CS5691: Pattern recognition and machine learning
Programming Assignment 1
Course Instructor : Arun Rajkumar.
Release Date : Sep-05, 2019
Submission Date: On or before 5 PM on Sep-23,2019

SCORING: There are 3 questions in this assignment. Each question carries 5 points. The total points obtained will be multiplied by $\frac{2}{3}$ as contribution towards your final grades. The points will be decided based on the report provided, code submitted and a final oral examination that covers all the assignments together.

DATASETS Each question has an associated data-set indexed by the question number (Eg: Dataset1 corresponds to Question 1).

WHAT SHOULD YOU SUBMIT? You should submit a zip file titled 'Solutions_ rollnumber1 _rollnumber2.zip' where rollnumber1 and rollnumber2 are roll numbers of the members of the group. Your assignment will NOT be graded if it does not contain all of the following:

- A text file titled 'Participants.txt' with names and roll numbers of members.
- A PDF file which includes explanations regarding each of the solution as required in the question. Title this file as 'Report.pdf'
- Source code for all the programs that you write for the assignment clearly named.

CODE LIBRARY: You are expected to code all algorithms from scratch. You cannot use standard inbuilt libraries for **computations**. The only allowed library are those that compute the Eigenvectors and Eigenvalues of matrices. If your code calls any other library function for computation, it will fetch 0 points. 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 equal to the number of days your submission is late by. For instance if you score 12 points out of 15, then your non-penalized score out of 10 would be 8. However, If you submit it after 5 PM on Sep-23,2019 and before 5 PM on Sep-24, 2019, your score will be $8 - 1 = 7$ points. If you submit it after 5 PM on Sep-24 and before 5 PM on September 25, your score will be $8 - 2 = 6$ points and so on. If you obtain negative points after penalty (or don't turn in your assignment), it will be considered as 0 points.

QUESTIONS

(1) You are given a data-set with 2000 data points each in $\mathbb{R}^2$.

i. Plot this data-set. What distribution might have generated this data-set? Why did you conclude so?

ii. Write a piece of code to obtain the maximum likelihood estimate of the parameters of the distribution that you think generated this data-set.

iii. What is the log-likelihood value of observing this data-set for the parameters that you estimated in (ii)?

iv. Assume that the above data-set was generated from a Gaussian distribution with unknown mean $\mu \in \mathbb{R}^2$ and a known co-variance of $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. What is the maximum likelihood estimator for μ?.

v. Under assumption (iv), plot the log likelihood of observing this data-set as a function of μ where each component of μ belongs to $\{-10, \ldots, 10\}$. Compare this graph with the log-likelihood value you obtained in (iii). What can you conclude?

(2) You are given a data-set with 500 data points each in $\mathbb{R}$. This data is generated from a Gaussian mixture model with an unknown number of mixtures k and unknown mixing co-efficients and parameters (mean and Covariance) of individual Gaussians.

i. Write a piece of code for estimating the parameters of a uni-variate Gaussian mixture model given the number of mixtures k.

ii. For every choice $k \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, run your code to estimate the parameters of the model. Tabulate all the values you obtain.

iii. Plot the log-likelihood of the data-set for the parameters estimated as a function of the number of mixtures k. What can you conclude about the process that generated this data-set using this plot?

(3) You are given a data-set with 1000 data points each in $\mathbb{R}^2$.

i. Write a piece of code to run the PCA algorithm on this data-set. How much of the variance in the data-set is explained by each of the principal components?

ii. Write a piece of code to implement the Kernel PCA algorithm on this dataset. Use the following kernels :

A. $\kappa(x, y) = (1 + x^T y)^d$ for $d = \{2, 3\}$

B. $\kappa(x, y) = \exp \frac{-(x-y)^T(x-y)}{2\sigma^2}$ for $\sigma = \{0.1, 0.2, \ldots, 1\}$

Plot the projection of each point in the dataset onto the top-2 components for each kernel. Use one plot for each kernel and in the case of (B), use a different plot for each value of σ.

iii. Which Kernel do you think is best suited for this dataset and why?