CS 274A Homework 1

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2026

Due: Noon Tuesday January 13th, 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 (10 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).
- Please be sure to read the academic integrity policy on the course Web page, including in particular the policies on restrictions for use of generative AI for homework assignments.
- 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 may 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, do not look at (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 submit 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 should provide a reference in terms of where the result is from, e.g., "based on material in Section 2.2 in ....." or a URL (e.g., Wikipedia).
- Please read each problem carefully. If you believe there is a typo, or some information is missing, or the problem is unclear, please post a question on the Ed discussion board.

Suggested 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.
- Appendices A, B, and C in *Understanding Deep Learning* (MIT Press, Simon Prince) are well worth reading as general background on probability, linear algebra, and other relevant topics. These appendices may be especially helpful if you are a little out of practice with these topics. This text is available online at `http://udlbook.com`.
- Chapter 6.1 to 6.5 and Chapter 8.5 in *Mathematics for Machine Learning* (MML) go beyond what is covered in the Note Sets and may be useful for some of the problems.

Problem 1: Properties of 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 purchases that a user makes during a visit to a particular Web site. A well-known simple model for this type of "count data" is the Poisson distribution, with

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

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 2: 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 or remember what the 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, each with 1000 simulated values for Y. Generate histogram plots of the 4 sets (one histogram for each value of n, producing 4 histograms). Please plot all 4 histograms on a single page (makes it easier for grading). Use $30 \approx \sqrt{1000}$ bins for each histogram. No need to submit your code.

3. From the definition of the properties of a uniform random variable (e.g., its mean and variance as a function of a and b—you can look up the definitions if you don't know them) and the definition of the central limit theorem, state what the mean and variance of Y should be as a function of n.

4. Evaluate how well your empirically simulated distributions from Part 2 match what the theory predicts from Part 3, e.g., show 1 or 2 tables, with different values of n for the rows, and where (in the columns) you compare the mean and variance of the simulated data with the values that theory predicts.

Problem 3: 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 α_k are non-negative and sum to 1, i.e.,

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

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 k 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" aspect of 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 a valid probability 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, also provide an intuitive interpretation in words of your interpretation of the equations you derived for each of the mean and the variance.

3. Now assume that $K = 2$ and that both components are Gaussian densities with $\mu_1 = 0$ and $\mu_2 = 5$. Plot $p(x)$ 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). If possible please put all 4 plots on a single page in your submission, e.g., using a 2×2 grid (this will make it easier for grading).

   No need to submit your code.

Problem 4: 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 M 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 M^d cells, with one cell for every possible combination of $x_1, \ldots, x_d$ values. (In practice, for large values of M 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 M^d cells and let the probability of any particular cell j be $P_j = \alpha_j / M^d, \alpha_j \geq 0$ and $\sum_j \alpha_j = M^d$. If all the α_j's are equal to 1 we get a uniform distribution over all if the possible M^d outcomes. How far α_j is from a value of 1 provides an indication of how much more (or less) likely outcome j is relative to a uniform distribution.

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

   (b) Let $\beta_j = \frac{N\alpha_j}{M^d}$. 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 M and/or d grow. For example, for modeling the probabilities of word-level trigrams in a language model we would have $d = 3$ and we could have on the order of $M = 10^5$ 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 5: Logistic Function

Let X be a d-dimensional real-valued (vector) random variable taking values $\underline{x}$ and let Y be a binary random variable taking values 0 or 1. Say we would like to model the conditional probability $P(Y = 1|\underline{x})$ as a function of $\underline{x}$. One well-known approach is to assume that $P(Y = 1|\underline{x})$ is defined as a logistic function (this is the basis of the logistic regression classifier in machine learning and statistics):

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

where α_0 is a real-valued scalar and $\underline{\alpha}^T$ is the transpose of a $d \times 1$ vector of real-valued coefficients $\alpha_1, \ldots, \alpha_d$. In this setup Y is typically referred to as the "class": 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 $\log \frac{P(Y=1|\underline{x})}{P(Y=0|\underline{x})}$ is an affine function of $\underline{x}$, i.e., that the log-odds can be written as $\underline{a}^T \underline{x} + b$ for some vector $\underline{a}$ and scalar b. State clearly what $\underline{a}$ and b are in your solution.

2. Say we know that $P(\underline{x}|Y = 1) = N(\underline{\mu}_1, \Sigma)$ and $P(\underline{x}|Y = 0) = N(\underline{\mu}_0, \Sigma)$ (i.e., we know that the densities for each class are multivariate Gaussian), where $\underline{\mu}_1$ and $\underline{\mu}_0$ are the d-dimensional means for each class and Σ is a common covariance matrix. Prove that, under these assumptions, $P(Y = 1|\underline{x})$ is in the form of a logistic function. See Section 4 in Note Set 2 for information about the multivariate Gaussian density.

   - Hint 1: one way to prove this is to make use of the result from part 1 of this problem).
   - Hint 2: it may be helpful in your solution to use the fact that $\underline{\mu}^T \Sigma^{-1} \underline{x} = \underline{x}^T \Sigma^{-1} \underline{\mu} = c$, where c is some scalar value. This is true as long as Σ^{-1} is symmetric (which it is, given that by definition a covariance matrix Σ is symmetric, and given the fact that the inverse of a symmetric matrix is also symmetric).

Problem 6: (Naive Bayes Classification Model)

The naive Bayes model is a very simple classification model. We have a class variable Y taking K possible values $\{1, \ldots, K\}$ and we wish to predict Y as a function of d features $X_1, \ldots, X_d$, where Y and X's are all random variables. The key assumption of a naive Bayes model is that each feature X_j is assumed to be conditionally independent of all the other features given Y.

For example, Y 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.

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

1. Write down an expression for the joint distribution $P(Y, X_1, \ldots, X_d)$ for the naive Bayes classification model and draw a picture of the graphical model for the case of $d = 3$. (No need to use plate diagrams or data here, just show how the Y and X variable related given the information provided).

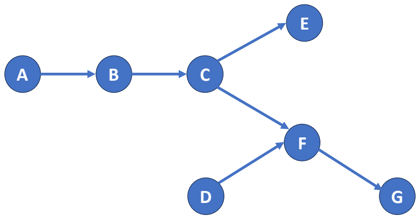
2. Specify exactly how many parameters are needed for this model in the general case, as a function of M, K, and d. A *parameter* in this context is any probability value or conditional probability value that is needed to specify the model.

3. Now say that each of the d features $X_1, \ldots, X_d$ are real-valued and that we assume that the conditional density for each feature given the class, $p(X_j|y = k)$ 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.

4. Say we are given a naive Bayes model where the parameters are known. Say we observe a set of values $x_1, \ldots, x_d$ for the features, but the value class variable Y is unknown. Using the structure of the model, show clearly and precisely (with equations) how one can use the model to compute the conditional distribution $P(y = k|x_1, \ldots, x_d), 1 \leq m \leq M$. Note that the solution to this problem is general in the sense that it doesn't depend on whether the x's are discrete or real-valued, or whether they are Gaussian or some other distribution if they are real-valued.

Problem 7: Inference in Graphical Models

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



Answer the following questions:

1. Write an equation for $P(a, b, c, d, e, f, g)$ that factorizes the joint probability in a manner that reflects the conditional independence assumptions in this graphical model.
2. List all of the different conditional and marginal probability distributions in this graphical model and define precisely how many parameters in total are required to specify the model. Take into account the fact that all distributions must sum to 1. "Parameter" here means a marginal or conditional probability in a probability table. Express your final answer in the form of a polynomial in M.
3. Consider computing the probability $P(g^*|a^*)$, where g^* and a^* are some specific values of G and A respectively. Describe (step by step, for all steps) the most efficient way to compute this conditional probability, using only 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 lowest time complexity in M as a function of the number of summations. For example, to compute $P(c|a^*) = \sum_b P(c|b)p(b|a^*)$ requires $O(M)$ summations for each value of c, and thus requires $O(M)$ of these summations to compute the distribution $P(c|a^*)$ for all M values of C, for a time complexity of $O(M^2)$ in general.
4. Now consider computing the probability $P(e^*|g^*)$, where e^* and g^* are some specific values of E and G. As in the last question, describe the most efficient way to compute this conditional probability using the lowest complexity in terms of M. One way to do this is to first compute the joint probability $P(e^*, g)$ for each possible value of g; and to then compute the conditional probability of interest, $P(e^*|g^*)$ via Bayes rule.

In the last 2 problems above you will want to start by using the law of total probability (LTP) and introducing variables that lie on the path between the nodes in the expression you are trying to compute. For example, in part 3 it may be helpful to first write $P(g^*|a^*)$ as a sum over F's values using LTP, and then proceed by seeing what needs to be computed next, and so on.