CS 274A Homework 1

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2019

Due Date: 11am Monday January 14th, 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.

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: (Expectation and Variance)

The expected value of a real-valued 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). In the questions below a and b are scalar constants (i.e., not random variables).

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

Problem 2: (Multiple Random Variables)

Let X and Y be two real-valued random variables. In the equations below the expectation on the left is with respect to $p(x, y)$ and the expectations on the right are with respect to $p(x)$ and $p(y)$ respectively.

1. Prove that $E[aX + bY] = aE[X] + bE[Y]$.
2. Prove that if X and Y are independent that $var(aX + bY) = a^2 var(X) + b^2 var(Y)$.

Problem 3: (Poisson Distribution)

Let Y be an integer-valued random variable taking values $y = 0, 1, 2, \ldots,$. For example, Y could be the number of clicks that users make during visits to a particular Web site. A well-known simple model for this type of "count data" is the Poisson distribution, with

$$P(Y = y) \ = \ e^{-y} \frac{\lambda^y}{y!}, \qquad y = 0, 1, 2, \ldots$$

where $\lambda > 0$ is the parameter for the distribution.

1. Prove that $\sum_{y=0}^{\infty} P(y) = 1$
2. Prove that $E[Y] = \lambda$
3. Prove that the variance $var[Y]$ is also equal to λ

Problem 4: (Entropy)

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 $\sum_m P(X = m) = 1$. The entropy of the random variable X is defined as the scalar-valued function $H(X) = \sum_{m=1}^{M} P(X = m) \log \frac{1}{P(X=m)}$. The base for the log function can be any base, but base 2 (bits) or base e ("nats") are commonly used. If $P(X = m) = 0$ for some value m, then the corresponding term in the entropy function is defined to be $P(X = m) \log \frac{1}{P(X=m)} = 0 \log 0 = 0$.

1. Explain clearly how the entropy can be viewed as an expectation with respect to $P(X)$.
2. Prove that $H(X) \geq 0$ and explain under what conditions this lower bound will be achieved.
3. Prove that $H(X) \leq \log(M)$ and explain under what conditions this upper bound will be achieved.
4. Generate a contour plot of $H(X)$ over the 2-dimensional simplex that defines the 2-dimensional region $P(X = 1), P(X = 2), P(X = 3)$, 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) or draw it manually as long its clear and reasonably accurate.

Problem 5: (Graphical Models)

Consider a directed graphical model with random variables A, B, C, D, E, F where F has parent E, E has parents C and A, D has parent B, 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 are in this graphical model. A parameter in this context is defined as any conditional probability or marginal (unconditional) probability that is needed to specify the model.
3. How many parameters would be required if we had a fully saturated model? (i.e., a model where no conditional independencies assumed).
4. Use both (a) the law of total probability, and (b) the structure of the graphical model, to show (in a series of equations) how one could use the structure of the model to compute the conditional distribution $p(e|a, b)$ for all values e and where a and b are some fixed observed values for A and B.

5. Show how you would update your answer in the previous question if you wanted to compute $p(e|a, b, d)$

6. Show how you would update your answer to part 4 if you wanted to compute $p(e|a, b, f)$

Problem 6: (Naive Bayes Classification Model)

The naive Bayes model is a probability model with a class variable C taking M possible values $\{1, \dots, M\}$ and d features $X_1, \dots, X_d$, where C and X's are all random variables. The key aspect of naive Bayes model is that each feature X_j is assumed to be conditionally independent of all the other features given C.

For example, C might represent different possible states of a patient in a medical diagnosis problem and the X's could be symptoms or features that could be measured for a patient. In classification problems we are usually interested in making a prediction about the value of the class C given observed values for the X's (the features).

Initially below we will assume that each of the X_j variables are discrete and each takes K possible values $x_j \in \{1, \dots, K\}$.

1. Draw a picture of the graphical model for the case of $d = 3$.

2. Write down an expression for the joint distribution $P(C, X_1, \dots, X_d)$ for this model.

3. Specify exactly how many parameters are needed for this model in the general 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.

4. Say we are given a naive Bayes model where the parameters are known. Say we observe a set of values $x_1, \dots, x_d$ for the features, but the value class variable C is unknown. Using the structure of the model, show clearly (with equations) how one can use the model to compute the conditional distribution $P(c = m|x_1, \dots, x_d), 1 \leq m \leq M$.

5. Now say that each of the d features $X_1, \dots, X_d$ are real-valued and that we assume that the conditional density for each feature given the class, $p(X_j|c = m)$ is a univariate Gaussian. Specify precisely how many parameters are needed for this Gaussian version of a naive Bayes, in the general case as a function of M, K, and d.

Problem 7: (Bayes Rule with Gaussians)

1. For 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 equations 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).

2. Say that $\mu_1 = 10, \mu_2 = 20$ and $P(a_1) = P(a_2) = 0.5$. Plot the logistic functions for each of the cases (1) $\sigma = 10$, (2) $\sigma = 5$, and (3) $\sigma = 1$. Put all 3 functions on the same plot centered around the point where $P(a_1|x) = 0.5$.

   Briefly comment on the shapes of the functions.

Problem 8: (Multivariate Gaussians)

Ad defined in Note Set 2, the joint multivariate (multidimensional) Gaussian density, for a d-dimensional vector of real-valued random variables $X_1, \ldots, X_d$, is defined as:

$$p(\underline{x}) = p(x_1, x_2, \ldots, x_d) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})}$$

where $\underline{x}$ is a d-dimensional vector (a set of possible values for $X_1, \ldots, X_d$), $\underline{\mu} = (\mu_1, \ldots, \mu_d)$ is a $d \times 1$ dimensional vector of mean values and Σ is a $d \times d$ positive semi-definite symmetric covariance matrix with entries $\sigma_{ij} = \sigma_{ji} = cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$ with $1 \leq i, j \leq d$. The diagonal terms $\sigma_{ii} = cov(X_i, X_i)$ are the variances $\sigma_i^2 > 0$ for each of the individual X_i variables.

In the two-dimensional case, $d = 2$, with $\underline{x} = (x_1, x_2)$ we can write the covariance matrix in a form such that $\sigma_{12} = \sigma_{21} = r\sigma_1\sigma_2$, where σ_1 and σ_2 are the standard deviations of X_1 and X_2 respectively and where r is the well-known linear correlation coefficient with $-1 < r < 1$.

Prove that if X_1 and X_2 are jointly Gaussian that they are then also each marginally Gaussian (its sufficient to show for $P(X_1)$ given $P(X_1, X_2)$, since we can show the same for X_2 by the same argument). Hint: you can do this proof by defining $P(X_1)$ via the law of total probability and then integrating out X_2 from the joint density. It will also be helpful to know that the inverse covariance matrix Σ^{-1}, for a two-dimensional Gaussian, can be written as:

$$\Sigma^{-1} = \frac{1}{1-r^2} \begin{pmatrix} \frac{1}{\sigma_1^2} & \frac{-r}{\sigma_1\sigma_2} \\ \frac{-r}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}$$

(A more complicated proof (not required here) is needed for the general statement that any subset of d variables is jointly Gaussian if the full set of d variables is jointly Gaussian. The same holds for conditional densities, i.e., any subset is jointly Gaussian conditioned on the values of any other subset, if the union of the two sets are jointly Gaussian).