CS 274A Homework 1

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2018

Due Date: Wednesday January 17th, submit hardcopy at the start of class

Instructions and Guidelines for Homeworks

- Please answer all of the questions and submit a **hardcopy** of your written solutions **at the start of class on the due date** (either hand-written 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 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.

If you need to brush up on your knowledge of probability, reading Note Sets 1 and 2 from the class Web page is recommended before attempting the problems below.

Problem 1: (Linearity of Expectation)

The expected value of a continuous random variable X, taking values x, is defined as $\mu_x = E[X] = \int p(x)\ x\ dx$ where $p(x)$ is the probability density function for X. The variance is defined as $var(X) = E[(X - \mu_x)^2] = \int p(x)(x - \mu_x)^2 dx$ (often also denoted as σ_x^2).

1. Prove that expectation is linear, i.e., that $E[aX + b] = aE[X] + b$ where a and b are constants.
2. Prove that $var(cX) = c^2 var(X)$ where c is a constant.
3. Prove that $var(X) = E[X^2] - (E[X])^2$.

Problem 2: (Uniform Density)

Let X be a continuous random variable with uniform density $U(a, b)$, with $a < b$, i.e.,

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

if $a \leq x \leq b$ and $P(x) = 0$ otherwise.

1. Derive an expression for $E[X]$.
2. Derive an expression for $var(X)$.

Problem 3: (Geometric Model)

Suppose that we repeatedly toss a coin (with no memory in the coin, so that tosses are independent) until we get the outcome of "heads." Let θ be the probability of heads on any toss. Let X be the number of such tosses until a heads occurs. This type of model can be used to describe "independent trials" and is frequently used in science and engineering to model various simple repetitive phenomena (e.g., how many times one uses a device until it breaks, the number of consecutive days of rainfall at a given location, and so on).

In this situation X is a discrete random variable with a geometric distribution taking values $x \in \{\{1, 2, 3, \ldots, \}$, with a probability distribution defined as $P(x) = (1 - \theta)^{x-1}\theta$. Here θ is the parameter of the geometric model (e.g., the probability of heads in coin-tossing) and $0 < \theta < 1$.

1. Prove that $\sum_{x=1}^{\infty} P(x) = 1$.
2. Derive an expression for the expected value of X, $\mu_X = E[X]$.
3. Derive an expression for the variance of X, where the variance is $\sigma_x^2 = E[(X - \mu_x)^2]$.

Problem 4: (Central Limit Theorem)

Let $X_1, \ldots, X_n$ be a set of independent and identically distributed real-valued random variables each with the same density $P(X)$ where $P(X)$ has mean μ and variance σ^2.

1. State the central limit theorem as it applies to $X_1, \ldots, X_n$ (if you don't know the central limit theorem you will need to look it up)
2. Let $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$ where each X_i has a uniform distribution between 0 and 1. Simulate 1000 values of Y for each of the following values of n, $n = 10^2, 10^3, 10^4, 10^5$. (So you should end up with 4 sets of Y values, each with 1000 values). For example you can use the `rand.m` function in Matlab to do this, or similar functions in R or Python. Generate histogram plots of the 4 results (e.g., using the `hist.m` function in MATLAB) for each value of n (this will produce 4 histograms). Use $\sqrt{1000} \approx 30$ bins in your histograms.
3. Based on visual inspection of the histograms, comment on how your simulated data matches the central limit theorem.
4. Quantitatively evaluate how well your empirically simulated distributions match what the theory predicts (e.g., compare the mean and variance of the simulated data with that from theory).

Problem 5: (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), where the density function for X is a *finite mixture* defined as

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

where the α_k's are the *mixing weights*, with $0 \leq \alpha_k \leq 1$ and $\sum_{k=1}^{K} \alpha_k = 1$, and where the $f_k(x; \theta_k)$'s are each probability densities (known as the *components* of the mixture model) with parameters θ_k. For example, a mixture of two Gaussians would have $K = 2$, with f_1 and f_2 each being Gaussians, and where each Gaussian density could have its own mean and variance parameters.

1. If each component densitiy $f_k(x)$ is unimodal, what is the minimum number of modes that a mixture can have? and what is the maximum number?
2. Express the expected value μ of the mixture model f in terms of the mixing weights α_k and the component means μ_k, $1 \leq k \leq K$.
3. Express the standard deviation σ of the mixture model f in terms of the mixing weights α_k and the component means μ_k and standard deviations σ_k, $1 \leq k \leq K$.

Problem 6: (Bayes Rule with Gaussians)

In Example 7 in Note Set 1 (two Gaussians with equal variance) prove that:

$$P(a_1|x) \;=\; \frac{1}{1+e^{-(\alpha_0+\alpha x)}}$$

and derive expressions for each of α_0 and α as a function of the two means μ_1 and μ_2, the variance σ^2, and the probabilities $P(a_1)$ and $P(a_2)$.

The expression for $P(a_1|x)$ above is known as the logistic function and shows up frequently in machine learning (even when we don't make Gaussian assumptions about x). Say that $\mu_1 = 5, \mu_2 = 10$ and $P(a_1) = P(a_2) = 0.5$. Plot the logistic functions for each of the cases (1) $\sigma = 5$, (2) $\sigma = 3$, and (3) $\sigma = 1$. Put all 3 functions on the same plot centered around the point where $P(a_1|x) = 0.5$. Comment on the shapes of the functions.

Problem 7: (Markov Model)

Let $X_1, X_2, \ldots, X_T$ be a set of discrete-valued random variables, each taking K possible values from 1 to K. For example, X_i might be the word in the ith sequential position in a piece of text consisting of T occurrences of words. Or X_i might be the state of some physical system at discrete time i, e.g., whether it rained or not on day i.

The variables $X_i, i = 1, \ldots, T$ are said to obey the first-order Markov property if

$$P(X_i|X_{i-1}, X_{i-2}, \ldots, X_1) = P(X_i|X_{i-1}), \qquad i = 2, \ldots, T,$$

i.e., the only information about the next word X_i (from the "history" of random variables preceding X_i) is contained in X_{i-1}. Or, equivalently, given knowledge of X_{i-1}, there is no additional information about X_i in any of the preceding words.

Say we have observed the words $x_1, \ldots, x_i$ (but not any words after x_i in the sequence), and we wish to predict X_{i+m}, for some $m > 1$.

1. Use the law of total probability and the first-order Markov property to derive an efficient way to compute $P(X_{i+m}|x_i, \ldots, x_1)$.
2. What is the time complexity of this computation as a function of m and K? (in "big O" notation)?

In answering this problem you can assume that the Markov chain is *homogeneous*, i.e., that the transition probabilities $P(X_i|X_{i-1})$ are the same for all values of i.

Problem 8: (Naive Bayes Classification Model)

The naive Bayes model is a probability model with a class variable C taking M possible values $c \in \{1, \ldots, M\}$ and d features $X_1, \ldots, X_d$. For simplicity we will assume that each of the X_j variables are discrete and each takes K possible values $x_j \in \{1, \ldots, K\}$. Each feature is conditionally independent of all the other features given C.

1. Write down the correct expression for the joint distribution $P(C, X_1, \ldots, X_d)$ for this model.
2. Draw a picture of the graphical model for the case of $d = 3$.
3. Specify exactly how many parameters are needed for this model in the genereal case, as a function of M, K, and d. A *parameter* in this context is any probability or conditional probability value that is needed to specify the model.

Problem 9: (Graphical Model 1)

Consider a directed graphical model with random variables A, B, C, D, E, F where F has parent E, E has parents C and B, D has parent C, C has parents A and B, and A and B each have no parents. Assume that each variable can take K values, $K \geq 2$.

1. Draw a diagram showing the structure of this graphical model and write down an expression for the joint distribution $P(a, b, c, d, e, f)$ as represented by this graphical model.
2. Specify precisely how many parameters (probabilities) are needed to specify this model. A parameter is defined (for this problem) as a conditional probability or a marginal (unconditional) probability.
3. How many parameters would be required if we had a saturated model? (i.e., a model with no conditional independencies assumed).

Problem 10: (Graphical Model 2)

Consider another directed graphical model with discrete random variables X, Y, V, Z where X has no parents, Y and V each have X as a parent, and Z has Y as a parent. You can assume that each variable takes K values, $K \geq 2$.

1. Suppose a value x is observed for X and we don't know the values of any of the other variables. Given that x is known, show how one would use the structure of the graphical model to compute $P(z|x)$ for any value z of the variable Z. In particular, prove that this computation does not depend in any way on the conditional probability table $P(V|X)$.
2. Now say that both x and v (some value for V) are observed. How does this change the answer to part 1?

3. Now say v is observed but x is not. Show how you could compute $P(z|v)$ in a step by step manner by first computing $P(x|v)$, then computing $P(y|v)$, and finally computing $P(z|v)$, at each stage using the information from the previous step.