

Solutions to CS 274A Homework 1

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

January 27, 2009

Problem 1:

A , B , C and D , are four discrete-valued random variables, each taking K values. Let $P(a, b, c, d)$ be the probability that $A = a, B = b$, etc, where a, b , etc, are some particular values for each of the 4 variables.

1. How many numbers do we need in general to specify the full “joint” distribution for the 4 variables, $P(a, b, c, d)$, i.e., for all possible values of a, b, c, d ?
2. In general, if we know the “full joint distribution” on any set of random variables, we can calculate any conditional probability (e.g., $P(b|a)$) or “marginal” probability (e.g. $P(a)$) of interest. In this problem you will go through some exercises to convince yourself that this is true. Show how each of the following quantities can be calculated by starting from $P(a, b, c, d)$. Use the basic definitions of conditional probability and standard results such as Bayes’ rule and the law of total probability. You need to show clearly how every term in your final expression (equation) can be derived from $P(a, b, c, d)$. For any of the terms below you can reuse results you have derived for a different expression if appropriate.
 - (a) $P(a)$
 - (b) $P(a, b)$
 - (c) $P(c|a, b)$
 - (d) $P(d|b)$
 - (e) $P(b, c|a)$.
3. Say that we are now told that the 4 random variables are all independent of each other . Write down a general expression for the joint distribution $P(a, b, c, d)$. Precisely how many numbers are needed to specify this joint distribution under the independence assumption?

SOLUTION:

1. The full joint distribution table on the 4 variables must have a dimension for each variable. Since all variables take on K values, each of these dimensions will have K cells. Thus, the table will have size K^4 . Since the table must sum to 1, the table only needs to specify $K^4 - 1$ values, as the last value can be inferred from those.
2. For notational simplicity, I will denote summing over all possible values of the variable d as

$$\sum_d$$

in this problem. For example,

$$P(x) = \sum_y P(x, y)$$

will be used to denote summing a joint distribution over all of y ’s values in order to obtain the marginal distribution of x .

- (a) To calculate $P(a)$ from the joint distribution, we can simply sum over all values of the variables besides a , as shown below.

$$P(a) = \sum_b \sum_c \sum_d P(a, b, c, d)$$

- (b) Similarly to part (a), to calculate $P(a, b)$ from the joint distribution, we will sum over all possible values of c and d .

$$P(a, b) = \sum_c \sum_d P(a, b, c, d)$$

- (c) By the definition of conditional probability, we have

$$P(a, b, c) = P(c|a, b)P(a, b)$$

Thus, we know that

$$P(c|a, b) = \frac{P(a, b, c)}{P(a, b)}$$

We can compute both $P(a, b, c)$ and $P(a, b)$ by summing over the other variables in the joint distribution, as we've done in parts (a) and (b). This gives us the final value of $P(c|a, b)$.

$$P(c|a, b) = \frac{\sum_d P(a, b, c, d)}{\sum_c \sum_d P(a, b, c, d)}$$

- (d) By the definition of conditional probability, we know

$$P(b, d) = P(d|b)P(b)$$

or, equivalently

$$P(d|b) = \frac{P(b, d)}{P(b)}$$

We calculate $P(b, d)$ with the same method as part (b) and $P(b)$ with the same method as part (a), resulting in the following equation.

$$P(d|b) = \frac{\sum_a \sum_c P(a, b, c, d)}{\sum_a \sum_c \sum_d P(a, b, c, d)}$$

- (e) By the definition of conditional probability, we relate $P(b, c|a)$ to $P(a, b, c)$.

$$P(b, c|a) = \frac{P(a, b, c)}{P(a)}$$

Via the same methods as before, we can sum out over the unused variables and get our answer,

$$P(b, c|a) = \frac{\sum_d P(a, b, c, d)}{\sum_b \sum_c \sum_d P(a, b, c, d)}$$

3. If the four random variables are independent of each other, then we can use the definition of independence to factorize the joint distribution:

$$P(a, b, c, d) = P(a)P(b)P(c)P(d)$$

We now only need to specify each variable's marginal distribution to be able to calculate the joint distribution for any set of assignments to the four variables. This will require K entries for each of the four variables. However, the K -th value of each of these marginal tables can be inferred, since $\sum_a P(a) = 1$ by definition. Thus, the table needs $4 \cdot (K - 1)$ entries, with the K -th entry in each marginal distribution implied by the others.

Problem 2:

Note: the problem below is slightly different to that given in the homework, but the concepts are the same.

In a medical diagnosis problem a random variable A can take 2 values, 0 meaning a patient *does not have the disease* and 1 meaning *has the disease*. The random variable B represents the outcome of a test for the disease, where B can take values 0 (negative) and 1 (positive). Assume we know (based on past medical data or prior experience) the following:

$$P(B = 1|A = 0) = 0.01, \quad P(B = 1|A = 1) = 0.9,$$

from which we can deduce that $P(B = 0|A = 0) = 0.99$ and $P(B = 0|A = 1) = 0.1$. In words this tells us (for example) that for healthy people ($A = 0$) the test is negative 99% of the time—or equivalently there is a 1% chance of a false positive.

To use Bayes' rule to calculate $P(A|B)$ we will also need to know what $P(A)$ is—so suppose we know that $P(A = 1) = 0.001$ (i.e., only 1 in a thousand people on average have the disease).

1. Given this information use Bayes' rule to compute the probability that a person has the disease given (a) that the test outcome is negative (has value 0), and (b) given that the test outcome is positive (has value 1).
2. Repeat the calculation in the previous part, but now assume that the probability of a randomly selected person having the disease is much higher, $P(A = 1) = 0.2$. Comment on how $P(A|B)$ changed given this change in the marginal probability for A .

SOLUTION:

- First, let's write all of the information we know about the probabilities:

$$\begin{aligned} P(B = 1|A = 0) &= 0.01 & P(B = 1|A = 1) &= 0.9 \\ P(B = 0|A = 0) &= 0.99 & P(B = 0|A = 1) &= 0.1 \\ P(A = 1) &= 0.001 & P(A = 0) &= 0.999 \end{aligned}$$

Using the law of total probability, we can find the marginal probabilities for B:

$$P(B = 0) = \sum_{i=0}^1 P(B = 0|A = i)P(A = i) = (0.99)(0.999) + (0.1)(0.001) = 0.98911$$

$$P(B = 1) = \sum_{i=0}^1 P(B = 1|A = i)P(A = i) = (0.01)(0.999) + (0.9)(0.001) = 0.01089$$

So, the probability that a person has the disease given the test outcome is negative is:

$$\begin{aligned} P(A = 1|B = 0) &= \frac{P(B = 0, A = 1)P(A = 1)}{P(B = 0)} \\ &= \frac{(0.1)(0.001)}{0.98911} \\ &= 1 \times 10^{-4} \end{aligned}$$

The probability that a person has the disease given the outcome is positive is:

$$\begin{aligned} P(A = 1|B = 1) &= \frac{P(B = 1, A = 1)P(A = 1)}{P(B = 1)} \\ &= \frac{(0.9)(0.001)}{0.01089} \\ &= 0.082645 \end{aligned}$$

- Next, we need to consider what happens if the probability of a randomly selected person having the disease is higher, $P(A = 1) = 0.2$ and $P(A = 0) = 0.8$. This also requires us to recalculate the probabilities for B:

$$P(B = 0) = \sum_{i=0}^1 P(B = 0|A = i)P(A = i) = (0.99)(0.8) + (0.1)(0.2) = 0.812$$

$$P(B = 1) = \sum_{i=0}^1 P(B = 1|A = i)P(A = i) = (0.01)(0.8) + (0.9)(0.2) = 0.188$$

The updated conditional probabilities are:

$$\begin{aligned} P(A = 1|B = 0) &= \frac{P(B = 0, A = 1)P(A = 1)}{P(B = 0)} \\ &= \frac{(0.1)(0.2)}{0.812} \\ &= 0.0246 \end{aligned}$$

$$\begin{aligned} P(A = 1|B = 1) &= \frac{P(B = 1, A = 1)P(A = 1)}{P(B = 1)} \\ &= \frac{(0.9)(0.2)}{0.188} \\ &= 0.957 \end{aligned}$$

For the marginal probability first given for A , it was very unlikely that a person had the disease and both $P(A = 1|B = 0)$ and $P(A = 1, B = 1)$ are accordingly small, especially since 10% of the time, the test will come out negative even when a person has the disease. However, when the disease becomes more prevalent, the joint probabilities, particularly for $P(A = 1|B = 1)$, are much larger given the fact that more people at random are more likely have the disease.

Problem 3:

In Example 7 in Note Set 1 (two Gaussians with equal variance) prove that:

$$P(a_1|x) = \frac{1}{1 + e^{-(\alpha_0 + \alpha x)}}$$

and derive expressions for α_0 and α .

SOLUTION: As Example 7 states, we start with the following facts.

$$p(x|a_1) \sim N(\mu_1, \sigma)$$

$$p(x|a_2) \sim N(\mu_2, \sigma)$$

In addition, we assume the value of a_1 is known. Using Bayes Theorem, we get

$$p(a_1|x) = \frac{p(x|a_1)p(a_1)}{p(x)}$$

From the law of total probability, we know

$$\sum_a p(x|a)p(a) = p(x)$$

So, substituting this in, we get

$$p(a_1|x) = \frac{p(x|a_1)p(a_1)}{p(x|a_1)p(a_1) + p(x|a_2)p(a_2)}$$

Since probabilities must sum to 1, we can replace $p(a_2)$ with $1 - p(a_1)$ and $p(x|a_2)$ with $1 - p(x|a_1)$.

$$\begin{aligned} p(a_1|x) &= \frac{p(x|a_1)p(a_1)}{p(x|a_1)p(a_1) + p(x|a_2)(1-p(a_1))} \\ p(a_1|x) &= \frac{1}{1 + \frac{p(x|a_2)(1-p(a_1))}{p(x|a_1)p(a_1)}} \end{aligned}$$

Substituting the definition of $p(x|a_1)$ and $p(x|a_2)$,

$$\begin{aligned} p(a_1|x) &= \frac{1}{1 + \frac{1-p(a_1)}{p(a_1)} e^{-\frac{(x-\mu_2)^2}{2\sigma^2} + \frac{(x-\mu_1)^2}{2\sigma^2}}} \\ p(a_1|x) &= \frac{1}{1 + \frac{1-p(a_1)}{p(a_1)} e^{-\frac{\mu_1^2 - \mu_2^2}{2\sigma^2} + \frac{2(\mu_2 - \mu_1)}{2\sigma^2} x}} \\ p(a_1|x) &= \frac{1}{1 + e^{\ln(\frac{1-p(a_1)}{p(a_1)}) + \frac{\mu_1^2 - \mu_2^2}{2\sigma^2} + \frac{\mu_2 - \mu_1}{\sigma^2} x}} \\ p(a_1|x) &= \frac{1}{1 + e^{-\left(\frac{\mu_2^2 - \mu_1^2}{2\sigma^2} - \ln\left(\frac{1-p(a_1)}{p(a_1)}\right) + \frac{(\mu_1 - \mu_2)}{\sigma^2} x\right)}} \end{aligned}$$

Finally, we can rewrite this as

$$p(a_1|x) = \frac{1}{1 + e^{-(\alpha_0 + \alpha x)}}$$

Where α_0 and α are given by,

$$\begin{aligned} \alpha_0 &= \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} - \ln\left(\frac{1-p(a_1)}{p(a_1)}\right) \\ \alpha &= \frac{\mu_1 - \mu_2}{\sigma^2} \end{aligned}$$

Problem 4:

In the problems below assume that X and Y are real-valued random variables:

- Prove that $E[ax + b] = aE[x] + b$ where expectation E is defined with respect to $p(x)$ and a and b are arbitrary scalar constants.
- Prove that $E[x + y] = E[x] + E[y]$, where the expectation $E[x + y]$ is defined with respect to $p(x, y)$, and $E[x]$ and $E[y]$ are each defined with respect to $p(x)$ and $p(y)$ respectively. Be careful to explain clearly and precisely how you get from one line to the next in your proof.

SOLUTION:

1. By definition, we know

$$E[g(x)] = \int g(x)p(x)dx$$

Thus, plugging in $g(x) = ax + b$, we get

$$E[ax + b] = \int (ax + b)p(x)dx = \int ax \cdot p(x)dx + \int b \cdot p(x)dx$$

Pulling the constants out of the integrals, and using the fact that the pmf p integrates to 1, we get,

$$E[ax + b] = a \int x \cdot p(x)dx + b \int p(x)dx = a \cdot E[x] + b \cdot 1 = a \cdot E[x] + b$$

2. Plugging our equation into the definition of expected value, we have

$$E[x + y] = \int \int (x + y)p(x, y)dxdy = \int \int x \cdot p(x, y)dxdy + \int \int y \cdot p(x, y)dxdy$$

Moving integrals to the inside when possible,

$$E[x + y] = \int x \int p(x, y)dydx + \int y \int p(x, y)dxdy$$

Now, using the fact that $\int p(a, b)da = p(b)$, we get

$$E[x + y] = \int x \cdot p(x)dx + \int y \cdot p(y)dy$$

By the definition of expectation, this gives,

$$E[x + y] = E[x] + E[y]$$

Problem 5:

We have 5 random variables A, B, C, D, E . Draw the appropriate directed graphical model for each of the cases below. In each case you should assume that only the dependencies that are stated are present (i.e., if the information given does not say or imply that 2 variables are dependent then you should assume they are independent (or conditionally independent if appropriate)).

- B, C , and D are each dependent on A but are conditionally independent given A . A depends on E .
- B, C, D , and E are marginally independent, and A depends on all four.
- A depends on B and C depends on D . E is independent of all of the others.

SOLUTION: Lists of the directed edges in each graphical model:

1. $E- > A, A- > B, A- > C, A- > D$.
2. $B- > A, C- > A, D- > A, E- > A$,
3. $B- > A, D- > C, E$ has no edges.

Problem 6:

(20 points) We have 3 binary random variables A, B, C . We observe 1000 random samples from the joint distribution $P(A, B, C)$. The observed counts are as follows:

A	B	C	Count
0	0	0	810
0	0	1	81
0	1	0	0
0	1	1	9
1	0	0	81
1	0	1	2
1	1	0	9
1	1	1	8

Your goal is to find the graphical model with the fewest edges that exactly explains this data. By “exactly explains” we mean that if you replace the counts with frequency-based estimates of probabilities (divide each of the counts by 1000), then the probabilities produced by the graphical model are directly proportional to the counts above¹. Your answer should include the following information:

- a brief explanation in words and/or equations describing the procedure you used to determine your graphical model.
- a diagram of the graphical model
- an equation representing the factorization of the joint probability implied by the graphical model
- each of the probability tables needed to specify the graphical model.

SOLUTION:

The directed graphical model that corresponds to this data has A and C as parents of B and no other edges. This is the simplest model (2 edges rather than the maximum of 3) that can exactly reproduce the data above. We will discuss the solution of this problem in class, but essentially these types of “structure learning” problems can only be solved (except in certain special cases) by exhaustive search over all possible graphs, or by heuristic search if we want the search to be computationally tractable. Many of you noticed from the data that A and C were marginally independent, and proceeded from there.

The factorization corresponding to the graphical model is:

$$P(A, B, C) = P(A)P(B|A, C)P(C)$$

Below is the probability of each variable taking on the value 1, given all possible values of its parents. The corresponding probabilities of taking the value 0 are just 1 minus these values.

¹In later classes we will discuss better ways to estimate these probabilities but for now this simple method is fine.

$$\begin{aligned}P(A = 1) &= 0.1 \\P(C = 1) &= 0.1 \\P(B = 1|A = 0, C = 0) &= 1.0 \\P(B = 1|A = 1, C = 0) &= 0.1 \\P(B = 1|A = 0, C = 1) &= 0.1 \\P(B = 1|A = 1, C = 1) &= 0.8\end{aligned}$$

Problem 7: MATLAB Assignment

(30 points)

For your MATLAB assignment please turn in (a) clearly documented printout of your code and (b) any plots or write-ups requested below (all attached to your written problem solutions above).

1. See the class Web page for instructions on how to access MATLAB at UCI via various PC and Unix workstations (either is fine). Please take at least one or two hours to go through some of the online tutorials (see the class Web page) and online documentation to familiarize yourself with MATLAB.
2. You are to write a simple MATLAB function called `twogaussian.m`. The purpose of this exercise is primarily to get you working with MATLAB, to help you get some intuition about what data from a 2-dimensional Gaussian distribution looks like, and to write a relatively simple function in MATLAB. A template (that contains 90% of the necessary code) is available for you to download from the Web page, as is the function `mvnrnd.m` that you will also need. You are to fill in the few lines (denoted with “dots”) that are missing in `template.m`. Use “help plot”, “doc plot”, “help mvnrnd”, etc., to learn more about how the various functions work. Feel free to experiment with the code to generate different Gaussian shapes beyond what is required.

This is the header for the MATLAB function:

```
function [data1, data2] = twogaussian(n1,mu1,cov1,n2,mu2,cov2);
%
% [data1, data2] = twogaussian(n1,mu1,sigma1,n2,mu2,sigma2);
%
% Function to simulate data from 2 Gaussian densities in d dimensions
% and to plot the data in the first 2 dimensions
%
% INPUTS:
%     n1, n2: two integers, size of data set 1 and 2 respectively
%     mu1, mu2: two vectors of dimension 1 x d, means
%               for data set 1 and 2
%     cov1, cov2: two matrices of dimension d x d, covariance
%                 matrices for data set 1 and 2 respectively
%
% OUTPUTS:
%     data1:  n1 x d matrix of data for data set 1
%     data2:  n2 x d matrix of data for data set 2
```

Be sure to test your function with some simple cases to validate that it is working correctly.

3. Produce two-dimensional “scatter plots” for the following four cases. For all cases let n_1 and n_2 be 500 each (the number of points plotted for data set 1 and data set 2) and let $\mu_1 = [0 \ 0]$ and $\mu_2 = [2 \ 2]$.

- $\text{cov1} = \text{cov2} = \mathbf{I}$, where \mathbf{I} is the identity matrix.
- $\text{cov1} = \mathbf{I}$, $\text{cov2} = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}$
- $\text{cov1} = \mathbf{I}$, $\text{cov2} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$
- $\text{cov1} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$, $\text{cov2} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$

Note color plots are **not** necessary for your printouts that you hand in with your homework—but if your plots are not in color then you need to modify the code so that the data points from “data1” are plotted with one symbol (e.g., “+”) and data points from “data2” are plotted with a different symbol (e.g., “o”). If you would like to save paper feel free to print multiple plots on a single page (e.g., using the subplot function).

SOLUTION:

Sample code (relevant part):

```
% Call the function mvnrnd.m to generate the two data sets
data1 = mvnrnd(mu1,cov1,n1); data2 = mvnrnd(mu2,cov2,n2);

% Now plot the two data sets as a two-dimensional scatter plot
% if d = 2: plot dimension1 on the xaxis and dimension 2 on the
% yaxis. Plot the points from data1 as green dots 'g.', and the
% points from data2 as red dots 'r.'.
if plotflag==1
    figure % open a figure window
    plot(data1(:,1),data1(:,2),'g.') % now plot data1
    hold on; % hold the figure to overlay a 2nd plot
    plot(data2(:,1),data2(:,2),'r.') % now plot data 2
    xlabel('Dimension 1');
    ylabel('Dimension 2');
    title('Simulation of data from two Gaussians in two dimensions');
    plot_gauss_parameters(mu1,cov1,1,2,'g');
    plot_gauss_parameters(mu2,cov2,1,2,'r');
end
```