Introduction to machine learning
Assignment 2
Course Instructor : Arun Rajkumar.
Release Date : April 5,2022
Submission Date: On or before 5 PM on April 24, 2022

SCORING: There are 2 questions in this assignment. The contribution of points scored in this assignment towards your final grades will be 25 points

The points will be decided based on the clarity and rigour of the report provided and the correctness of the code submitted.

DATASETS The data-sets are in the corresponding google drive folder shared in Moodle.

WHAT SHOULD YOU SUBMIT? You should submit a zip file titled 'Solutions_rollnumber.zip' where rollnumber is your institute roll number. Your assignment will NOT be graded if it does not contain all of the following:

- A text file titled 'Details.txt' with your name and roll number.
- A PDF file which includes explanations regarding each of the solution as required in the question. Title this file as 'Report.pdf'
- Clearly named source code for all the programs that you write for the assignment .

CODE LIBRARY: You are expected to code all algorithms from scratch. You cannot use standard inbuilt libraries for **algorithms** taught in class. You are free to use inbuilt libraries for plots. You can code using either Python or Matlab or C.

GUIDELINES: Keep the below points in mind before submission.

- Plagiarism of any kind is unacceptable. These include copying text or code from any online sources. These will lead to disciplinary actions according to institute guidelines.
- Any graph that you plot is unacceptable for grading unless it labels the x-axis and y-axis clearly.
- Don't be vague in your explanations. The clearer your answer is, the more chance it will be scored higher.

LATE SUBMISSION POLICY You are expected to submit your assignment on or before the deadline to avoid any penalty. Late submission incurs a penalty in points equal to the number of days your submission is late by. Any late submission post three days of the deadline would not be graded and will fetch 0 points.

QUESTIONS

(1) You are given a data-set with 1000 data points generated from a mixture of some distribution in the file A2Q1.csv.

i. (i) Determine which probabilisitic *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.

ii. (ii) Assume that the same data was infact 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.

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

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

(2) You are given a data-set in the file A2Q2Data_train.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).

i. Obtain the least squares solution $\mathbf{w}_{ML}$ to the regression problem using the analytical solution.

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

iii. 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 your observations?

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 $\mathbf{w}_R$. Compare the test error (for the test data in the file A2Q2Data_test.csv) of $\mathbf{w}_R$ with $\mathbf{w}_{ML}$. Which is better and why?