COL774: Machine Learning- Assignment 1

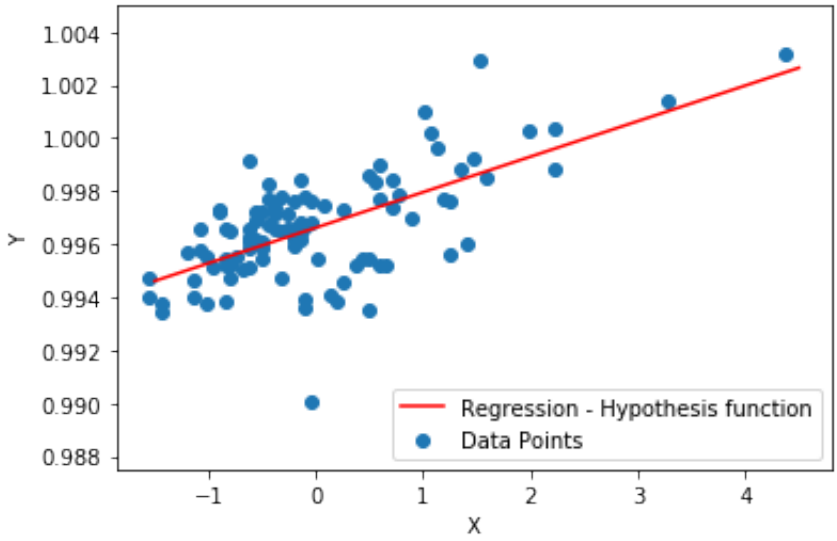
Rajat Jaiswal (2017CS50415)

10th February 2020

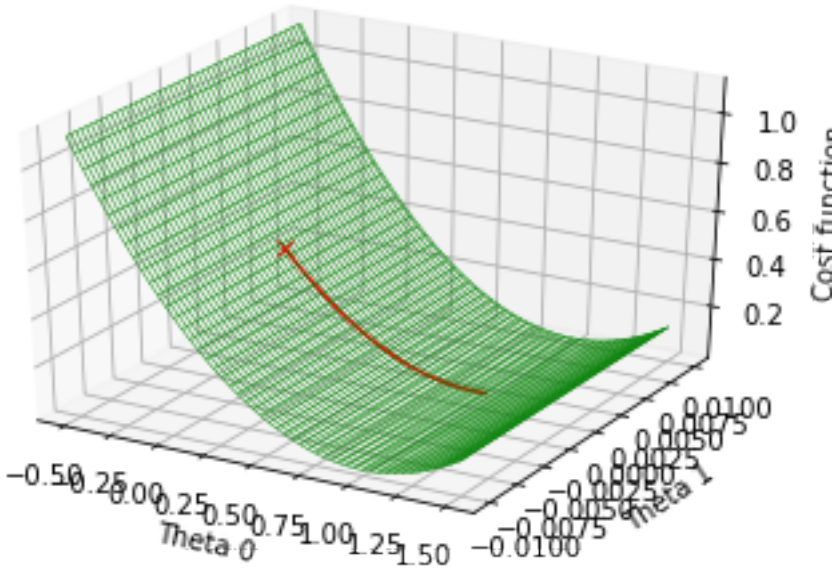
1 Linear Regression

(a) The learning rate(η) = 0.001, and parameter $\theta = \begin{bmatrix} 0.99659363 \\ 0.00134016 \end{bmatrix}$. The stopping criteria used was number of iterations = 100

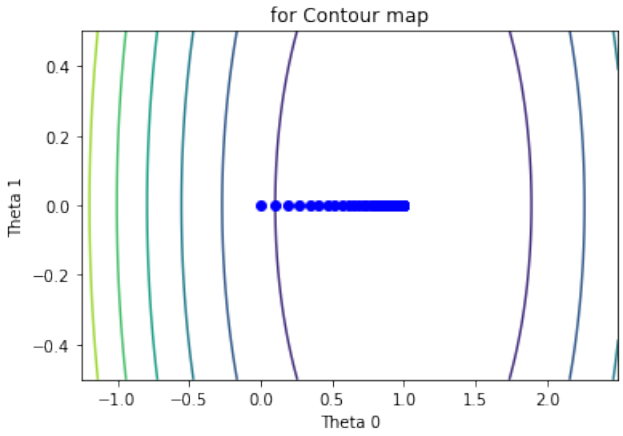
(b) The corresponding plot of training data and hypothesis function is:



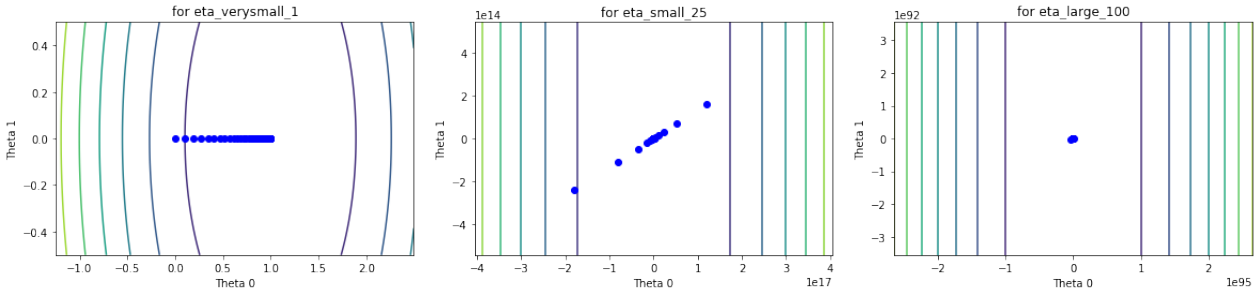
(c) The animation is submitted in form of a video, the final picture is as below:



(d) The animation is submitted in form of a video, the final picture is as below:



(e) The animations are submitted in form of a video, the final picture is as below for different learning rates:



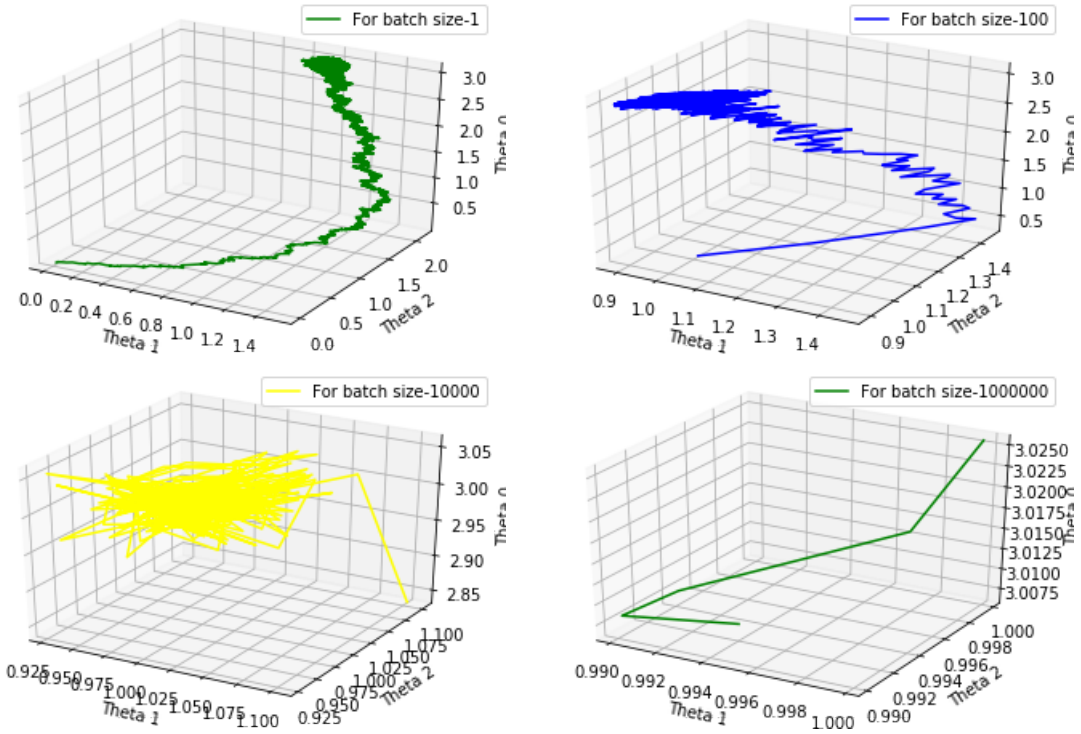
2 Sampling and Stochastic Gradient Descent

(a) The data was sampled as per the given specifications.

(b) Coefficients learnt are $\theta_1 = \begin{bmatrix} 3.01129451 \\ 1.0168542 \\ 1.99710636 \end{bmatrix}$, $\theta_{100} = \begin{bmatrix} 3.01129451 \\ 1.0168542 \\ 1.99710636 \end{bmatrix}$, $\theta_{10000} = \begin{bmatrix} 3.01129451 \\ 1.0168542 \\ 1.99710636 \end{bmatrix}$, and $\theta_{1000000} = \begin{bmatrix} 3.01129451 \\ 1.0168542 \\ 1.99710636 \end{bmatrix}$

(c) Yes, the parameter θ converge to the same value for varying values of batch size. They are almost same as the parameters of the original hypothesis function from which the data was generated. The number of iterations taken to converge is same in all the cases. Though the batch size 1 takes relatively more time to run which is due to cache miss and cache coherence fault in the memory since the data size is very large. Since the parameter learnt in each case is same, the error reported in each case is also the same which is 0.9974710884444786. Note that this error is average value.

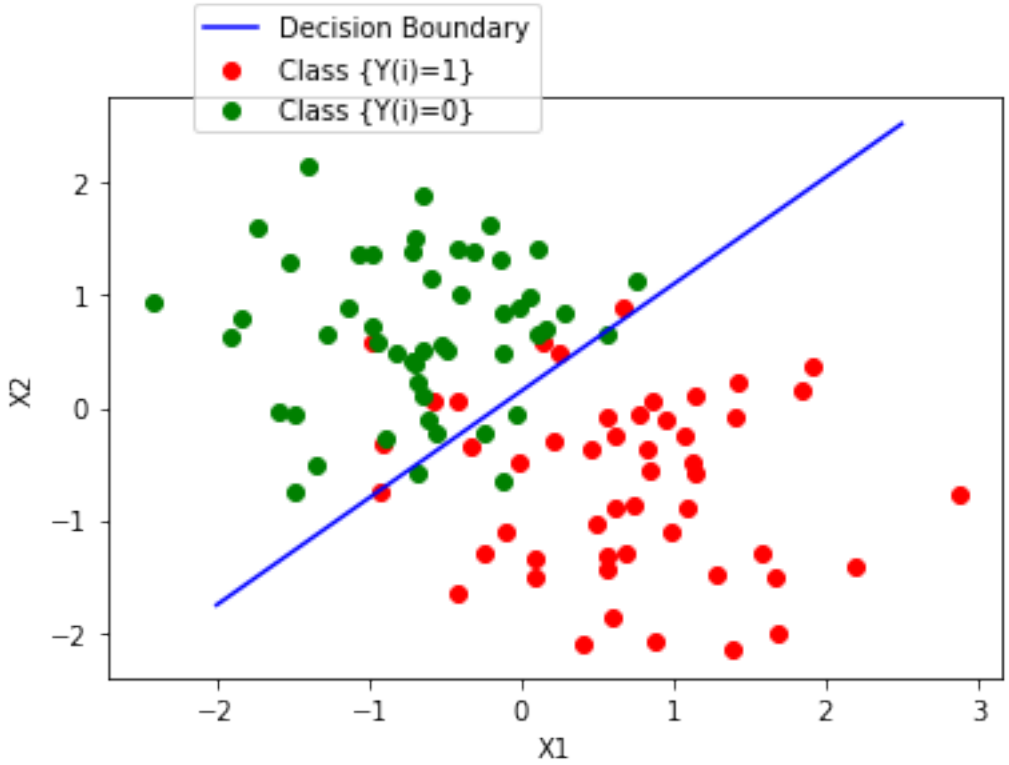
(d) The shape of movement is almost same for the case of 1 and 100. From 1 to 10000, the movement near the optimum values of theta is more random as the model converges and then roams around the same value of theta for a longer period of time, hence the distortion in the plots. For the batch size 1 million, the movement is very smooth, rather it moves in straight lines, because the batch size is same as the size of data set. So this case is that of batch gradient descent. The plots for different batch sizes are as follows:



3 Logistic Regression

(a) The coefficients $\theta = \begin{bmatrix} 0.40125316 \\ 2.5885477 \\ -2.72558849 \end{bmatrix}$

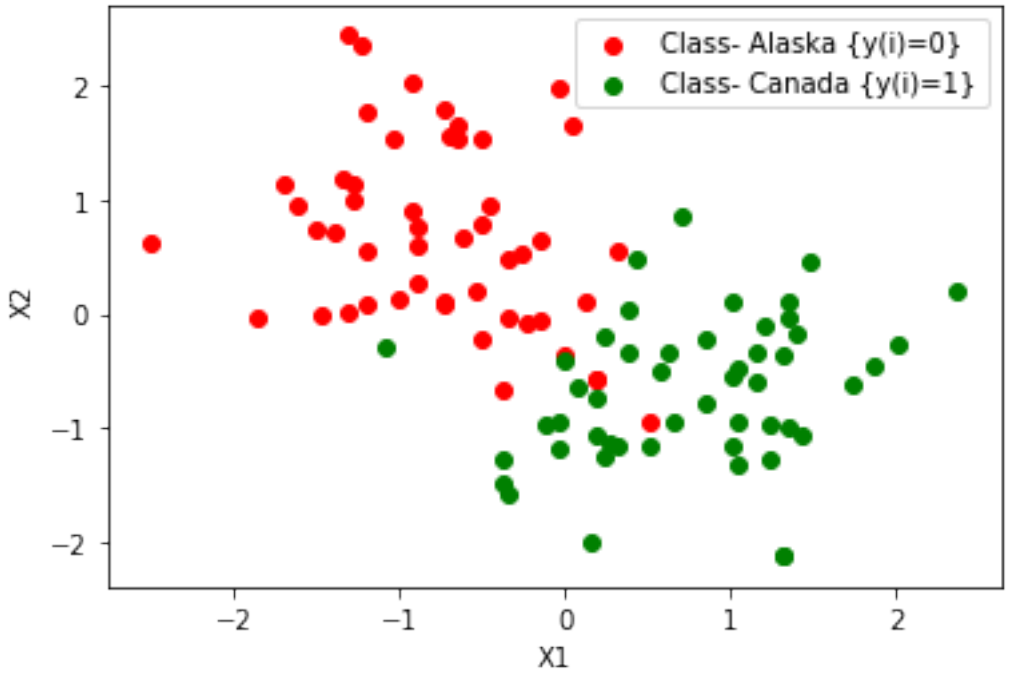
(b) The corresponding plot of training data and decision boundary is:



4 Gaussian Discriminant Analysis

(a) Value of $\mu_0 = \begin{bmatrix} -0.75529433 \\ 0.68509431 \end{bmatrix}$, $\mu_1 = \begin{bmatrix} 0.75529433 \\ -0.68509431 \end{bmatrix}$, and $\Sigma_0 = \Sigma_1 = \Sigma = \begin{bmatrix} 0.42953048 & -0.02247228 \\ -0.02247228 & 0.53064579 \end{bmatrix}$

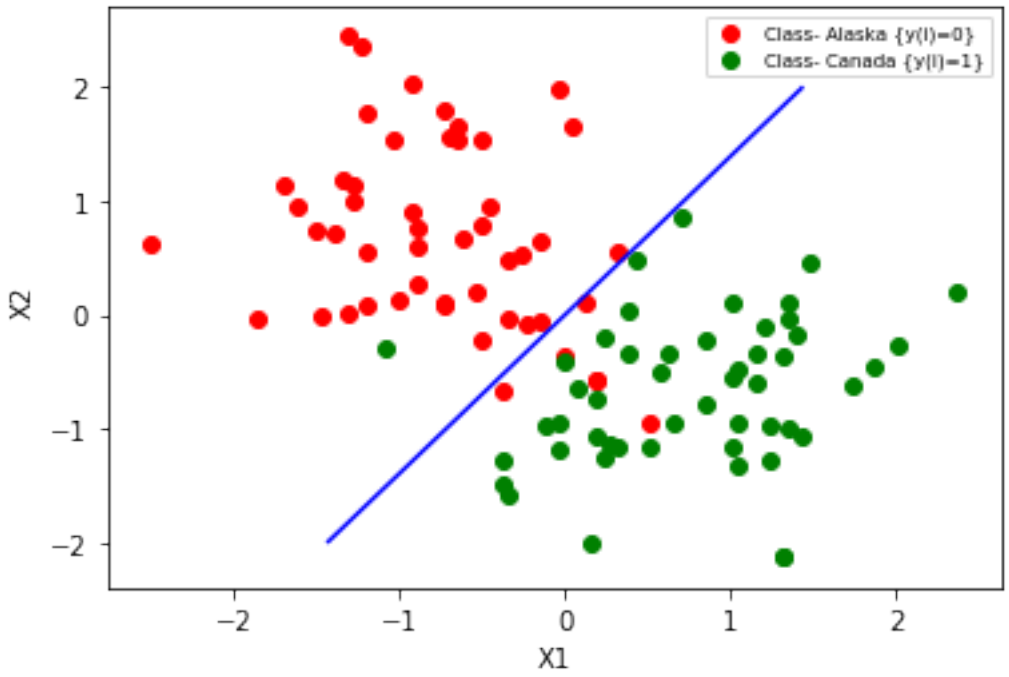
(b) The corresponding plot of training data is:



(c) The equation is as follows:

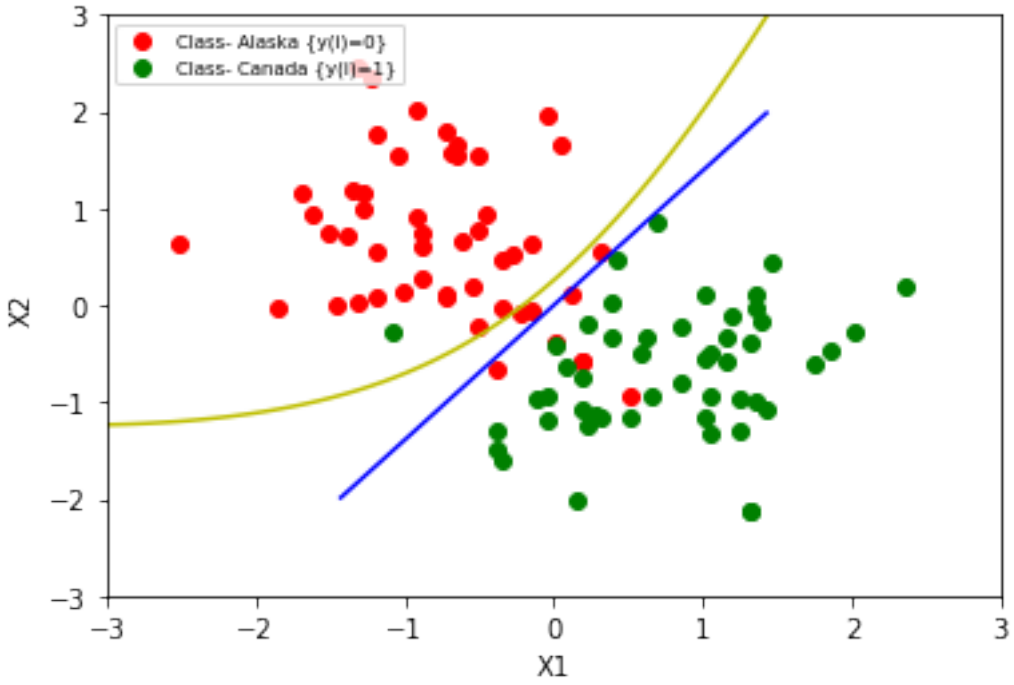
$X^T\Sigma^{-1}(\mu_1 - \mu_0) + (\mu_1^T - \mu_0^T)\Sigma^{-1}X + \mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1 + 2\log\frac{\phi}{(1-\phi)} = 0$, where $X = \begin{bmatrix} X1 \\ X2 \end{bmatrix}$

The corresponding plot of training data and linear decision boundary is:



(d) $\mu_0 = \begin{bmatrix} -0.75529433 \\ 0.68509431 \end{bmatrix}$, $\mu_1 = \begin{bmatrix} 0.75529433 \\ -0.68509431 \end{bmatrix}$, $\Sigma_0 = \begin{bmatrix} 0.38158978 & -0.15486516 \\ -0.15486516 & 0.64773717 \end{bmatrix}$, and $\Sigma_1 = \begin{bmatrix} 0.47747117 & 0.1099206 \\ 0.1099206 & 0.41355441 \end{bmatrix}$

(e) The equation is as follows:
$X^T(\Sigma_0^{-1}-\Sigma_1^{-1})X+\mu_0^T\Sigma_0^{-1}\mu_0-\mu_1^T\Sigma_1^{-1}\mu_1+(\mu_1^T\Sigma_1^{-1}-\mu_0^T\Sigma_0^{-1})X+X^T(\Sigma_1^{-1}\mu_1-\Sigma_0^{-1}\mu_0)+2\log\frac{\phi}{(1-\phi)}+\log\frac{|\Sigma_0|}{|\Sigma_1|}=0,$
where $X=\begin{bmatrix}X1\\X2\end{bmatrix}$
The corresponding plot of training data, linear decision boundary and quadratic decision boundary is:



(f) In this case the linear boundary seems to be separating the data in a better manner than the quadratic boundary. Initially, We had normalized the entire data under one mean and variance, therefore it is more likely that the entire data has same co-variance and hence in this case the linear boundary seems to be more fitting. The quadratic boundary assumes that the data for both the classes have different co-variance which is likely not the case here and hence it is not that a good separator as its linear counterpart.