CS 274A Homework 2

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2009

Due Date: Thursday January 29th, in class

All problems are worth 10 points unless otherwise stated.

Required Reading

Class Note Set 3 (available on the class Web page).

Comments on Homeworks

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 work through the problems yourself. If you can't solve a problem, you can discuss it with another person, but the basic rule is that you are can only discuss the problem verbally. You cannot exchange or view written material related to the homeworks (e.g., notes on solutions, written-up solutions, MATLAB code, etc) with anyone else. Homeworks will usually consist of a mix of review problems related to what we are discussing in class and some programming assignments in MATLAB so that you can actually try out some of the ideas we discuss in class with actual data and learning algorithms.

Note: the homework problems are very important to help you better learn and reinforce the material we discuss in class. If you don't do the homeworks you will likely have difficulty in the exams later in the quarter.

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, no "hand-waving" allowed!).

Please answer each of the following questions and submit your written solutions (either hand-written or typed are fine). 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.

Hand in a hardcopy of your solutions in class on the date due.

In the problems below assume that the data are IID (i.e., conditionally independent given the parameters in the model) unless stated otherwise.

Problem 1:

Let X be a Gaussian random variable with unknown parameters $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. Given a data set $D = \{x(1), \ldots, x(N)\}$, derive maximum likelihood estimators for θ_1 and θ_2.

Problem 2:

(20 points)

Assume you are working for a large search engine company and you wish to build a probabilistic model for the number of search results that a user clicks on. Let X be a random variable taking values $x = 0, 1, 2, \ldots ...$, where x represents the number of clicks. Assume that two different models are being considered for this problem:

- The geometric distribution, where
$$P(X = x) = (1 - p)p^x,$$
where p is the parameter of the model and $0 < p < 1$.
- The Poisson distribution, where
$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!},$$
where $\lambda > 0$ is the parameter of the model.

Assume that you have a data set $D = \{x(1), \ldots, x(N)\}$.

1. Write down expressions for the log-likelihood for both the geometric and Poisson distributions.
2. Derive the maximum likelihood estimate of p for the geometric distribution.
3. Derive the maximum likelihood estimate of λ for the Poisson distribution.
4. Let the data set D consist of the following table of values:

value	number of occurrences
0	15
1	30
2	25
3	16
4	7
5	3
6	2
7	0
8	1
9	1

Plot the log-likelihood for each of the geometric and Poisson models as a function of their respective parameters for this data set.

5. Using the maximum likelihood estimates of the parameters, on a single plot with the x-axis running from 0 to 10, plot the following:
   - The empirical probability (from the data) of each value
   - The probability distribution for the geometric model
   - The probability distribution for the Poisson model
6. Is the geometric or Poisson model a better fit to this data? Explain the reasoning behind your answer

You may want to use MATLAB to generate the results for the parts 4 and 5 if you wish (you do not need to hand in any code if you use MATLAB for this problem).

Problem 3:

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 the sufficient statistics for this model).

Derive the maximum likelihood estimates for each θ_k for this model.

Problem 4:

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.

1. Derive maximum likelihood estimators for a and b (think carefully about how to do this).
2. Write 2 or 3 sentences suggesting why these maximum likelihood estimates might not necessarily be the best estimates.
3. Suggest an alternative method for estimating the parameters.

Problem 5:

Consider two data sets D_1 and D_2, each consisting of scalar measurements $x_i, i = 1, \ldots, N_1$ for D_1 and $x_j, j = 1, \ldots, N_2$ for D_2. Assume that each set of measurements comes from a Gaussian distribution. The two Gaussian distributions share a common variance σ^2 and the mean μ_2 of the Gaussian for data set D_2 is known to be twice the value of the mean μ_1 for the first data set D_1. μ_1, μ_2 and σ^2 are all assumed unknown.

- Draw a graphical model in the form of a plate diagram for this problem.
- Define the log-likelihood for this problem.
- Derive the maximum likelihood estimators for the unknown parameters.

Problem 6:

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 μ, and each Gaussian has a different variance $\sigma_i^2, 1 \leq i \leq N$, where these N variances are known.

- Define the log-likelihood for this problem
- Derive the maximum likelihood estimator for μ
- 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 7:

Let $\mathbf{X}$ be a d-dimensional real-valued random variable taking values $\underline{x}$. Assume $\mathbf{X}$ has a multivariate Gaussian distribution with unknown mean $\underline{\mu}$ and a *diagonal* covariance matrix Σ.

- How many parameters are there for this model?
- Given a data set $D = \{\underline{x}(1), \ldots, \underline{x}(N)\}$ write down an expression for the log-likelihood that is expressed in terms of the parameters.
- Derive maximum likelihood estimates for the unknown parameters.

Problem 8:

Write a MATLAB function called `binlikelihood.m` that takes in arguments r and n and plots the likelihood function for the binomial model. Use your function to generate plots for (a) $r = 5$, $n = 10$, (b) $r = 50, n = 100$, and (c) $r = 5, n = 50$, over the range $[0, 1]$. Submit copies of both your plots and your code.

Problem 9:

Write a MATLAB function called `gausslogL.m` that takes in a data set in the form of an $n \times 1$ vector called `data`. The function should generate two plots:

1. the log-likelihood as a function of the mean μ, with σ fixed to its maximum likelihood value;
2. the log-likelihood as a function of σ, with μ fixed to its maximum likelihood value.

For μ, the range of the x-axis should correspond to the range of the data points supplied to the function. For σ, use a fixed range of 0.1 to 5 times the maximum likelihood estimate of σ.

Submit your code and your plots for the following two cases:

1. `data` is a vector of 10 draws from a Gaussian with mean $\mu = 10$ and $\sigma = 1$.
2. `data` is a vector of 1000 draws from the same distribution.