CS 274A Homework 1

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2023

Due: 11:59pm Tuesday January 17th, submit via Gradescope

Instructions and Guidelines for Homeworks

- Please answer all of the questions and submit your solutions to Gradescope (either hand-written or typed are fine as long as the writing is legible).
- All problems are worth equal points unless otherwise stated. All homeworks will get equal weight in computation of the final grade for the class (with lowest-scoring homework being dropped).
- The homeworks are intended to help you better understand the concepts we discuss in class. It is important that you solve the problems yourself to help you 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.
- In problems that ask you to derive or prove a result 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 can't solve a problem, you can discuss the high-level concepts *verbally* with another student (e.g., what concepts from the lectures or notes or text are relevant to a problem). However, you should not discuss any of the details of a solution with another student. In particular note that 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, etc. The work you hand in should be your own original work.
- If you need to you can look up standard results/definition/identities from textbooks, class notes, textbooks, other reference material (e.g., from the Web). If you base any part of your solution on material that we did not discuss in class, or is not in the class notes, or is not a standard known result, then you may want to rovide a reference in terms of where the result is from, e.g., "based on material in Section 2.2 in ....." or a URL.

Recommended Reading for Homework 1

- Note Sets 1 and 2 from the class Web page, for a review of basic concepts in probability, conditional independence, Gaussian models, etc.
- Chapter 6.1 to 6.5 in *Mathematics for Machine Learning* (MML) is recommended for additional details beyond what is covered in the Note Sets.

Problem 1: Properties of the 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, where $p(x)$ is the PDF.

1. Derive an expression for the expected value $E[X]$.
2. Derive an expression for the variance $var(X)$, where $var(X) = \sigma_x^2 = E[(X - \mu_x)^2]$.

Problem 2: Expectations/Variance with Two Random Variables

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 $\sigma_x^2 = var(X) = E[(X - \mu_x)^2] = \int p(x)(x - \mu_x)^2 dx$. In the questions below a and b are scalar constants (i.e., not random variables).

1. Prove that $var(X) = E[X^2] - \big(E[X]\big)^2$.

   In the next two questions let X and Y be two real-valued random variables, each one-dimensional (i.e., scalar-valued). In the equations below the expectation on the left is with respect to the joint density $p(x, y)$ and the expectations on the right are with respect to $p(x)$ and $p(y)$ respectively. Be sure to be clear in each line of your derivation and don't skip steps.

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

Problem 3: High-dimensional Data

Answer the following problems:

1. Consider a d-dimensional discrete random (vector) variable $X = (X_1, X_2, \ldots, X_d)$, where each component random variable $X_i, 1 \leq i \leq d$ can take one of K values. Let $P(\underline{x})$ be a probability distribution for X where $\underline{x} = (x_1, \ldots, x_d)$ represents a d-dimensional vector of possible values of X.

   Assume we have a data set consisting of N random (independent) samples from $P(\underline{x})$. This dataset can be represented as counts in a d-dimensional table consisting of K^d cells, with one cell for every possible combination of $x_1, \ldots, x_d$ values. (In practice, for large values of K and d, we would likely use a sparse matrix/array representation to list the non-zero counts, rather than storing everything with a full array).

   Let j be an index over the K^d cells and let the probability of a particular cell j be $P_j = \alpha_j / K^d, \alpha_j \geq 0$ and $\sum_j \alpha_j = K^d$. For example, if all the α_j's are equal to 1 we get a uniform distribution over the K^d outcomes. How far the value of α_j is from 1 provides an indication of how much more (or less) likely outcome j is relative to a uniform distribution.

   (a) For a particular cell j, and with N independent random samples from $P(\underline{x})$, derive an expression involving α_j, K, d, N for the probability that at least 1 of the N samples lies in cell j.

   (b) Let $\beta_j = \frac{N\alpha_j}{K^d}$. Prove that if $\beta_j \ll 1$ then the probability that cell j has no samples will be approximately equal to $1 - \beta_j$. (Hint: a Taylor series approximation using the result from part 1 would be one possible approach here).

   (c) Comment briefly (1 or 2 sentences) on the implications of this result for estimation of distributions as K and/or d grow. For example, for modeling the probabilities of word-level trigrams in a language model we would have $d = 3$ and K could be on the order of 10^5 words.

2. Consider a d-dimensional hypercube whose edges are of length $2r$. Now consider a d-dimensional hypersphere which has radius r and is inscribed within the hypercube. The hypercube and hypersphere have their centers in the same location.

   (a) Derive a general expression for the ratio of the volume of the hypersphere to the volume of the hypercube. (You don't need to derive the equation for the volume of a hypersphere in d dimensions, you can just look it up).

   (b) Compute numerically (e.g., using a calculator or computer) the value of this ratio for $d = 1, 2, \ldots, 10$. You won't need to know the value of r to do this.

   (c) Comment briefly on what the numbers in the table tell you about where "data lives" (at least under a uniform distribution) in high-dimensional spaces.

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 each X_i has mean μ and variance σ^2. (Note that the density $p(x)$ could be any probability density function, it need not be Gaussian).

1. State precisely the central limit theorem as it applies to $X_1, \ldots, X_n$ (if you don't know what central limit theorem is 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 $U(a, b)$ with $a = 0, b = 1$. Simulate 1000 values of Y (using any language such as Python, R, Matlab, C, etc) for each of the following values of n: $n = 10^2, 10^3, 10^4, 10^5$. You should end up with 4 sets of Y values, each with 1000 values. Generate histogram plots of the 4 results for each value of n (this will produce 4 histograms). Please make sure that all 4 histograms are plotted on a single page (makes it easier for grading). Use $\sqrt{1000} \approx 30$ bins for each histogram.
3. Based on visual inspection of the histograms, comment on the qualitative nature (e.g., shape, nature of the distribution) for 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: Logistic Function

Let X be a d-dimensional real-valued (vector) random variable taking values $\underline{x}$ and let C be a binary random variable taking values 1 or 2. Say we would like to model the conditional probability $P(C = 1|\underline{x})$ as a function of $\underline{x}$. One well-known approach is to assume that $P(C = 1|\underline{x})$ is defined as a logistic function (this is the basis of the logistic regression classifier):

$$P(C = 1|\underline{x}) \;=\; \frac{1}{1 + exp(-\alpha_0 - \alpha^T \underline{x})}$$

where α_0 is a real-valued scalar and α^T is the transpose of a d-dimensional vector ($d \times 1$) of real-valued coefficients $\alpha_1, \ldots, \alpha_d$. In machine learning C is typically referred to as the "class" variable: its the variable we want to predict given $\underline{x}$.

1. Prove that the definition of the logistic function above implies that the log-odds $\log \frac{P(C=1|\underline{x})}{P(C=2|\underline{x})}$ is an affine function of $\underline{x}$.
2. Say we know that $P(\underline{x}|C = 1) = N(\underline{\mu}_1, \Sigma)$ and $P(\underline{x}|C = 2) = N(\underline{\mu}_2, \Sigma)$ (i.e., we know that the densities for each class are multivariate Gaussian), where $\underline{\mu}_1$ and $\underline{\mu}_2$ are the d-dimensional means for each class and Σ is a common covariance matrix. Prove that, under these assumptions, $P(C = 1|\underline{x})$ is in the form of a logistic function. (Hint: you may find the algebra to be easier if you utilize the result from part 1).

Problem 6: Finite 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 of a mixture model is to define a density (or distribution) $p(x)$ that is a weighted mixture of K component probability density functions $p_k(x|Z = k)$, where the weights are non-negative and sum to 1, i.e.,

$$p(x) \;=\; \sum_{k=1}^{K} p_k(x|Z = k)P(Z = k)$$

where

- Z is a discrete indicator random variable taking values from 1 to K, indicating which of the K mixture components generated data point x.
- The mixture weights $\alpha_k = P(Z = k)$ are the marginal probabilities of data point x being generated by component k, with $\sum_{k=1}^{K} \alpha_k = 1, \;\; 0 \leq \alpha_k \leq 1$.
- for each value of k, $p_k(x|Z = k)$ is itself a probability density function with its own parameters θ_k. For example, if a component is Gaussian then $\theta_k = \{\mu_k, \sigma_k^2\}$.

The full set of parameters for a mixture model consists of both (a) the K weights, and (b) the K sets of component parameters θ_k for each of the K mixture components. (Note that the "finite" in finite mixture models comes from the fact that K is finite. There are also infinite mixture models where K is unbounded, but we will not consider those here).

1. Given the definition above for a finite mixture model, 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. Derive general expressions for the (a) mean μ of $p(x)$, and (b) 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 final expression for each of the mean and the variance.
3. If we now assume that $K = 2$ where both components are Gaussian densities and $\mu_1 = 0$ and $\mu_2 = 5$, plot the density of $p(x)$ (e.g., with isocontours) as a function of x for each of the following cases:
   (a) $\alpha_1 = 0.5, \sigma_1 = 3, \sigma_2 = 3$
   (b) $\alpha_1 = 0.5, \sigma_1 = 2, \sigma_2 = 2$
   (c) $\alpha_1 = 0.5, \sigma_1 = 2, \sigma_2 = 1$
   (d) $\alpha_1 = 0.1, \sigma_1 = 2, \sigma_2 = 2$

Let x range from -5 to 10 in your plots. Its fine to write some code to generate the plots (in fact this is preferred since generating these plots accurately by hand would be tricky to do).

Problem 7: Conditionally Independent Experts

Let Y be a binary class variable taking values $y \in \{0, 1\}$. Let X_i be a feature taking feature values x_i (potentially vector-valued) with $i = 1, \ldots, M$. Associated with each of the M features X_i is an "expert" that given a feature value x_i produces a prediction $P(y = 1|x_i)$. Individual experts could for example correspond to machine learning models or humans.

Now consider a decision-maker that wishes to compute $P(y = 1|x_1, \ldots, x_M)$ but that doesn't know the x_i values directly. Instead the decision-maker is only given the expert predictions $P(y = 1|x_i), i = 1, \ldots, M$. In addition the decision-maker knows the marginal probability $P(y = 1)$. This could model a situation for example where there are multiple different pieces of medical information x_i relevant to predicting disease status Y for a patient, but the information x_i can't be provided to the decision-maker for privacy reasons—however, all the individual expert predictions $P(y = 1|x_i)$ can be provided.

We will analyze the case where the decision maker assumes that the X_i variables are conditionally independent given Y. Also say that the decision maker assumes that each expert i is providing the true probability $P(y = 1|x_i)$ rather than an estimate of this probability.

1. Derive an equation that shows how the decision-maker can compute the odds, $\frac{P(y=1|x_1,\ldots,x_M)}{P(y=0|x_1,\ldots,x_M)}$ based on the information provided above.

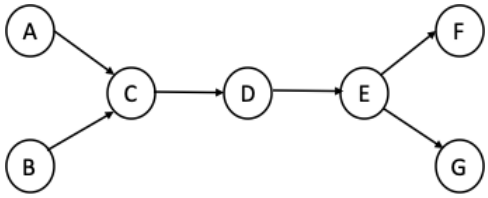
2. Show that the log-odds in part (1) can be written as a linear function of the log-odds from the individual experts, plus an additional term that depends on the marginal probability of Y.

3. Interpret your result for the case $M = 1$ and explain in words what is qualitatively different to the case for $M > 1$.

4. There are multiple other ways the decision-maker could combine information from the M experts (such as averaging the predictions or using voting). For example, say the decision-maker were to threshold the individual probabilities of each expert, i.e., $z_i = 1$ if $P(y = 1|x_i) \geq 0.5$ and $z_i = 0$ otherwise (so the z_i in effect correspond to the votes of individual experts), and the decision maker then computes $P(y = 1|x_1, \ldots, x_M) \approx \frac{1}{M} \sum_i z_i$, i.e., takes the average of the votes. Provide an example for $M = 3$ that shows that illustrates clearly why this strategy of combining information (given the assumptions above) is suboptimal compared to the solution you derived in part 1 above.

Problem 8: Inference in Graphical Models

Consider the directed graphical model in the figure below. All variables are discrete and all take $K \geq 2$ values.



Answer the following questions:

1. Write an equation for the joint probability, $P(a, b, c, d, e, f, g)$ that represents the conditional independence relations in this graphical model.

2. Precisely how many parameters are required to specify this graphical model? Express your answer as a function of K. Take into account the fact that distributions sum to 1. "Parameter" here means a probability in a probability table. Express your final answer in the form of a polynomial in K.

3. Consider the probability $P(d^*|a^*)$, where d^* and a^* are specific values of D and A respectively. Describe (step by step, for all steps) the most efficient way to compute this conditional probability, starting from the marginal and conditional probability tables that are specified in the graphical model. You can interpret "most efficient" to mean a method that requires the least number of summations as a function of K, e.g., $\sum_{x,y} P(x, y, z^*)$ would involve a sum over the values x and y with the number of summations being of order $O(K^2)$.

4. Now consider the probability $P(d^*|a^*, g^*)$, where g^* is a specific value of G. As in the last question, describe the most efficient way to compute this conditional probability. The most straightforward way to do this is to first compute the joint probability $P(d, g^*|a^*)$ for each value of d; and to then compute the conditional probability of interest, $P(d^*|a^*, g^*)$, from the joint probability via Bayes rule.