CSE/ISyE 6740-A Midterm Exam

Computational Data Analytics
Fall 2024

Oct 10th 2024, 12:30pm - 1:45pm ET (75min)

Personal Information

- Name:
- GTID:
- Email:
- Course section (please select CSE, ISyE, or unsure):

Please remember to fill in your personal information. No grade will be given if we cannot identify your identity. To be safe, please write your name on all pages as well.

Instructions

- The full score for the midterm exam is **100 points**. If you are not sure if you miss any questions, please double check if the sum is 100 points.
- There are **5 questions** in **9 pages** (excluding the instruction page), each of which may ask you to answer some smaller questions. Please remember to answer all of them before moving onto the next one.
- The exam is closed book. However, you are allowed to bring **one double-sided A4 handwritten note** as your cheat sheet for reference. Any other materials are not allowed. Calculator is neither needed nor allowed.
- You can write your solution on the question pages, or on any additional papers that you may want to use. But **please clarify which questions the answers correspond to and insert your answers into the question pages following the order of the questions**.
- Try your best to be as clear as possible. No credit will be given to unreadable writing.
- We will upload your solutions on behave of you to Gradescope immediately after the exam. After the grading is done, you can review it and submit regrade requests.
- Please return the question pages together with your answers, and do not share the questions outside of the class.
- Good luck and enjoy!

1 K-means Clustering [20 points]

Given a dataset with n samples $\{x^1, x^2, ..., x^n\}$ where each $x^i \in \mathbb{R}^d$ $\forall i$, let us assume K cluster centers, and the cluster centers are denoted by $\{c^1, c^2, \cdots, c^K\} \in \mathbb{R}^d$. We will perform k-means clustering on the samples to update the cluster centers.

(a) K-means clustering algorithm [10 points]

As we have shown in class, the clustering problem can be written as a minimization problem:

$$\min_{\pi, c} \frac{1}{n} \sum_{i=1}^{n} \left\| x^i - c^{\pi(i)} \right\|^2 \tag{1}$$

where the cluster assignment is denoted by $\pi(i) \in \{1, 2, \cdots, K\}$.

Please write down the pseudocode of the k-means clustering algorithm and explain why each step of the k-means algorithm can reduce the clustering objective.

(b) Computation cost and memory cost [5 points]

Please analyze the computation cost of each iterative step of the k-mean clustering algorithm. Your answer should be in terms of the number of samples n, the dimensionality of the sample d, and the number of clusters K.

(c) Memory-efficient k-means algorithm [5 points]

Suppose now that the size of the dataset n is too large for our memory to load. So, every time you can only load a small subset of the samples $n' \ll n$ into the memory. Which part of the k-means algorithm can go wrong when the dataset is too large? Could you modify the k-means algorithm so that it works for a large dataset? Please analyze and explain your solution.

2 EM Algorithm and Exponential Mixture Model [35 pts]

You are the owner of a cafe, and you want to better understand the waiting times between the arrivals of customers throughout the day. From your observations, you notice that there are peak hours (during breakfast and lunch times) and off-peak hours, with customer arrivals following different patterns. You hypothesize that the waiting times between customer arrivals can be modeled by a mixture of two exponential distributions.

During peak hours, customers arrive more frequently (with a shorter average waiting time). During off-peak hours, customers arrive less frequently (with a longer average waiting time). Based on what we learn from probability course, the probability density function of an exponential distribution with parameter λ is given by:

$$P(t|\lambda) = \lambda e^{-\lambda t}, \quad t \geq 0 \tag{2}$$

Therefore, you assume that the peak hours follow an exponential distribution with parameter λ_1, and off-peak hours follow an exponential distribution with parameter λ_2 You assume that the probability of being at peak hours is π_1, and the probability of being at off-peak hours is π_2.

In the past few months, you collected a dataset of n waiting time $D = \{x^1, x^2, \cdots, x^n\}, x^i \in \mathbb{R}\ \forall i$, but you did not record if each sample was collected during peak hours or off-peak hours. Please help yourself to find the parameters of the exponential mixture model using EM algorithm.

(a) Likelihood function [10 points]

Given parameters $\theta = (\pi_1, \pi_2, \lambda_1, \lambda_2)$, please compute the log-likelihood function $\log P(D|\theta)$ as a function of $\{x^i\}_{i=1,2,\cdots,n}$ and $\{\pi_k, \lambda_k\}_{k=1,2}$[1]. Please also explain why the log-likelihood function is hard to optimize.

[1] If you prefer, you can assume there are K distributions each with probability π_k and parameter λ_k for $k \in \{1, 2, \cdots, K\}$.

(b) Expectation step [15 points]

In the following two questions, please establish the EM algorithm for the exponential mixture model as we have shown in the Gaussian mixture model to solve the maximum likelihood problem.

(b-1) [5 points] Please compute the latent posterior probability $\tau_k^i = P(z^i = k|x^i, \theta)$ as a function of the model parameters $\{\pi_k, \lambda_k\}_{k=\{1,2\}}$ and the data points x^i, where z^i denotes which exponential distribution the data sample i was drawn. You can use τ_k^i to denote this posterior probability after this step.

(b-2) [10 points] Please derive the lower bound of the log-likelihood function $\log P(D|\theta)$ using Jensen's inequality. Please also prove under what conditions the lower bound touches the log-likelihood function $\log P(D|\theta)$, i.e., the equality of the Jensen's inequality is attained.

(c) Maximization step [10 points]

Please write down the maximization step of the exponential mixture model to maximize the lower bound. Note that you can use $\tau_k^i = P(z^i = k|x^i, \theta)$ to simplify the derivation as we did in class.

3 Decision Tree [10 pts]

Given a dataset $\{(x^i, y^i)\}_{i=1,2,\cdots,n}$ where $y^i \in \{0, 1\}$ and $x^i \in \mathbb{R}^d$ for all $i \in 1, 2, \cdots, n$, that is, continuous features and binary labels, please answer the following questions about the decision tree algorithm.

(a) Decision tree algorithm [5 points] Please describe how the decision tree algorithm works for continuous features. Please also explicitly explain how to compute the entropy term and the conditional entropy from the dataset.

(b) Computation cost of each split [5 points] Please analyze the computation cost of finding the best feature X_j and the best threshold t to split at $X_j > t$ for the dataset D. The computation cost should be expressed in terms of the number of samples n and the dimensionality of the features d.

4 Logistic Regression [15 points]

Given a dataset $D = \{(x^i, y^i)\}_{i=1,2,\cdots,n}$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{0, 1\}$ for all $i \in \{1, 2, \cdots, n\}$, please answer the following questions about logistic regression.

(a) Log-likelihood and concavity [10 points]

Please derive the log-likelihood function of using logistic regression to fit the dataset D, and show that the log-likelihood function is a concave function in terms of the logistic regression parameter.

(b) Regularization [5 points]

To prevent overfitting in logistic regression, regularization techniques such as L1 and L2 regularization are often used as well. Please describe the difference and effects of the regularization of L1 and L2 in logistic regression.

5 Support Vector Machine [20 pts]

Given a dataset $\{(x^i, y^i)\}_{i=1,2,\cdots,n}$ where $x^i \in \mathbb{R}^d$ and $y^i \in \{-1, 1\}$ for all $i \in \{1, 2, \cdots, n\}$, please answer the following questions about SVM.

Soft-margin SVM

The primal problem of SVM is derived from a margin-based minimization problem:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi^i \tag{3}$$

$$\text{s.t.} \quad y^i(w^\top x^i + b) \geq 1 - \xi^i \quad \forall i \tag{4}$$

$$\xi^i \geq 0 \tag{5}$$

The dual problem is given by:

$$\max_{\alpha} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y^i y^j ({x^i}^\top x^j) \tag{6}$$

$$\text{s.t.} \quad C \geq \alpha_i \geq 0 \quad \forall i \tag{7}$$

$$\sum_{i=1}^{n}\alpha_i y^i = 0 \tag{8}$$

(a) Dual problem derivation and concavity [10 points]

Please show how to derive the dual problem from the primal problem, and prove that the dual problem objective is a concave function in terms of the dual variable α.

(b) Support vectors and prediction [10 points]

Please explain the meaning of the variable ξ^i and α_i in the case of margin support vectors, non-margin support vectors, and non support vectors, respectively. You can use the following example to explain this. Please also explain how to use the dual solution α to make a prediction when a new feature x is given.