Likelihood plot of data with respect to number of clusters, step size :0.01, Epochs:2500

3. **Stochastic Gradient Descent plot $||\mathbf{w^t} - \mathbf{w_{ML}}||$:**

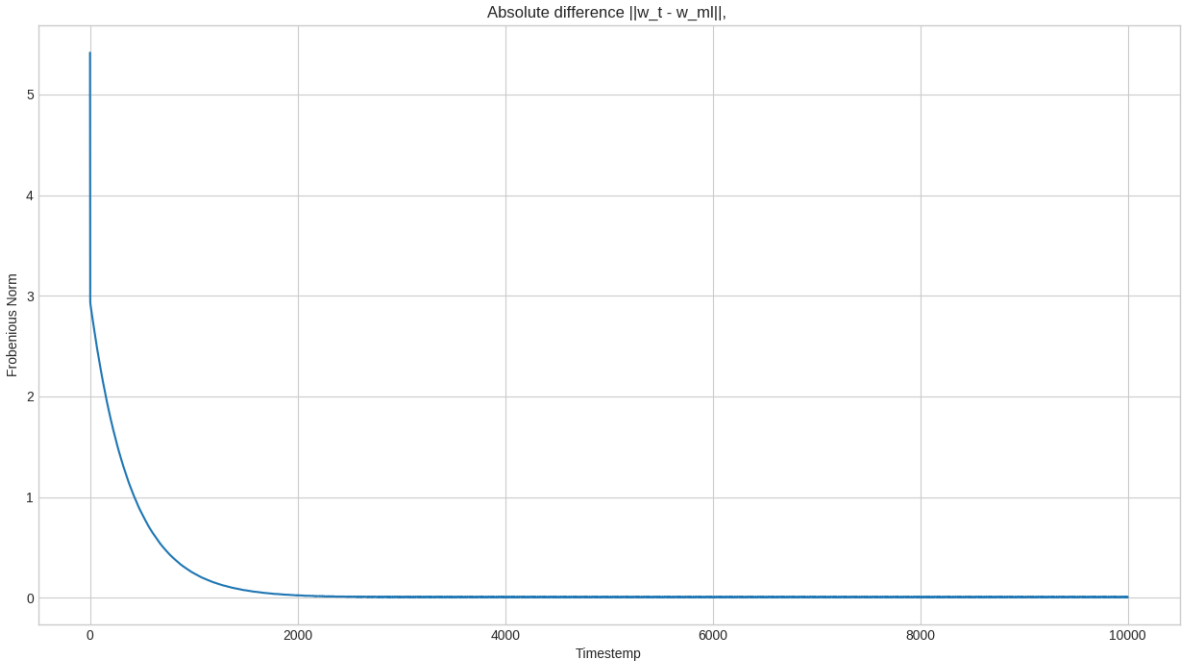


Figure 2.2: Log Likelihood plot of data with respect to number of clusters, step size :0.01, Epochs:100

4. **Observations :** As, times-temps increases the values of W_t are approaching the values of W_{ML}, the closed form solution. Because it reaches to global optima in both of the cases.
   Stochastic Gradient descent take less epochs compare to vanilla gradient descent to converge. Because it provides stable learning path in attain global optima(in this case) and computationally effective a update happen after accumulating total average of a batch instead of updating over whole data set! So, SGD generalize faster than vanilla Gradient Descent.

Chapter 3

Regularized Regression

3.1 Ridge Regression

1. **Cross-validation plot for different λ :**
   Best value of $\lambda = 0.7$

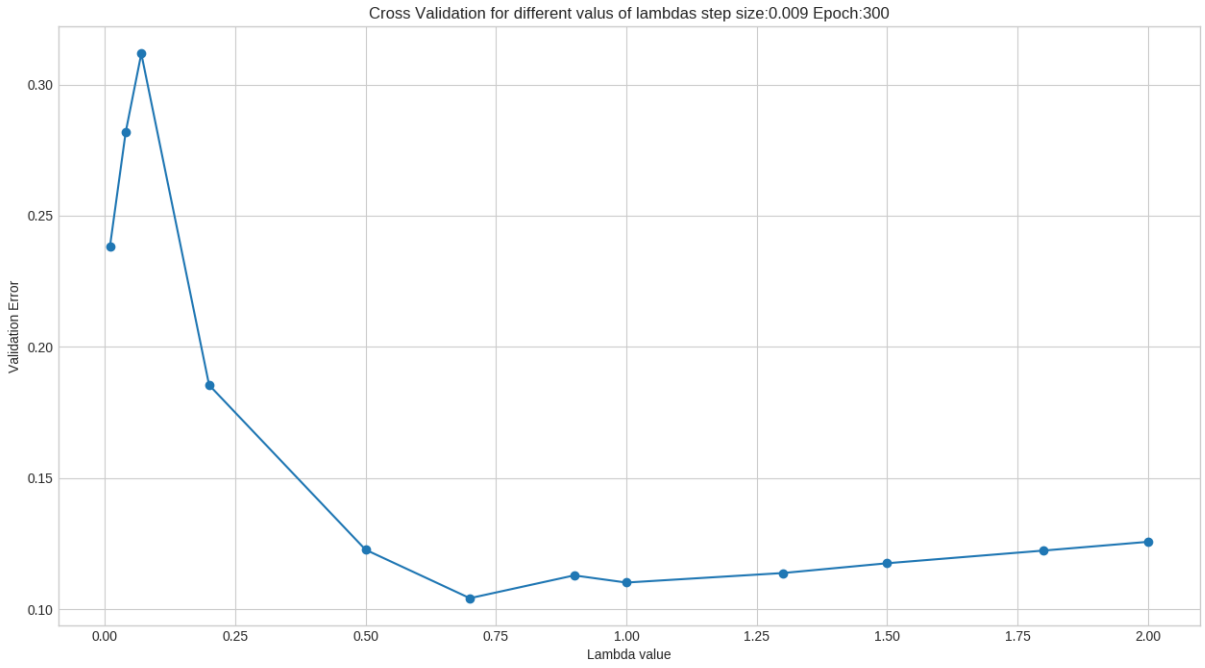


Figure 3.1: Cross Validation for best λ value

2. **Test Error Comparison for best parameter configuration:**
   - W_R : 11.17%
   - W_{ML} : 18.32%
3. W_R is better than W_{ML} as shown above. Main reason is closed form solution is completely fit the distribution of training data and very sensitive to the the data which is different than training data. So, high variance and low bias. In simple terms, over-fits the training data
   When we add regularization, it add bias, so reducing variance when features are and try to avoid over fit on training data.

3.2 Lasso Regression

1. **Cross-validation plot for different λ :** Best value of $\lambda = 0.007$ Validation Error : 2.01%

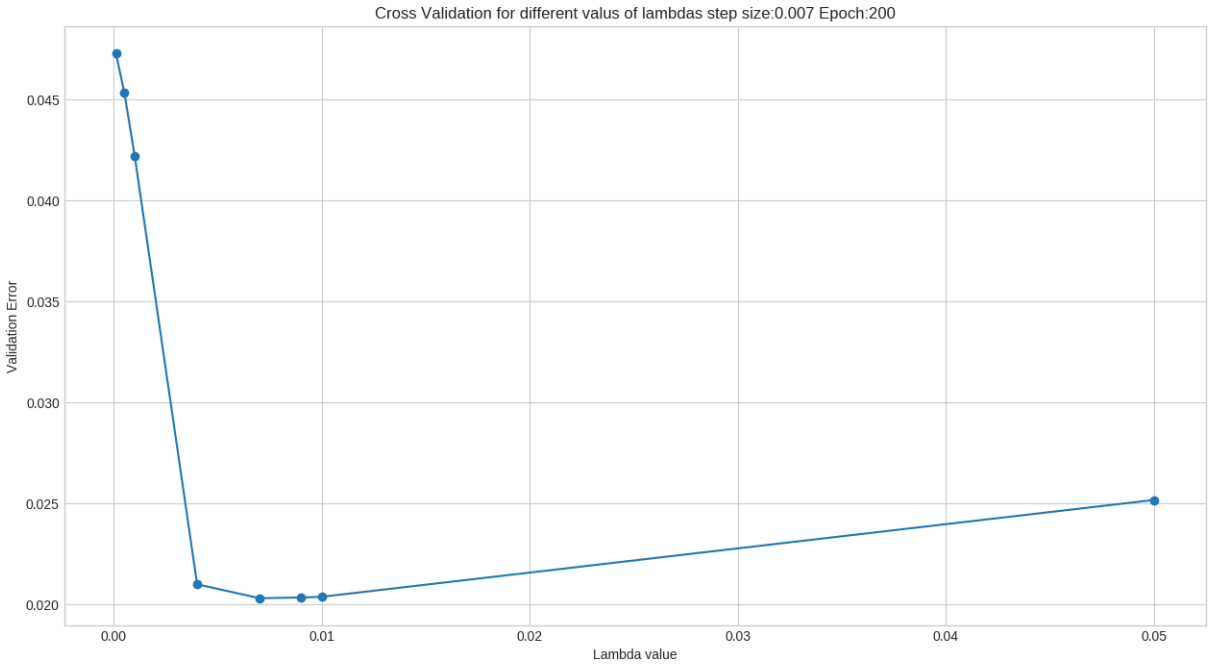


Figure 3.2: Cross Validation for best λ value

2. **Test Error Comparison :**
   - W_{LASSO} : 18.057%
   - W_{ML} : 18.58%
   - W_R : 11.17%
3. As described above, Ridge regression gives less error on test data than lasso regression. We have found the best hyper parameters using cross-validation for both methods. Both shrinkage the value of weights and tackle over fitting.

   Both methods has its pros-cones. Ridge regression is computationally efficient than Lasso. Lasso reduces the unnecessary parameters by making coefficient values to zeros.

Bibliography

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] C. S. Ong M. P. Deisenroth, A. A. Faisal. *Mathematics for Machine Learning*. Cambridge University, 2019.