CS 274A Homework 1

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

Due Date: Wednesday January 18th, 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 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.

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:

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:

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

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

Problem 3:

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:

Let X be a random variable taking values $x \in \{0, 1, 2, 3, \ldots, n\}$, where X has a binomial distribution, $X \sim Bin(n, \theta)$, with parameters n and θ. Here $n \in \{1, 2, 3, \ldots\}$ is the number of independent trials, and θ is the probability of success on each trial with $0 \leq \theta \leq 1$. The value of X is the number of successes in n independent trials. A simple example would be tossing a coin n times with $\theta = 0.5$, where X would be the number of successes (e.g., "heads") that are observed.

1. Define the probability distribution $P(x), x = 0, 1, 2, \ldots, n$ and explain clearly why this definition is the correct one.
2. Prove that $\sum_{x=0}^{n} P(x) = 1$.
3. Derive an expression for the expected value of X, $\mu_X = E[X]$.
4. Derive an expression for the variance of X, where the variance is $\sigma_x^2 = E[(X - \mu_x)^2]$.

Problem 5:

Let X be a categorical random variable taking values $\{1, \ldots, M\}$ and with probability distribution $P(X = 1), P(X = 2), \ldots, P(X = M)$ where $\sum_k P(X = k) = 1$.

1. Define the entropy of X, $H(X)$.
2. Explain clearly how the entropy can be viewed as an expectation with respect to $P(X)$.
3. Prove that $H(X) \geq 0$ and explain under what conditions this lower bound will be achieved.
4. Prove that $H(X) \leq \log(M)$ and explain under what conditions this upper bound will be achieved.
5. Generate and submit a contour plot of $H(X)$ over the 2-dimensional simplex that defines $P(X = 1), P(X = 2), P(X = 3)$ for a random variable X taking one of the $M = 3$ values. The simplex is the 2-dimensional triangular region defined by the following two constraints:

   (a) $0 \leq P(X = k) \leq 1, \forall k$,

   (b) $\sum_{k=1}^{M} P(X = k) = 1$.

   Indicate on the plot where in the simplex $H(X)$ attains both its maximum and minimum values. You can use Matlab, R, Python, etc., to generate the plot (no need to submit your code).

Problem 6:

Let $X_1, \ldots, X_n$ be a set of independent and identically distributed random variables each with the same distribution $P(X)$.

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 7:

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 α_0 and α.

Problem 8:

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 $d = 4$.

3. Specify exactly how many parameters are needed for this model 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:

Let A, B, C, D, E, F, G be discrete random variables, each taking K possible values. You are given a graphical model with the following information: D and F have no parents, D is the parent of B and of E, F is the parent of E, B is the parent of A and G, and E is the parent of C and G.

1. Draw a diagram showing the structure of this graphical model and write down an expression for the joint distribution $P(a, b, c, d, e, f, g)$ as represented by this graphical model.

2. How many parameters (probability values) in total are required to specify this model? How many would be required if we had a saturated model?

3. In the rest of the problem assume that we have observed the value d for variable D, and that we wish to compute $P(A = a|D = d)$ for some value a of the variable A. Assume that the values of all of the other variables are unknown. If were given a table with the full joint distribution $P(a, b, c, d, e, f, g)$ as a table (and did not know anything about the graphical model structure) write down an equation that shows how we would use this table to compute $P(a|d)$ for specific values a and d. What is the time complexity ("big O" notation) of computing this as a function of K?

4. Now assume you know the graphical model structure and the associated set of conditional probability tables (expressed in the form of $P(var|parents)$). Show clearly, using equations in a step by step manner, how you can use the graph structure and the tables to compute $P(a|d)$ as efficiently as possible. Please be sure to show all steps. You will need to use Bayes rule and the law of total probability. What is the time complexity of computing $P(a|d)$ using the most efficient procedure (for this graph)?

Problem 10:

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.

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 $m > 1$.

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