DS403 - Introduction to Statistical Learning

Assigned Date: 20/10/2023

11:59pm, Due Date:5/11/2023

data: https://drive.google.com/drive/folders/1c4NYGUZCLrIWY-eJydKvEyJp4xOGLka-?usp=sharing

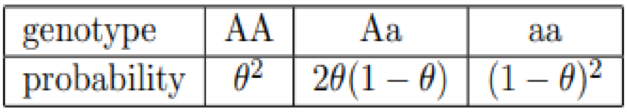
1 Maximum Likelihood Estimation

1.1 Theory [15 Marks]

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]

3. Suppose that a particular gene occurs as one of two alleles (A and a), where allele A has frequency in the population. That is, a random copy of the gene is A with probability θ and a with probability $1-\theta$. Since a diploid genotype consists of two genes, the probability of each genotype is given by:
   Suppose we test a random sample of people and find that k1 are AA, k2 are Aa, and k3 are aa. Find the MLE of θ [5 Marks]



genotype	AA	Aa	aa
probability	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

1.2 Programming [10 Marks]

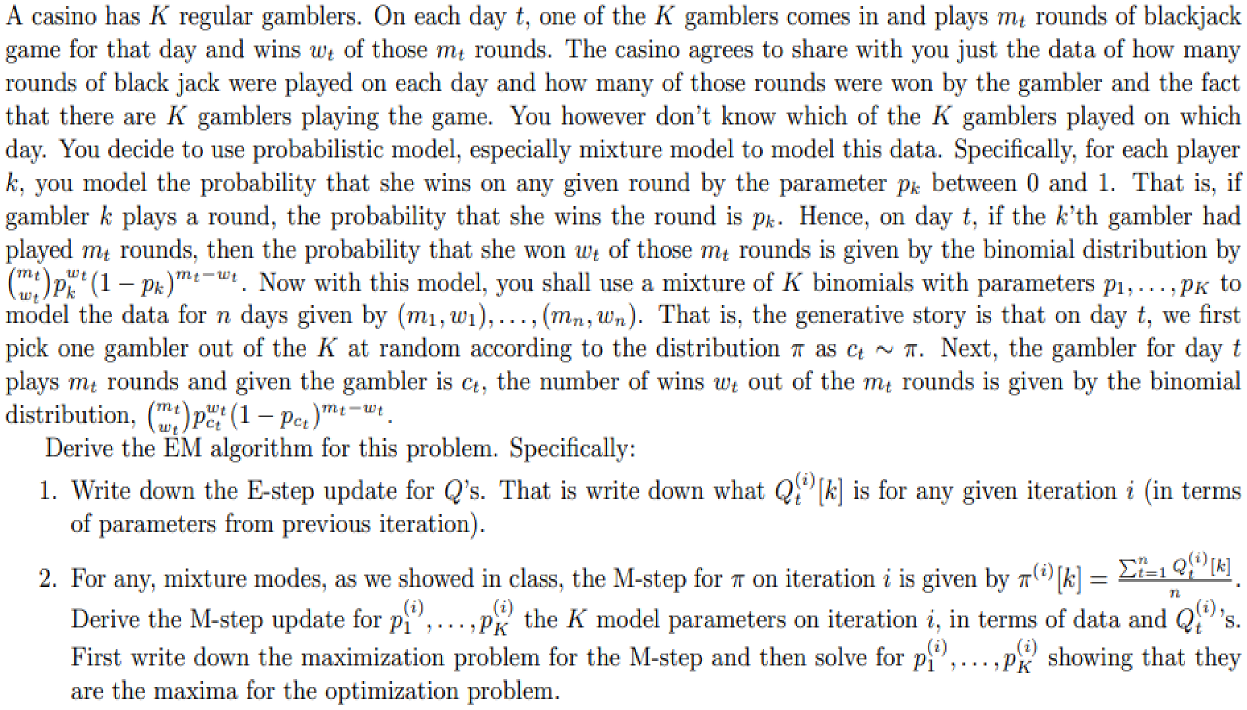
1. Write a program to a Bayes classifier and use it to predict class labels of test data. Laplacian smoothing should be used. The learned classifier should be tested on test instances and the accuracy of prediction for the test instances should be printed as output. A single program should train the classifier on the training set as well as test it on the test set.
   Data Set Description: The task is to predict whether a citizen is happy to live in a city based on certain parameters of the city as rated by the citizens in a scale of 1-5 during a survey. Attribute Information: download link
   D = decision/class attribute (D) with values 0 (unhappy) and 1 (happy) (Column 1 of file) X1 = the availability of information about the city services (Column 2 of file) X2

= the cost of housing X3 = the overall quality of public schools X4 = your trust in the local police X5 = the maintenance of streets and sidewalks X6 = the availability of social community events Attributes X1 to X6 have values 1 to 5.

2 Gaussian Mixture Modeling- EM Algotithm

2.1 Theory [25 Marks]

Question-1 [10 Marks]



A casino has K regular gamblers. On each day t, one of the K gamblers comes in and plays m_t rounds of blackjack game for that day and wins w_t of those m_t rounds. The casino agrees to share with you just the data of how many rounds of black jack were played on each day and how many of those rounds were won by the gambler and the fact that there are K gamblers playing the game. You however don't know which of the K gamblers played on which day. You decide to use probabilistic model, especially mixture model to model this data. Specifically, for each player k, you model the probability that she wins on any given round by the parameter p_k between 0 and 1. That is, if gambler k plays a round, the probability that she wins the round is p_k. Hence, on day t, if the k'th gambler had played m_t rounds, then the probability that she won w_t of those m_t rounds is given by the binomial distribution by $\binom{m_t}{w_t} p_k^{w_t}(1-p_k)^{m_t-w_t}$. Now with this model, you shall use a mixture of K binomials with parameters $p_1, \ldots, p_K$ to model the data for n days given by $(m_1, w_1), \ldots, (m_n, w_n)$. That is, the generative story is that on day t, we first pick one gambler out of the K at random according to the distribution π as $c_t \sim \pi$. Next, the gambler for day t plays m_t rounds and given the gambler is c_t, the number of wins w_t out of the m_t rounds is given by the binomial distribution, $\binom{m_t}{w_t} p_{c_t}^{w_t}(1-p_{c_t})^{m_t-w_t}$.

Derive the EM algorithm for this problem. Specifically:

1. Write down the E-step update for Q's. That is write down what $Q_t^{(i)}[k]$ is for any given iteration i (in terms of parameters from previous iteration).

2. For any, mixture modes, as we showed in class, the M-step for π on iteration i is given by $\pi^{(i)}[k] = \frac{\sum_{t=1}^{n} Q_t^{(i)}[k]}{n}$. Derive the M-step update for $p_1^{(i)}, \ldots, p_K^{(i)}$ the K model parameters on iteration i, in terms of data and $Q_t^{(i)}$'s. First write down the maximization problem for the M-step and then solve for $p_1^{(i)}, \ldots, p_K^{(i)}$ showing that they are the maxima for the optimization problem.

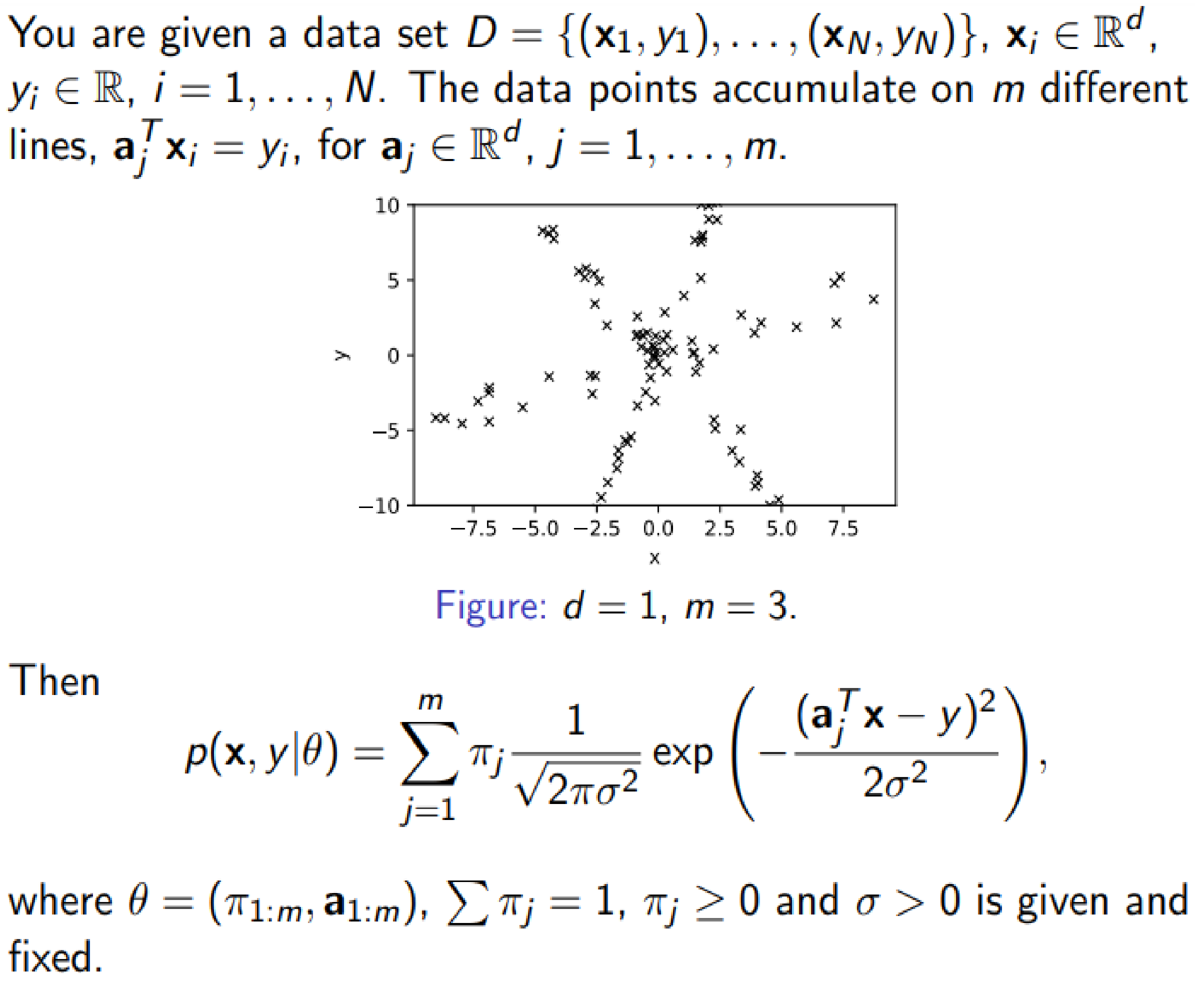
Question-2 [15 Marks]

Please the terms for the given probability density function in the Figure below.

- a) Find the responsibilities in the E-step of EM, $\gamma_{nk}^{t} = p(z_n = k|x_n, y_n, \theta^{(t-1)})$.. t = iteration number for EM algorithm.

- b) Write down the optimization objective for the M-step in EM for $\pi_j^{(t)}$ and $a_j^{(t)}$ in the terms of the responsibilities $\gamma_{nk}^{(t)}$.

- c) Derive the explicit update rules for $\pi_j^{(t)}$ and $a_j^{(t)}$ used in the M-step of EM. Hint: You can assume that the matrix $\Sigma_{n=1}^{N} \gamma_{nk}^{(t)} x_n x_n^T$ is invertible for all $k = 1, ..., m$



You are given a data set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, $i = 1, \ldots, N$. The data points accumulate on m different lines, $\mathbf{a}_j^T \mathbf{x}_i = y_i$, for $\mathbf{a}_j \in \mathbb{R}^d$, $j = 1, \ldots, m$.

10
5
0
−5
−10
y
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
x

Figure: $d = 1$, $m = 3$.

Then

$$p(\mathbf{x}, y|\theta) = \sum_{j=1}^{m} \pi_j \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{a}_j^T \mathbf{x} - y)^2}{2\sigma^2}\right),$$

where $\theta = (\pi_{1:m}, \mathbf{a}_{1:m})$, $\sum \pi_j = 1$, $\pi_j \geq 0$ and $\sigma > 0$ is given and fixed.

Figure 1: Figure for Question2

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

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)

Speaker Identification System: Question -2 [Bonus 5 Marks in endsem]

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.

3 Support Vector Machine

3.1 Theory [10 Marks]

1. Comment on computational complexity of SVM?

3.2 Programming [60 Marks]

Question-1 [20 Marks] - **Social Network Ads data**

1. Suppose a company is going to launch a new campaign for their new brand of car and want to know which category of people are likely to buy their brand new car so they can have the ads that target those peoples. For this they contacted a social network advertising company which have the data from another similar successful campaign. Now, they want to make a model which helps achieve their goal.

2. Dataset: The dataset contains 400 entries which contains the userId, gender, age, estimatedsalary and the purchased history. The matrix of features taken into account are age and estimated salary which are going to predict if the user is going to buy new car or not(1=Yes, 0=No)

3. plot the 2-d data and visualize, comment whether a linear model can seperate the data or not. Split the data into

4. Fit SVM model using a python package.

5. Identify support vectors -data points and indicate on the plot. Compute the confusion matrix to find the test accuracy.

6. i) Apply RBF kernel ii) fit the SVM model and iii) Compute the confusion matrix.

7. i) Apply Gaussian kernel ii) fit the SVM model and iii) Compute the confusion matrix.

8. i) Apply Polynomial kernel ii) fit the SVM model and iii) Compute the confusion matrix.

9. Compare the SVM model with and without Kernels.

Question-2 - [20 Marks] **SVM Multi-dimension data and Binary Classification - News Data**

1. Implement SVM.Use a subset of the ”20 Newsgroups” dataset (news.mat). There are two topics of documents for classification taken from the talk.politics.misc and talk.religion.misc, for a total of 842 documents. X_{train} is a sparse matrix of training data, where each row is a document and each column is a feature (a word). X_{ij} = 1 corresponds to the document i containing the j-th word. Y_{train} is a vector of labels (1 or 1). X_{test} and Y_{test} follow the same pattern.

2. Load Matfile and convert to suitable format. Split the data into training and testing

3. Use Quadprog or Python Package to construct a solver for SVM

4. Calculate the mis-classification rate on the training and test data. Compute confusion matrix.

Question-3 [20 Marks] -**SVM Multi-dimension data and Multi Class Classification - MNIST data**

1. Load MNIST training and testing data.
2. Implement SVM using any python package to perform digit classification.
3. Compute the Confusion Matrix for test data.
4. Comment on your classifier. Is it using one Vs one classifiers or one vs rest classifiers?

4 Comparing the classifiers

4.1 Theory [20 Marks]

1. Compare the following classifiers
   - Logistic Regression
   - Linear Classifier Estimated using Least Squares.
   - Bayes Classifier
   - Support Vector Classifier

4.2 Programming - [Bonus 2 Marks will be added to your endsem marks]

1. Take any real time classification problem of your choice from your own domain.Get the data. Describe the problem statement.
2. Solve that particular using the following classifiers.
   - Logistic Regression
   - Linear Classifier Estimated using Least Squares.
   - Bayes Classifier
   - Support Vector Classifier
3. Implement all of these classifiers and compare the performance in terms of test accuracy.
4. Explain your results.