CS 274A Homework 2

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2023

Due: 11:59pm Sunday January 29th, submit via Gradescope

Instructions and Guidelines for Homeworks

- Please answer all of the questions and submit your solutions to Gradescope (either hand-written or typed are fine as long as the writing is legible).
- All problems are worth equal points unless otherwise stated. All homeworks will get equal weight in computation of the final grade for the class (with lowest-scoring homework being dropped).
- The homeworks are intended to help you better understand the concepts we discuss in class. It is important that you solve the problems yourself to help you learn and reinforce the material from class. If you don't do the homeworks you will likely have difficulty in the exams later in the quarter.
- In problems that ask you to derive or prove a result you should submit a complete mathematical proof (i.e., each line must follow logically from the preceding one, without "hand-waving"). Be as clear as possible in explaining your notation and in stating your reasoning as you go from line to line.
- If you can't solve a problem, you can discuss the high-level concepts *verbally* with another student (e.g., what concepts from the lectures or notes or text are relevant to a problem). However, you should not discuss any of the details of a solution with another student. In particular note that you are not allowed to view (or show to any other student) *any written material* directly related to the homeworks, including other students' solutions or drafts of solutions, solutions from previous versions of this class, etc. The work you hand in should be your own original work.
- If you need to you can look up standard results/definition/identities from textbooks, class notes, textbooks, other reference material (e.g., from the Web). If you base any part of your solution on material that we did not discuss in class, or is not in the class notes, or is not a standard known result, then you may want to rovide a reference in terms of where the result is from, e.g., "based on material in Section 2.2 in ....." or a URL.

Recommended Reading for Homework 2: Note Set 3

Problem 1: Maximum Likelihood for the Multinomial Model

Consider building a probabilistic model for how often words occur in English. Let W be a random variable, taking values $w \in \{w_1, \ldots, w_M\}$, where M is the number of words in the vocabulary. In practice M can be very large, e.g., $M = 100,000$ is not unusual (there are more words than this in English, but many rare words are not modeled). In other words, W is a discrete-valued random variable taking M possible values, with M probabilities that sum to 1. The parameters of the model are $\theta = \{\theta_1, \ldots, \theta_M\}$, where $\theta_m = P(W = w_m)$, and where $\sum_{m=1}^{M} \theta_m = 1$. If we generate samples in an IID manner from W then we refer to this as a *multinomial model*.

The *multinomial likelihood* (under the assumptions above) is similar to the binomial likelihood for tossing coins, but instead of two possible outcomes there are now M possible outcomes for each observation. Let the observed data be $D = \{r_1, \ldots, r_M\}$, where r_m is the number of times word w_m occurred, $m = 1, \ldots, M$ (these are known as the sufficient statistics for this model).

1. Show how the likelihood function can be defined for this problem
2. Derive the maximum likelihood estimates for each of the θ_m's for this model.

Problem 2: Visualization of Likelihoods

Consider a dataset $D = \{11, 5, 9, 52, 13, 25, 3, 6, 7, 12\}$ assumed to be generated in an IID manner from a probability density with parameters θ. (Even though these are integer values above, in what follows below we will investigate probability density functions for the data D, since its easier to specify a set of integers for this problem than a set of real-valued numbers).

Answer the following questions:

1. Using your favorite computing environment (Python, R, Matlab, etc) generate graphs of the log-likelihood function $l(\theta)$, as a function of θ, for the following models:
   (a) an exponential density with a mean (or scale) parameter $\theta = \frac{1}{\lambda}$, where $p(x) = \frac{1}{\theta} e^{-\frac{1}{\theta}x} = \lambda e^{-\lambda x}$ with $\lambda > 0, x \geq 0$.
   (b) a Gaussian density where $\theta = \mu$ (the unknown mean of the Gaussian). Assume that you know σ^2 and set it equal to the empirical variance defined as $\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$, where $m = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the empirical mean and where each x_i is one of the data points in D.

   Note: for this problem, when computing $l(\theta)$ include all terms in the density functions whether they include θ or not (i.e., don't drop normalization terms): this is important since you will be comparing likelihoods for two different densities.

Plot both functions $l(\theta)$ on the same graph, clearly indicating which function is which (e.g., use color, different linestyles, etc). Plot both functions over the range $\theta \in [3, 40]$. Use natural log (i.e., log to the base e) in your calculations. Put a grid in the background of the plot to make it easier to read (e.g., `grid on` in Matlab).

Also put on your graph the two log-likelihood values for a uniform density $U(a, b)$, with $a = 0$ and (i) $b = 60$, and (ii) $b = 100$. These are each single numbers so just plot each of them as a flat horizontal line corresponding to the log-likelihood value on the y-axis. Be sure to clearly indicate which line corresponds to which value of b.

2. Write a few sentences interpreting what you see in the graphs.

3. Now regenerate the same types of graph (with plots for exponential and Gaussian) for this new dataset $D_2 = \{11, 33, 19, 44, 13, 25, 31, 26, 37, 22\}$, computing σ^2 now using D_2, and again overlaying the two uniform logL values. Comment briefly on this new graph.

Problem 3: Maximum Likelihood: Geometric Distribution

Consider a data set $D = \{x_1, \ldots, x_n\}$, where $x_i \in \{1, 2, 3, \ldots\}$. Assume that the integer-valued data x_i were generated in an IID manner from a geometric model, where the geometric distribution with parameter θ is defined as

$$P(x_i = k) \;=\; (1-\theta)^{k-1}\theta\,, \quad k = 1, 2, 3, \ldots, \quad 0 < \theta < 1$$

1. Define the likelihood function for this problem

2. Derive an equation for the maximum likelihood estimate for θ

3. Now consider a different problem where our data consists of IID samples in the form of pairs (x_i, c_i) where $c_i \in \{1, 2\}$ are possible values of a discrete random variable. The density for $P(x)$ is a mixture model: $P(x) = P(x|c = 1)P(c = 1) + P(x|c = 2)P(c = 2)$, where $P(x|c = 1)$ and $P(x|c = 2)$ are both geometric distributions with parameters λ_1 and λ_2 respectively. Derive maximum likelihood estimates for λ_1, λ_2, and $\alpha = P(c = 1)$.

4. Now consider a different version of the previous problem where the constraint has been added that $\lambda_2 = 2\lambda_1$. Derive the maximum likelihood estimates for λ_1 for this constrained problem (for this part of the problem you can assume that $\alpha = P(c = 1)$ is known).

Problem 4: Method of Moments and the Uniform Model

The method of moments is an alternative parameter estimation method to maximum likelihood. Theoretically its' properties are not in general as good as maximum likelihood, but it can nonetheless be useful for some problems (e.g., where the likelihood function is not easy to optimize but the method of moments is easier to work with).

The method works as follows: Given a probability model (e.g., a Gaussian, a uniform, etc) with K parameters we write down K equations that express the first K moments as functions of K parameters. The moments are defined as $E[X^k], k = 1, \ldots, K$. Given a data set with N data points $x_1, \ldots, x_N$, we then plug in the empirical estimates of these moments (from the data, e.g., the average value of x_i, of x_i^2, etc) into these equations and get K equations with K unknown parameters. We can think of this method as "moment matching,", i.e., it is trying to find parameters such as the moments of the model (with its estimated parameters) match the empirical moments in the observed data.

Let X be uniformly distributed with lower limit a and upper limit b, where $b > a$, i.e.,

$$p(x) = \frac{1}{b-a}$$

for $a \leq x \leq b$ and $p(x) = 0$ otherwise. Assume we have a data set D consisting of n scalar measurements $x_i, 1 \leq i \leq n$, where the x_i are conditionally independent given a and b.

1. Derive estimators for a and b using the method of moments. Since there are 2 unknown parameters you will need two equations, involving the first and second moment.

2. Now derive the maximum likelihood estimators for a and b (think carefully about how to do this: it is somewhat different conceptually to the examples we did in class).

3. Write 2 or 3 sentences comparing the properties of the maximum likelihood estimates with the method of moment estimates. You can use the following simple data set $D = \{12, 4, 4, 10, 7, 5, 9, 10\}$ to provide some intuition for your answer.

Problem 5: Maximum Likelihood for a Simple Model of Graphs

Consider an undirected graph G with $N > 1$ nodes (or vertices) and with r undirected edges. Note: in this problem the graph is **not** a graphical model, just a standard graph. The edges in G are denoted by $e_{i,j} = e_{j,i}, 1 \leq i, j \leq N$. If $e_{i,j} = 1$ there is an edge between nodes i and j, and if $e_{i,j} = 0$ there is no edge between i and j. You can assume there are no self-edges, i.e., that $e_{i,i} = 0$, for $1 \leq i \leq N$. We can think of the graph as being represented by an $N \times N$ binary adjacency matrix $D = \{e_{i,j}\}, 1 \leq i \leq N, 1 \leq j \leq N$ where D is the data representing relations between N nodes.

We would like to fit a probabilistic model to this data D where we have a single parameter $\theta = p(e_{i,j} = 1)$. In terms of a generative model, edges $e_{i,j}$ in the graph are generated independently with probability θ.

1. Precisely define the likelihood $p(D|\theta)$ for this problem in terms of the information provided above. Try to reduce the likelihood to as simple an expression as possible.

2. Derive the maximum likelihood estimate for θ using the information provided above.

3. Briefly mention one significant limitation of this graph model if we wanted to use it to model real-world graphs.

4. Now say we have K different graphs, where the data for each graph is assumed to be generated independently with a single parameter θ in the same manner as above. Each graph G_k has N_k nodes and r_k edges, $k = 1, \ldots, K$, and each G_k has its own adjacency matrix D_k. Note that N_k is allowed to be different for different graphs. Our data is now the set $D = \{D_1, \ldots, D_K\}$ and we have a single parameter θ.

   Consider the following estimator for θ, $\hat{\theta} = \frac{1}{K}\sum_{k=1}^{K} \hat{\theta}^{(k)}$, where $\hat{\theta}^{(k)}$ is a maximum likelihood estimate defined for each graph separately (i.e., just using data D_k for each graph), using your answer from part 1. Is this the maximum likelihood estimator of θ or not? Provide a justification of your answer.

Problem 6: Markov Chains with Missing Data

Consider the following problem. We have 3 binary random variables A, B, C with the following graphical model: $A \rightarrow B \rightarrow C$, i.e., the 3 variables form a Markov chain of length 3.

Each variable takes values 0 or 1. Assume that the conditional probability table (CPT) for $P(B|A)$ and $P(C|B)$ are the same, i.e., the entries are the same. (This is known as a homogeneous Markov chain). These CPTs are often referred to as "transition matrices" when discussing Markov chains.

In addition we will assume that each CPT (or transition matrix) is such that $P(B = 0|A = 0) = \alpha$, $P(B = 1|A = 1) = \alpha$, i.e., the "self-transition" probabilities conditioned on $A = 0$ or $A = 1$ are the same and equal to α. And because the chain is homogeneous, we also have that $P(C = 0|B = 0) = \alpha$, $P(C = 1|B = 1) = \alpha$. To complete our notation let $P(A = 0) = \phi$. So, for this simple special case, our Markov chain with 3 variables just has 2 parameters ϕ and α.

Assume that we have observed data D consisting of n independently drawn random samples from the chain, i.e., $D = \{a_i, b_i, c_i\}_{i=1}^{n}$. Now, further assume that the b_i's have all been removed from the dataset, resulting in a dataset $D' = \{a_i, c_i\}_{i=1}^{n}$ that only has the a_i and c_i observations and the corresponding b_i values are missing. (This problem is motivated by a real-world problem involving missing data, in the context of remote-sensing and satellite data).

1. Show step by step how the log-likelihood $L(\phi, \alpha) = P(D'|\phi, \alpha)$ for this problem can be defined as a function of ϕ, α, and n_S and n. Here $n_S \leq n$ is defined as $\sum_{i=1}^{n} I(a_i, c_i)$ where $I(a_i, c_i) = 1$ if $a_i = c_i$ and $I(a_i, c_i) = 0$ if $a_i \neq c_i$: in other words, n_S counts the number of samples (out of n) where the values of a_i and c_i are the same (i.e., the chain starts and ends with the same value).

2. Now assume that $\phi = 0.5$ (and everything else is the same as before). Derive a closed form expression for finding a maximum likelihood estimate $\hat{\alpha}_{ML}$. Note that there need not be a unique solution.

3. Letting $n = 100$, plot the log-likelihood (with ϕ fixed at 0.5) as a function of α on the x-axis, with 1000 values of α uniformly spaced between 0 and 1. Generate a separate graph for each of the following values of $n_S : n_S = 90, n_S = 70, n_S = 50, n_S = 10$. Comment on the shapes that you see for the 4 plots. Please put all 4 plots on a single page to help with grading and a standard grid in the background of your plot.