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 equal 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.
- 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 it verbally with another student. However, please 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, and so forth. The work you hand in should be your own original work.
- While it is strongly recommended that you try to solve the problem yourself, without looking up solutions elsewhere, you are allowed to use reference materials in your solutions, such as class notes, textbooks, other reference material (e.g., from the Web). **However, you must clearly state at the start of the problem that you are doing this and cite your source** (unless its already in the class notes, or discussed in class, or a standard known result such as a closed-form expression for an infinite sum). Note that you may not get full points if your solution relies on other material.
- 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. And please feel free to submit the .tex (as well as the .pdf) file for your solutions —it may be helpful to us when we compile solutions (we sometimes find that students come up with more elegant solutions than the ones we already have!).
Recommended Reading for Homework 2:

- Note Set 3 on the class Web page should be helpful.
- Section 8.3 in Chapter 8 in *Mathematics for Machine Learning* (MML)
- Section 4.1 and 4.2 in the new online edition of Kevin Murphy’s book

Note: In most of the problems below, unless otherwise stated, the observations in a data set $D$ are assumed to be conditionally independent given parameters $\theta$.

**Problem 1: Maximum Likelihood for the Multinomial Model**

Consider building a probabilistic model for how often words occur in English. Let $W$ be a random variable, taking values $w \in \{w_1, \ldots, w_V\}$, where $V$ is the number of words in the vocabulary. In practice $V$ can be very large, e.g., $V = 100,000$ is not unusual (there are more words than this in English, but many rare words are not modeled).

The *multinomial model* for $W$ is essentially the same as the binomial model for tossing coins, where we have independent trials, but instead of two possible outcomes there are now $V$ possible outcomes for each “trial”. The parameters of the multinomial are $\theta = \{\theta_1, \ldots, \theta_V\}$, where $\theta_k = P(W = w_k)$, and where $\sum_{k=1}^{V} \theta_k = 1$. Denote the observed data as $D = \{r_1, \ldots, r_V\}$, where $r_k$ is the number of times word $k$ occurred in the data (these are known as the sufficient statistics for this model).

1. Define the likelihood function for this problem
2. Derive the maximum likelihood estimates for the $\theta_k$’s for this model.

**Problem 2: Maximum Likelihood for the Poisson Model**

Consider a data set $D = \{x_1, \ldots, x_n\}, x_i \in \{0, 1, 2, \ldots\}$. Assume a Poisson model for the data, defined as

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!},$$

with parameter $\lambda > 0$ and where $k \in \{0, 1, 2, 3, \ldots\}$.

1. Define the likelihood function for this problem
2. Derive the maximum likelihood estimate for $\lambda$
Problem 3: Method of Moments and the Uniform Model

The method of moments is an alternative parameter estimation to maximum likelihood—theoretically its’ properties are not in general as good as maximum likelihood, but it can nonetheless be useful for some problems (e.g., where the likelihood function is not easy to optimize but the method of moments is easier to work with).

The method works as follows: Given a probability model (e.g., a Gaussian, a uniform, etc) with \( K \) parameters we write down \( K \) equations that express the first \( K \) moments as functions of \( K \) parameters. The moments are defined as \( E[X^k], k = 1, \ldots, K \). Given a data set with \( N \) data points \( x_1, \ldots, x_N \), we then plug in the empirical estimates of these moments (from the data, e.g., the average value of \( x_i, \) of \( x_i^2, \) etc) into these equations and get \( K \) equations with \( K \) unknown parameters. We can think of this method as “moment matching,” i.e., it is trying to find parameters such as the moments of the model (with its estimated parameters) match the empirical moments in the observed data.

Let \( X \) be uniformly distributed with lower limit \( a \) and upper limit \( b \), where \( b > a \), i.e.,

\[
p(x) = \frac{1}{b - a}
\]

for \( a \leq x \leq b \) and \( p(x) = 0 \) otherwise. Assume we have a data set \( D \) consisting of \( N \) scalar measurements \( x_i, 1 \leq i \leq N \).

1. Derive estimators for \( a \) and \( b \) using the method of moments. Since there are \( K = 2 \) unknown parameters this means that you will need two equations, involving the first and second moment.
2. Now derive the maximum likelihood estimators for \( a \) and \( b \) (think carefully about how to do this).
3. Write 2 or 3 sentences comparing the properties of the maximum likelihood solutions with the method of moment solutions. You can use the following simple data set \( D = \{12, 4, 5, 10, 6, 5, 8, 10\} \) to provide some intuition for your answer.

Problem 4: Maximum Likelihood for Finite Mixture Models

Consider a finite mixture model \( f(x) \), where \( x \) is real-valued and univariate, and where the \( K \) component densities \( f_k(x|Z = k) \) are Gaussian, and where the weights are non-negative and sum to 1, i.e.,

\[
f(x) = \sum_{k=1}^{K} f_k(x|Z = k)P(Z = k)
\]

(see Homework 1 for additional explanation about mixtures).

Given an IID data set \( D = \{x_1, \ldots, x_N\} \) describe a solution that maximizes the log-likelihood for this problem. Note that this will not necessarily be derived from calculus but instead by thinking about the nature of a mixture model and the definition of maximum likelihood. Your solution should only take 2 or 3 lines to describe. Note: please think about this problem first, don’t look it up online. If you end up not being able to figure it out and look it up online then cite your source.
Problem 5: Maximum Likelihood with Measurement Variance per Point

Consider a data set $D$ consisting of $N$ scalar measurements $x_i, 1 \leq i \leq N$, where each measurement is taken from a different Gaussian, such that each Gaussian has the same mean $\mu$, and each Gaussian has a different variance $\sigma_i^2, 1 \leq i \leq N$, where these $N$ variances are known. For example, this might be an astronomy problem where we are trying to estimate the brightness $\mu$ of a star and our data consists of measurements $x_i$ taken at different locations $i$ on the planet where noise $\sigma_i^2$ per datapoint varies due to the local atmosphere (in a known way) with location $i$.

- Define the log-likelihood for this problem.
- Derive the maximum likelihood estimator for $\mu$.
- Comment on the functional form of your solution: for example, can you interpret the result in the form of a weighted estimate? what are the weights?

Problem 6: Maximum Likelihood and Naive-Bayes/Gaussian Model

In the problems below let $X = (X_1, \ldots, X_d)$ be a real-valued $d$-dimensional vector random variable taking values $\mathbf{x} = (x_1, \ldots, x_d)$.

1. Let $X$ have a Gaussian (Normal) multivariate joint density $p(x) = N(\mu, \Sigma)$. Prove that $\Sigma$ is diagonal if and only if the $X_i$’s are independent.

2. In this part of the problem let $C$ be a discrete-valued random variable taking values $c \in \{1, \ldots, K\}, K \geq 2$. Consider a model parametrized as $P(X|C)P(C)$, which is in effect a mixture model with $K$ components, and assume that each of the $K$ components are Gaussian with means $\mu_c$ and covariances $\Sigma_c$. Further assume that this model is a naive-Bayes model where each variable $X_i$ is conditionally independent of all of the other variables in $X$, conditioned on $C$.

   (a) What is the total number of parameters in this model? (Assume here that no assumptions are made here about parameters being tied in some way or being dependent on each other).

   (b) Given $N$ IID observations, $(x_i, c_i), i = 1, \ldots, N$ from this model, write the likelihood in a compact way that reflects the various independence assumptions in our model. You can use $j = 1, \ldots, d$ to index across dimensions and $i = 1, \ldots, N$ to index across data points, e.g., $x_{ij}$ is the value of the $j$th component of the $i$th observed vector $\mathbf{x}_i$. Feel free to use expressions such as $\prod_{i:c_i=k}$ to indicate products over terms where the class label is $k$.

   (c) Derive maximum likelihood estimates for all of the parameters in the model. (Note that if you look at the structure of the likelihood you should be able to simplify this considerably and your answer need only take a few lines). Feel free to use notation such as $n_k$ for the number of data points in class $k$. And you can refer to class notes or earlier solutions in your answer.
Problem 7: Maximum Likelihood: Variable-Length Sequences

Consider a Markov chain with 3 states and the following transition matrix, where the rows indicate the transition probabilities out of states 1, 2, and 3 (top to bottom respectively):

\[
\begin{pmatrix}
\alpha & \beta & \gamma \\
\beta & \alpha & \gamma \\
0 & 0 & 1 \\
\end{pmatrix}
\]

Let \( \theta = \{\alpha, \beta, \gamma\} \) be the 3 unknown parameters. States 1 and 2 each have self-transition probabilities of \( \alpha \); and each has transition probability \( \beta \) of transitioning to the other.

State 3 is what is known as an “absorbing state.” In this model, once the Markov chain arrives at state 3 it halts and produces no more states, i.e., if we generate data from this model we will have finite length sequences where the last state is always state 3 (and the last state is the only occurrence of state 3 in a sequence), e.g., a possible sequence could be \( s = [1121222113] \).

Consider an observed data set \( D \) consisting of \( L \) sequences \( \{s_1, \ldots, s_L\} \) where the sequences can be of different lengths. Let sequence \( s_i, i = 1, \ldots, L \) have \( n_i + 1 \) symbols (and thus, has \( n_i \) transitions). You can assume that the initial state distribution is \( \pi = [0.5, 0.5, 0] \), i.e., a sequence has a 50% chance of starting in either state 1 or 2 and 0% chance of starting in state 3. Each sequence \( s_i \) is conditionally independent of all other sequences given the parameters \( \theta \).

In the problems below feel free to use additional variable definitions such as \( N = \sum_{i=1}^{L} n_i \) to simplify notation.

1. Clearly define the likelihood \( L(\theta) \) for this problem.

2. Derive maximum likelihood estimates for each of the 3 unknown parameters.