CS 274A Homework 2

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2017

Due Date: 5pm Friday January 27th (to be dropped off at the TA's office between 3pm and 5pm)

Instructions and Guidelines for Homeworks

- Please answer all of the questions and submit a **hardcopy** of your written solutions (either handwritten or typed are fine as long as the writing is legible). Clearly mark your name on the first page.
- 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 any material that we did not discuss in class, or is not in the class notes, then you need to 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.

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 a mixture of each.

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. In all the problems below you can assume the HMM is homogeneous.

In answering this question and the next one feel free to consult Note Set 5 on HMMs (on the class Website)—it uses different notation to the notation here. This is just an option—its not necessary to read Note Set 5 to answer these 2 questions.

1. Draw a picture of the graphical model and write down an equation for the joint distribution of X's and Y's based on the graphical model structure.
2. Say $P(Y_t|X_t)$ is Gaussian density (for real-valued y_t's) for 2 different cases: (a) diagonal covariance matrices, (b) full covariance matrices, and the y_t's a d-dimensional real-valued vector. In each case (a) and (b) define precisely how many parameters we needed to specify the HMM.

Problem 2: (Hidden Markov Models)

This is a continuation of the last problem.

1. Consider a small toy example with $T = 5$, i.e., sequences of length 5. Show systematically how one can compute $P(X_5|y_1, \ldots, y_5)$. Hint 1: work with joint probabilities and then normalize over Y_5's values 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.
2. Now say we know that $X_1 = x_1^*$ and $X_3 = x_3^*$ for some values of X_1 and X_3. Explain clearly how this would modify your solution to the last problem.
3. Describe briefly how you could compute $P(X_3|y_1, \ldots, y_5)$ given $X_4 = x_4^*$ and $X_1 = x_1^*$. (No need to show all the details, just clearly describe what the key steps are in doing this calculation).

Problem 3: (Mixture Models)

Finite mixture models show up in a wide variety of contexts in machine learning and statistics (we will discuss them in more detail in lectures later in the quarter). In this problem consider a real-valued random variable X taking values x (in general we can define mixtures on vectors, but here we will just consider the 1-dimensional scalar case). The basic idea is to define a density (or distribution) $p(x)$ that is a weighted mixture of component densities $f_k(x)$ where the weights are non-negative and sum to 1, i.e.,

$$f(x) = \sum_{k=1}^{K} \alpha_k f_k(x; \theta_k)$$

where

- the weights obey the following conditions: $\sum_{k=1}^{K} \alpha_k = 1, \ \ 0 \leq \alpha_k \leq 1$
- each $p_k(x; \theta_k)$ is itself a probability density function with its own parameters θ_k. For example, if a component is Gaussian then $\theta_k = \{\mu_k, \sigma_k^2\}$.

1. Given the properties above prove that a finite mixture $p(x)$ is itself a density function, i.e., it obeys all the necessary properties needed to be a density function.
2. If each of the components $p_k(x)$ are unimodal (i.e., have unique mode), how many modes in general can the mixture have (assume here that $\alpha_k > 0, 1 \leq k \leq K$.
3. Derive an expression for the (a) mean μ of $p(x)$, and (b) the the variance σ^2 of $p(x)$, as a function of the component weights, means and variances $\alpha_k, \mu_k, \sigma_k^2, 1 \leq k \leq K$. For each of μ and σ^2 provide an intuitive interpretation in words of your solutions.

Problem 4: (Maximum Likelihood)

Consider the following data set $D = \{4, 15, 6, 8, 9, 12, 10, 6, 9, 7\}$. 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 model with μ as the unknown parameter in the log-likelihood function and with a fixed standard deviation of $\sigma = 3$.
2. a uniform distribution with $a = 2$ and b as the unknown parameter in the log-likelihood function
3. an exponential distribution 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, e.g., if θ is the mode you could plot in the range $[0.2\theta, 2\theta]$. Comment on the shape of each of the 3 plots. Please hand in a hardcopy of your graphs with your homework.

Note: In the next several problems below assume that a data set $D = \{x_1, \ldots, x_n\}$ exists. You can also assume that the x_i's are conditionally independent of each other given the parameters of the model.

Problem 5: (Maximum Likelihood)

Let $f(x; \theta)$ be a Gaussian density function, i.e.,

$$f(x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Derive formulas defining the maximum likelihood estimates of μ and σ^2.

Problem 6: (Maximum Likelihood)

Derive the maximum-likelihood estimator for the geometric distribution with parameter p, where the geometric distribution is defined as

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

Here the $x_i \in \{1, 2, 3, \ldots, \}$

Problem 7: (Maximum Likelihood)

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 the sufficient statistics for this model).

Derive the maximum likelihood estimates for each θ_k for this model.

Problem 8: (Maximum Likelihood)

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.

1. Derive maximum likelihood estimators for a and b (think carefully about how to do this).
2. Write 2 or 3 sentences discussing these maximum likelihood estimators, e.g., do they make intuitive sense? if not, briefly describe an alternative method for estimating the parameters.

Problem 9: (Maximum Likelihood)

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 σ_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 due to the local atmosphere σ_i^2 varies (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?