CS 274A Homework 2

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2024

Due: 9am Monday 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 (10 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, do not look at (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 (e.g., Wikipedia).

Recommended Reading for Homework 2: Note Set 3

Problem 1: Maximum Likelihood for the Multinomial Model

Consider building a probabilistic model for the probability of words occurring in a particular context (e.g., in medical documents or patent documents). Let W be a discrete random variable, taking M values $w \in \{w_1, \ldots, w_M\}$. The values w_i could be different words for example, where M is the vocabulary size, and w_M represents "all other words" that are not in the set $\{w_1, \ldots, w_{M-1}\}$. In language and speech applications M can be very large, e.g., $M = 100,000$. 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 likelihood 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. Define the likelihood function 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 an unknown parameter θ, where $p(x) = \frac{1}{\theta} e^{-\frac{1}{\theta}x}$ with $\theta > 0, x \geq 0$. (The exponential density can also be written as $p(x) = \lambda e^{-\lambda x}$ with $\lambda = \frac{1}{\theta}$; in this problem you will be looking at the log-likelihood as a function of θ and not λ).
   (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 that this variance σ^2 (and the value or m) should be computed once and the resulting single value of σ^2 used in all your plot when plotting $l(\mu)$.

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 plot 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 generate new plots of the same type as above (with plots for exponential and Gaussian) for a different 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.

No need to submit your code for this problem, just the graphs and your comments with each graph.

Problem 3: Maximum Likelihood for the Poisson Model

Consider a data set $D = \{x_1, \ldots, x_N\}, x_i \in \{0, 1, 2, \ldots\}$, i.e., can take the value of any non-negative integer. The x_i's could for example be counts of the number of times a certain event occurs for each of a set of patients $i = 1, \ldots, n$ where each patient has at least one event. Assume a Poisson model for each x_i, defined as

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!},$$

with parameter $\lambda > 0$ and where $x \in \{0, 1, 2, 3, \ldots\}$.

1. Assume an IID likelihood and define the likelihood function for this problem

2. Derive the maximum likelihood estimate for λ

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 Graph Data

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 undirected 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 a single 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 (this is known as the Erdos-Renyi graph model). Once we fit this simple probabilistic model we could use it for example to simulate other graphs, test hypotheses about the graph, and so on.

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. Our data is now the set $D = \{D_1, \ldots, D_K\}$. In our model we will still use a single parameter θ for all graphs, i.e., each graph is assumed to be generated independently (given θ) in the same manner as earlier in this problem. 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.

   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, e.g., if this is not the maximum likelihood estimator then derive the correct one.

Problem 6: Maximum Likelihood with Measurement Variance per Point

Consider a data set D consisting of N scalar measurements $x_i, 1 \leq i \leq N$, where each measurement is taken from a different Gaussian, such that each Gaussian has the same mean μ, and each Gaussian has a different variance $\sigma_i^2, 1 \leq i \leq N$, where these N variances are known. For example, this might be an astronomy problem where we are trying to estimate the brightness μ of a star and our data consists of measurements x_i taken at different locations i on the planet where noise σ_i^2 per datapoint varies due to the local atmosphere (in a known way) with location i.

- Define the log-likelihood for this problem.
- Derive the maximum likelihood estimator for μ.
- Comment on the functional form of your solution: for example, can you interpret the result in the form of a weighted estimate? what are the weights?