CS 274A Homework 2

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2020

Due: Noon Wednesday January 29th, submit via Gradescope

Instructions and Guidelines for Homeworks

- Please answer all of the questions and submit a scanned copy of your written solutions to Gradescope (either hand-written or typed are fine as long as the writing is legible).
- All problems are worth 10 points unless otherwise stated. All homeworks will get equal weight in computation of the final grade for the class.
- The homeworks are intended to help you work through the concepts we discuss in class in more detail. It is important that you try to solve the problems yourself. The homework problems are important to help you better 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.
- If you can't solve a problem, you can discuss it *verbally* with another student. However, please note that before you submit your homework solutions 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, and so forth. The work you hand in should be your own original work.
- You are allowed to use reference materials in your solutions, such as class notes, textbooks, other reference material (e.g., from the Web), or solutions to other problems in the homework. It is strongly recommended that you first try to solve the problem yourself, without resorting to looking up solutions elsewhere. If you base your solution on material that we did not discuss in class, or is not in the class notes, then you need to clearly provide a reference, e.g., "based on material in Section 2.2 in ....."
- In problems that ask for a proof 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 wish to use LaTeX to write up your solutions you may find it useful to use the .tex file for this homework that is posted on the Web page. And please feel free to submit the .tex (as well as the .pdf) file for your solutions —it may be helpful to us when we send out solutions later on.

Suggested reading for Homework 2:

- Note Set 3 on the Course Web page

Problem 1: (Hidden Markov Models)

Hidden Markov models are widely used in speech recognition, language modeling, genomics, time-series modeling, and many other applications. To define a hidden Markov model (HMM) we have two types of random variables:

1. $X_1, \ldots, X_T$ the hidden states, and
2. $Y_1, \ldots, Y_T$ the observations

The index $t = 1, \ldots, T$ could be discrete-time (e.g., every day) or just be an index on the order or position in a sequence—we will just refer to t as time below. We will assume below that each X_t variable is discrete taking K values (sometimes these variables are referred to as "states"). In general each Y_t variable can be a vector of length d with components that can be discrete or continuous or some combination.

There are two conditional independence assumptions in an HMM. The first is a Markov assumption on the hidden X variables:

$$P(X_t|X_{t-1}, \ldots, X_1) = P(X_t|X_{t-1}), \quad t = 2, \ldots, T$$

and where $P(X_1)$ has its own distribution (referred to as the initial state distribution).

The second assumption is that each Y_t is conditionally independent of all other variables given X_t, i.e.,

$$P(Y_t|X_1, \ldots, X_T, \text{other Y's}) = P(Y_t|X_t), \quad t = 1, \ldots, T$$

Thus, each Y_t only depends on the state X_t at time t.

The parameters of an HMM consist of: (i) an initial state distribution $P(X_1)$, (ii) a transition matrix $P(X_{t+1} = j|X_t = i), 1 \leq i, j, \leq K$, and (iii) parameters for the conditional densities $P(Y_t|X_t = k)$ of the observations given each possible state value, $1 \leq k \leq K$. The model is called *homogeneous* when $P(X_{t+1} = j|X_t = i)$ and $P(Y_t|X_t = k)$ do not change as a function of t. Below you can assume the HMM is homogeneous.

1. Draw a picture of the graphical model for $T = 5$ and write down an equation for the joint distribution of X's and Y's based on the graphical model structure.
2. Let each $P(Y_t|X_t = k)$ be a Gaussian density (for real-valued y_t's) for two different cases: (a) diagonal covariance matrices, (b) full covariance matrices, where the y_t's are d-dimensional real-valued vectors. For each case (a) and (b) define precisely how many parameters we need in order to specify the HMM.

3. Using $T = 5$, show systematically how one can compute $P(x_5 = k|y_1^*, \ldots, y_5^*),\ 1 \leq k \leq K$ where $y_1^*, \ldots, y_5^*$ are some observed values for the y's.

   Hint 1: work with joint probabilities and then normalize to get the conditional probability of interest at the end.

   Hint 2: start with $P(X_2, y_1^*, y_2^*)$, use this to find $P(X_3, y_1^*, y_2^*, y_3^*)$, and so on.

   You can use $\sum_{x_t}$ or $\int_{y_t} dy_t$ to indicate summing or integrating over variables such as x_t or y_t respectively. Note that this uses similar ideas to those used to solve the last problem on HW1.

Note: In the next few problems below, unless otherwise stated, the observations in a data set D are assumed to be conditionally independent given parameters θ.

Problem 2: (Likelihood Functions)

Consider the following data set $D = \{4, 8, 6, 8, 9, 12, 10, 6, 9, 7\}$. Use MATLAB (or Python, or R, or something similar) to generate graphs of the log-likelihood function for each of the following cases:

1. a Gaussian density with μ as the unknown parameter in the log-likelihood function and with a fixed standard deviation of $\sigma = 3$.
2. a uniform density with $a = 3$ and b as the unknown parameter in the log-likelihood function
3. an exponential density with the exponential parameter as the unknown parameter in the log-likelihood function.

In each case you can the plot a range of values around the mode of the log-likelihood function, e.g., if θ is the mode you could plot in the range $[0.2\theta, 2\theta]$ (or something similar). Comment on the shape of each of the 3 plots.

Problem 3: (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 $w \in \{w_1, \ldots, w_V\}$, where V is the number of words in the vocabulary. In practice V can be very large, e.g., $V = 100,000$ is not unusual (there are more words than this in English, but many rare words are not modeled).

The *multinomial model* for W is essentially the same as the binomial model for tossing coins, where we have independent trials, but instead of two possible outcomes there are now V possible outcomes for each "trial". The parameters of the multinomial are $\theta = \{\theta_1, \ldots, \theta_V\}$, where $\theta_k = P(W = w_k)$, and where

$\sum_{k=1}^{V} \theta_k = 1$. Denote the observed data as $D = \{r_1, \ldots, r_V\}$, where r_k is the number of times word k occurred in the data (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 the θ_k's for this model.

Problem 4: (Maximum Likelihood for the Geometric Model)

Consider a data set $D = \{x_1, \ldots, x_n\}, x_i \in \{1, 2, \ldots\}$. Assume a geometric model for the data, where the geometric distribution with parameter θ is defined as

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

1. Define the likelihood function for this problem
2. Derive the maximum likelihood estimate for θ

Problem 5: (Maximum Likelihood for the Poisson Model)

Consider a data set $D = \{x_1, \ldots, x_n\}, x_i \in \{0, 1, 2, \ldots\}$. Assume a Poisson model for the data, defined as

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

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

1. Define the likelihood function for this problem
2. Derive the maximum likelihood estimate for λ

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?

Problem 7: (Maximum Likelihood: Comparing Models)

Assume you are working for a large search engine company and you wish to model the distribution of the number of search results that a user clicks on for a typical search. Let X be a random variable taking values $k = 0, 1, 2, \ldots ...$, where k represents the number of clicks. $P(X)$ represents the distribution of number of clicks for a randomly selected user.

1. Let the data set D consist of observations from $N = 100$ users, summarized by the following table of values:

value k	r_k, number of users with this value k
0	10
1	17
2	29
3	23
4	10
5	7
6	3
7	1

   Letting r_k be the number of users with k clicks, consider both a Poisson model and for a geometric model for this data (here our geometric model for X starts at 0 rather than 1, so the appropriate definition is $P(X) = (1 - \theta)^k \theta$).

2. Plot the log-likelihood for each of the geometric and Poisson models as a function of their respective parameters for this data set. Indicate clearly on each plot (i) where the maximum likelihood estimate is (on the x-axis) and (ii) what its value is (on the y-axis). You can use the maximum likelihood estimates of the parameters for this data (based on what you derived in the earlier problems, or by looking it up if you did not do the earlier problems).

3. Using (again) the maximum likelihood estimates of the parameters, on a single plot with the x-axis running from 0 to 10, i.e., $X \in [0, 10]$, plot the following:

   - The empirical probability (from the data) of each value
   - The probability distribution for the geometric model
   - The probability distribution for the Poisson model

4. Is the geometric or Poisson model a better fit to this data? Hint: you can use the maximizing value of log-likelihood of the observed data for each model as a metric to help assess which model has a better fit.

5. If we used another model that has two parameters instead of one parameter (e.g., a Negative Binomial distribution or a zero-inflated Poisson model), and fit such a model to the data and calculated log-likelihood using the maximum likelihood parameters, could we compare this log-likelihood value with those you computed above to select the "best" model out of the 3? Explain your answer in a few sentences. If your answer is "no" sketch what you believe would be a fair method to select best model (from the two models with a single parameter and the third model with 2 parameters).

You will likely want to use R or Python or MATLAB to generate the plots above if you wish (you do not need to submit any code with your solution but do submit your plots).

Problem 8: (Method of Moments and the Uniform Model)

The method of moments is an alternative parameter estimation 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 $x_i, 1 \leq i \leq N$.

1. Derive estimators for a and b using the method of moments. Since there are $K = 2$ unknown parameters this means that 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).

3. Write 2 or 3 sentences comparing the properties of the maximum likelihood solutions with the method of moment solutions. You can use the following simple data set $D = \{3, 4, 5, 10, 3, 5, 8, 6\}$ to provide some intuition for your answer.