CS 274A Homework 2

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2026

Due: Noon Monday January 26th, 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).
- Please be sure to read the academic integrity policy on the course Web page, including in particular the policies on restrictions for use of generative AI for homework assignments.
- 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 may 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 submit 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 should provide 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).
- Please read each problem carefully. If you believe there is a typo, or some information is missing, or the problem is unclear, please post a question on the Ed discussion board.

Recommended Reading for Homework 2: Note Set 3

Problem 1: Maximum Likelihood for the Multinomial Model

Consider building a probabilistic model for the marginal probability of words 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 represent different words from a predefined set called the vocabulary where M is the vocabulary size. 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 $\sum_{m=1}^{M} \theta_m = 1$. For simplicity we will assume that observations (individual words) are sampled in an IID manner from W: we will 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 outcome w_m occurred, $m = 1, \ldots, M$ in the data (these are known as the sufficient statistics for this model).

1. Define the likelihood function for this problem (you can ignore constants that don't depend on θ)

2. Derive the maximum likelihood estimates for each of the θ_m's for this model. Note that in the optimization you need to account for the constraint that $\sum_{m=1}^{M} \theta_m = 1$. You may want to read about the Lagrange multiplier method for constrained optimization, e.g., the Wikipedia page "Lagrange Multiplier" (e.g., Example 1).

Problem 2: 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). Assume an IID likelihood.
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 3: Visualization of Likelihoods

Consider a dataset $D = \{10, 5, 9, 48, 13, 29, 4, 6, 13, 11\}$, 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 per dataset and then fixed during plotting, with the fixed value of σ^2 used when plotting $l(\mu)$ curve for the Gaussian log-likelihood.

   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 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 the 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 $D_2 = \{3, 25, 11, 36, 5, 17, 23, 18, 29, 14, 20\}$, computing σ^2 now using D_2, and again overlaying the two uniform logL values. Again plot the log-likelihood functions over the range $\theta \in [3, 40]$. Comment briefly on this new graph, e.g., how its different from the first graph.

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

To check that your answers are on roughly the correct scale, for the first dataset the log-likelihoods should range between -60 and -35 and for the second dataset should range between -80 and -40 (roughly, across the set of models).

Problem 4: Bernoulli Parameters

Consider 2 data sets $D_1 = \{x_1, \ldots, x_{N_1}\}, x_i \in \{0, 1\}$ and $D_2 = \{x_1, \ldots, x_{N_2}\}, x_j \in \{0, 1\}$.

1. In the first part of the problem, the likelihood model for each data set is IID Bernoulli, where the Bernoulli parameter for $P(x_i = 1)$ in D_1 is θ and the Bernoulli parameter for $P(x_j = 1)$ in D_2 is $(1 - \theta)$. Imagine for example that we have IID data from two different experiments, where the probability of heads is exactly the opposite in the 2nd experiment relative to the first.

   (a) Clearly define the likelihood for this problem

   (b) Derive the maximum likelihood solution for θ_{ML}.

2. In the second part of the problem, the likelihood model for each data set is again IID Bernoulli, but now with parameter θ for $P(x_i = 1)$ in D_1 and parameter θ^2 for $P(x_j = 1)$ in D_2, i.e., IID data from two different experiments, where the Bernoulli parameter for the second is the square of the Bernoulli parameter for the first. Show that the solving for θ_{ML} requires obtaining the solution of a polynomial equation in θ: you don't need to solve for θ_{ML}, just show that the functional form of the solution is a polynomial.

3. In the third part of this problem, we just have a single dataset $D = \{x'_1, \ldots, x'_N\}$. This data was generated in the following way. A coin is tossed (in an IID manner) N times, with $P(x_i = 1) = \theta$. However, we don't directly observe whether the outcome for each toss is $x_i = 1$ or $x_i = 0$. Instead, for each toss, we have a noisy observer (could be a person or a measuring instrument) that reports x'_i, where $x'_i = x_i$ with probability α and $x'_i = 1 - x_i$ with probability $1 - \alpha$ and where we assume $0.5 < \alpha \leq 1$. We also assume that we know the value of α.

   (a) Clearly define the likelihood for this problem

   (b) Derive the maximum likelihood solution for θ_{ML}.

Problem 5: Maximum Likelihood for a Gaussian with a Scaled Mean

Let X and Y be two real-valued random variables, taking values x and y respectively, where $x \sim N(\mu, \sigma^2)$ and where $y \sim N(a\mu, \sigma^2)$, where $a > 0$ is some known constant. Assume that X and Y are conditionally independent given μ and σ^2. For reference, the Gaussian (Normal) density $z \sim N(\mu_z, \sigma_z^2)$, for some variable Z taking values z, is defined as

$$p(z) = \frac{1}{\sqrt{2\pi\sigma_z^2}} e^{-\frac{1}{2\sigma_z^2}(z-\mu_z)^2}$$

Say we have two datasets, $D_x = \{x_i\}, i = 1, \ldots, N$, and $D_y = \{y_j\}, j = 1, \ldots, M$, and where we use $D = \{D_x, D_y\}$ to denote all of the data. The x_i's and y_j's are assumed to be conditionally independent given the parameters θ.

Let $\theta_1 = \mu$ and $\theta_2 = \sigma^2$, with $\theta = \{\theta_1, \theta_2\}$. Answer the following questions:

1. Clearly define the likelihood $L(\theta) = P(D|\theta) = P(D_x, D_y|\theta)$, using the information provided above
2. Using the log-likelihood, derive maximum likelihood estimates for each of the parameters θ_1 and θ_2.

Problem 6: Maximum Likelihood with Measurement Variance per Point

Consider a data set D consisting of N scalar measurements (data) $x_i, 1 \leq i \leq N$, 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. You can assume that the data x_i are conditionally independent given μ and the σ_i's. As an 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?

Problem 7: Estimation of Mixtures of Densities

Let $x \geq 0$ be the value of a real-valued scalar random variable X that is distributed as a mixture of two exponentials:

$$p(x) \;=\; \alpha f(x; \lambda) + (1-\alpha) g(x; c\lambda)$$

where each of f and g are exponential densities, parametrized by λ and $c\lambda$ respectively, i.e.,

$$f(x; \lambda) \;=\; \lambda e^{-\lambda x}, \quad g(x; c\lambda) \;=\; (c\lambda) e^{-c\lambda x}$$

where α is a mixing weight, $0 \leq \alpha \leq 1$, and where $\lambda > 0$, and where $c > 0$ is some constant.

Say we have a data set $D = \{(x_i, z_i)\}$, $i = 1, \ldots, N$, consisting of IID draws from $p(x, y)$, where $z_i \in \{0, 1\}$ and $z_i = 1$ indicates that x_i was generated by component f, and $z_i = 0$ indicates that x_i was generated by component g. Let $n = \sum_{i=1}^{N} z_i$.

Answer the following questions:

1. Assume λ and α are unknown and that c is known. Clearly define the likelihood $L(\lambda, \alpha)$ for this problem in terms of the information provided above.

2. Using the likelihood (or log-likelihood) from the previous part, derive the maximum likelihood estimators for λ and α. Explain every step in your solution clearly.

Problem 8: Maximum Likelihood for Variable-Length Sequences

Consider a Markov chain with 3 states and the following transition matrix, where the rows indicate the transition probabilities from states 1, 2, and 3 to the next state:

$$\begin{pmatrix} \alpha & \beta & \gamma \\ \beta & \alpha & \gamma \\ 0 & 0 & 1 \end{pmatrix}$$

Let $\theta = \{\alpha, \beta, \gamma\}$ be the 3 unknown parameters. States 1 and 2 each have self-transition probabilities of α; and each has transition probability β of transitioning to the other. Because the parameters don't depend on sequence position, this is known as a *homogeneous* first-order Markov chain.

State 3 is what is known as an "absorbing state." In this model, once the Markov chain arrives at state 3 it halts and produces no more states, i.e., if we generate data from this model we will have finite length sequences where the last state is always state 3 (and the last state is the only occurrence of state 3 in a sequence), e.g., a possible sequence could be $s = [1121222113]$.

Consider an observed data set D consisting of M sequences $\{s_1, \ldots, s_M\}$ where the sequences can be of different lengths. Let sequence s_m have $n_m + 1$ states, i.e., the sequence has length $n_m + 1$, with $m = 1, \ldots, M$. For each sequence s_m, the last state (the $(n_m + 1)$th state) is always 3.

You can assume that the initial state distribution is $\pi = [0.5, 0.5, 0]$, i.e., a sequence has a 50% chance of starting in either state 1 or 2 and 0% chance of starting in state 3. Each sequence s_m is assumed to be conditionally independent of all other sequences given the parameters θ.

In the problems below use the notation $r_{m,i,j}$ to refer to the number of transitions from state i to j observed in sequence m. Terms such as $N = \sum_{m=1}^{M} n_m$ may also be useful to define to simplify notation.

1. Clearly define the likelihood $L(\theta)$ for this problem.

2. Derive maximum likelihood estimates for each of the 3 unknown parameters.

3. Explain in words why it would be suboptimal to (i) estimate maximum likelihood parameters separately for each sequence and then (ii) to obtain an overall estimate that is the average of these parameters (e.g., mention a simple example of a dataset to illustrate your point).