CS 274A Homework 1

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2025

Due: 9am Wednesday January 15th, 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).
- 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.

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.
- 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: 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 2: Properties of the Uniform Density

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

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

for $a \leq x \leq b$, and otherwise $p(x) = 0$, where $p(x)$ is the probability density function for X.

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

Your expressions for both the mean and variance should be functions of a and b.

Problem 3: Modeling Sequence Lengths in Language Models

Consider a large language model that can generate sequences of tokens (e.g, words, numbers, punctuation) in a stochastic manner as follows. The model generates the first token of the sequence by sampling from a probability distribution over all possible initial tokens. The model then samples a biased coin with probability θ of getting "heads" and probability $1 - \theta$ of getting "tails," where θ is some fixed number and $0 < \theta < 1$. If the outcome is heads, then the language model generates an end-of-sequence (EOS) token and the sequence ends. If the outcome is tails then the language model generates a 2nd token, by sampling from a conditional distribution that is a function of the tokens that have occurred earlier in the sequence.

The process continues in this manner: first a coin with parameter θ is sampled, where the sequence ends with an EOS token if the coin is "heads", or if the coin is "tails" another token is drawn using a conditional probability distribution over tokens that depends on the history of tokens already generated up to that point[1].

Let X be the discrete random variable representing the length of a generated sequence, not including the EOS token. Given the description above, X can take values $k \in \{1, 2, 3, \ldots, \}$. Answer the following questions:

1. Write down the general formula for $P(X = k)$ as a function of θ and $k = \{1, 2, 3, \ldots, \}$. Explain clearly in words (sentence or two) why this is the correct formula. What well-known parametric distribution does this correspond to?
2. Prove that $\sum_{k=1}^{\infty} P(X = k) = 1$ (you may need to consult standard results in sums and series).
3. Derive an expression as a function of θ for the expected value of X, $\mu_x = E[X]$.
4. Derive an expression as a function of θ for the variance of X, where the variance is $\sigma_x^2 = E[(X - \mu_x)^2]$.

Note that you no longer are required to submit a solution for 3.4, you can get full points by just submitting 3.1 to 3.3

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 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.

[1] Real-world large language models operate in this sequential fashion, generating each word (or token) based on earlier words in the sequence, until the EOS token is sampled. Here, for simplicity, we have made θ be a fixed quantity that doesn't depend on the words that came before it; in real language models, the probability of an EOS token will depend on the tokens already generated, e.g., in real models if the most-recently generated token is a punctuation token then the probability of ending the sequence may be much higher than if it is a non-punctuation token.

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 any particular cell j be $P_j = \alpha_j / K^d, \alpha_j \geq 0$ and $\sum_j \alpha_j = K^d$. If all the α_j's are equal to 1 we get a uniform distribution over all if the possible K^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, 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 we could have on the order of $K = 10^5$ 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.

Note that you no longer are required to submit a solution for 4.2 you can get full points by just submitting 4.1

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 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: Prediction with Missing Data

Consider a problem where we are predicting a binary class variable Y taking values $y \in \{0, 1\}$. We build a model (e.g., learn it from data) that takes as input a d-dimensional real-valued feature vector $\underline{x}$ and produces as output $P(y = 1|\underline{x})$. The model could be a logistic regression model or a feedforward neural network or some other classification model.

At prediction time, when we make predictions on new feature vectors $\underline{x}$, assume that only some of the features are observed and the rest of the feature values are missing. We will further assume that for every feature vector $\underline{x}$ that we wish to make a prediction for, some subset of $d' < d$ of the components of $\underline{x}$ are deleted and missing, and the selection of which feature values are missing is entirely random (e.g., for every $\underline{x}$ each component of $\underline{x}$ might be deleted or not in some stochastic manner that is independent of other values or other deletions in $\underline{x}$).

Notationally, we are making predictions with feature vectors of the form $\underline{x} = (\underline{x}_o, \underline{x}_m)$ where $\underline{x}_o$ and $\underline{x}_m$ are the observed and missing components of $\underline{x}$ respectively. Assume that when the model is learned there is no information available on what components $\underline{x}$ will be missing or observed at prediction time (e.g., the d' missing components are randomly selected for every new prediction vector $\underline{x}$). Assume also that the model needs a full d-dimensional vector of feature values to make a prediction $P(y = 1|\underline{x})$.

1. Derive an equation for the optimal method for making predictions $P(y = 1|\underline{x}_o)$ for any feature vector $\underline{x} = (\underline{x}_o, \underline{x}_m)$. (Hint: this derivation involves the law of total probability and is quite short).

2. Explain in words how this optimal method is different to substituting in specific values for the missing feature values (e.g., the mean values for these features).

3. Identify (in words) two significant practical issues with using the optimal method in real-world problems (e.g., write 1 or 2 sentences for each issue you identify)

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.

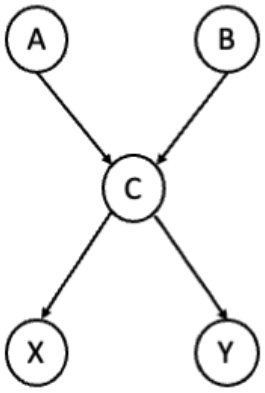
2. Show that the log-odds in part (1) can be written as a linear function of the log-odds from the individual experts, plus an additional term that depends on the marginal probability of Y.

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. The variables take values a, b, c, x, y respectively. A, B, C are discrete random variables with A and B each taking M values and C is a binary random variable. X and Y are real-valued random variables that are conditionally Gaussian, i.e., for each value c of C, there is a conditional Gaussian density for X with its own mean and variance (and same for Y).



Answer the following questions:

1. Write an equation for the joint probability, $p(a, b, c, x, y)$ that corresponds to this graphical model.

2. How many parameters precisely are required to specify this graphical model? Express your answer as a function of M. Note that the term "parameters" includes conditional and marginal probabilities need to fully specify the graphical model as well as any means, variances, etc.

3. Say we want to compute $p(Y = y^*|A = a^*)$, i.e., the value of the probability density at y^* conditioned on a^*, where y^* and a^* are two specific values of Y and A respectively. Show precisely how to perform this computation using the information provided by the graphical model, i.e., the probability tables and densities that are assumed to be provided with the graphical model.

   Clearly explain (with equations) how all intermediate steps in the computation would be performed. Note that you don't need to show the equation for a Gaussian density in your solution, you can just refer to it as a density function p that can be evaluated at $Y = y^*$.

4. Say we now have the same graphical model as indicated in the diagram above, but X and Y are each d-dimensional vectors instead of being scalars: so now $p(X|C = c)$ and $p(Y|C = c)$ are each multidimensional Gaussian densities. How many parameters precisely are required to specify this new graphical model? Express your answer as a function of M and d. (It may be helpful to review Section 4 on multivariate Gaussian models in Note Set 2 before working on this problem).