DS413 - Introduction to Statistical Learning

Assignment-2
Assigned Date: 11/09/2024
11:59pm, Due Date:30/09/2024

Instructions

- Work on the assignments on your own. You are free to discuss among your selves,but don't copy. If we find the assignments of a group of (two or more) students very similar, the group will get zero points towards this assignment. Plagiarism will be cheked with tools. Please use Python for writing code. You can submit the code as a Jupyter notebook and For the theory questions, please submit your work to TAs. Use slack to discuss, if you have any doubts.

1 Theory

1.1 Maximum Likelihood Estimation - Theory only

1. Professor decides to assign final grades in a subject by ignoring all the work the students have done and instead using the following probabilistic method: each student independently will be assigned an A with probability θ , a B with probability 3θ, a C with probability $\frac{1}{2}$, and an F with probability $\frac{1}{2} - 4\theta$. When the quarter is over, you discover that only 2 students got an A, 10 got a B, 60 got a C, and 40 got an F.Find the maximum likelihood estimate for the parameter θ that Professor used. Give an exact answer as a simplified fraction. [5 Marks]

2. Suppose that the lifetime of Badger brand light bulbs is modeled by an exponential distribution with (unknown) parameter γ. We test 5 bulbs and find they have lifetimes of 2, 3, 1, 3, and 4 years, respectively. What is the MLE for γ? [5 Marks]

2 Programming Assignment

2.1 Logistic Regression - implement from scratch

1. Go back to the second question in Assignment-1 (second question - 5th and 6th sub parts) and Learn the logistic regression model from scratch for the dataset with outliers and comment on the fit? Is it any better? [10 marks]

2. Create your own dataset with four 2-dimensional Gaussians such that the data distributed will have four classes at four corners. example: class means = (2,2),(-2,2),(2,-2,)(-2,-2) respectively and use small variance around 0.1. Fit a 4-class classifier using logistic regression model using Gradient Ascent or Descent method. Observe the model fit? print classification accuracy? [10 marks]

2.2 Programming [30 Marks + 5 Bonus Marks (will be directly added to endsem marks]

Dataset link

Implement GMM without any builtin functions: Question-1 [30 Marks]
The parameters of Gaussian Mixture Model (GMM) can be estimated via the EM algorithm.

1. Download the Old Faithful Geyser Dataset.link The data file contains 272 observations of (eruption time, waiting time). Treat each entry as a 2 dimensional feature vector. Parse and plot all data points on 2-D plane.

2. Implement a bimodal GMM model to fit all data points using EM algorithm. Explain the reasoning behind your termination criteria. For this problem, we assume the covariance matrix is spherical (i.e., it has the form of σ^2I for scalar σ) and you can randomly initialize Gaussian parameters. For evaluation purposes, please submit the following figures:

    - Plot the trajectories of two mean vectors in 2 dimensions (i.e., coordinates vs. iteration).
    - Run your program for 50 times with different initial parameter guesses. Show the distribution of the total number of iterations needed for algorithm to converge.

3. Repeat the task in (c) but with the initial guesses of the parameters generated from the following process:

    - Run a k-means algorithm over all the data points with K = 2 and label each point with one of the two clusters.
    - Estimate the first guess of the mean and covariance matrices using maximum likelihood over the labeled data points. Compare the algorithm performances of (c) and (d)

Real world Problem-Speaker Identification System with builtin function: Question -2 [Bonus 5 Marks in endsem] Dataset link

1. Check the audio files from the dataset for different speakers. Split the data into training and testing.

2. Extract features using MFCC function (Mel Frequency Cepstral Coefficients) (You can use python package to extract features from an Audio-file
(Ref- https://stackoverflow.com/questions/54160128/feature-extraction-using-mfcc)).

3. Implement GMM model for each speaker (Use sklearn package)

4. Take test speech sample (audio file) and identify the speaker.