18.05 Problem Set 5, Spring 2018

(due outside 4-174 Monday, Mar. 19 at 9:30 AM)

Problem 1. (40 pts.) **Maximum likelihood estimates**

(a) The Pareto distribution with parameter α has range $[1, \infty)$ and pdf

$$f(x) = \frac{\alpha}{x^{\alpha+1}}$$

The parameter α must be in the range $(0, \infty)$. Suppose the data 5, 2, 3, 2.5, 6 was drawn independently from such a distribution. Find the maximum likelihood estimate (MLE) of α.

(b) A game is played with 3 identical looking urns containing colored balls. Urn 1 contains 6 red, 7 green, 5 blue balls; urn 2 contains 4 red, 9 green, 3 blue balls; urn 3 contains 5 red, 10 green, 6 blue balls.

The rules are that a player picks one urn at random and draws 3 balls without replacement. They then have to guess which urn they chose. Suppose that in order a player picks a red, a green and a red ball. Which urn is the maximum likelihood estimate for the chosen urn?

(c) I'm always late for dinner. Exactly how late follows a uniform(0,b) distribution where b is not known. (Apparently there is a time beyond which even I won't be late.) My wife wanted to figure out the value of b so she recorded my lateness in minutes for 4 days. The results were 2.5, 19.75, 12.0, 7.0. What is the maximum likelihood estimate for b based on this data.

Hint: this is not a calculus problem.

(d) Linear regression. Bivariate data means data of the form

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n).$$

One typical analysis is to make a scatter plot and find the line that best fits the data. This is called a linear regression model. It turns out that, under some assumptions about random variation of measurement error, one way to find a "best" line is by solving a maximum likelihood problem.

The model assumes that the value of x_i is random and y_i has a linear relation to x_i plus some measurement error. That is

$$y_i = ax_i + b + \text{ random measurement error.}$$

The model assumes the measurement errors are independent and identically distributed and follow a $\mathrm{N}(0, \sigma^2)$ distribution.

This goal is to find the values of the parameters a and b that give the MLE for this model. To guide you we note that the model says that

$$y_i \sim \mathrm{N}(ax_i + b, \sigma^2).$$

Also remember that you know the density function for this distribution.

(i) For a general datum (x_1, y_1) give the likelihood and log likelihood functions (these will be functions of y_1, x_1, a, b, and σ.)

(ii) Consider the data $(1, 1)$, $(3, 3)$, $(1.5, 4)$. Assume that σ is a known constant and find the maximum likelihood estimate for a and b.

Note: since there are *two* variables a and b, in order to find a critical point you will have to take partial derivatives and set them equal to 0.

(iii) Use R to plot the data and the regression line you found in part (ii) The commands `plot(x,y, pch=19)` and `abline()` will come in handy. For `abline` be careful: the parameter `a` is the intercept and `b` is the slope – exactly the opposite of our usage. Print the plot and turn it in.

(iv) In order to compute the MLE in part (ii) we assumed that sigma was a known constant. Why did it turn out to be unnecessary to assume that the value of sigma was known?

Problem 2. (15 pts.) **Monty Hall: Sober and drunk.**

Recall the Monty Hall problem: Monty hosts a game show. There are three doors: one hides a car and two hide goats. The contestant Elan picks a door, which is not opened. Monty then opens another door which has a goat behind it. Finally, Elan must decide whether to stay with his original choice or switch to the other unopened door. The problem asks which is the better strategy: staying or switching?

To be precise, let's label the door that Elan picks by A, and the other two doors by B and C. Hypothesis H_A is that the car is behind door A, and similarly for hypotheses H_B and H_C.

(a) In the usual formulation, Monty is sober and knows the locations of the car and goats. So if the contestant picks a door with a goat, Monty always opens the other door with a goat. And if the contestant picks the door with a car, Monty opens one of the other two doors at random. Suppose that sober Monty Hall opens door B, revealing a goat. So the data is: 'Monty showed a goat behind B'. Our hypotheses are 'the car is behind the first door', etc. Make a Bayes table with prior, likelihood and posterior. Use the posterior probabilities to determine the best strategy.

(b) Now suppose that Monty is drunk, i.e. he has completely forgotten where the car is and is only aware enough to randomly open one of the two doors not chosen by the contestant. It's entirely possible he might accidentally reveal the car, ruining the show.

Suppose that drunk Monty Hall opens door B, revealing a goat. Make a Bayes table with prior, likelihood and posterior. Use the posterior probabilities to determine the best strategy. (Hint: the data is the same but the likelihood function is not.)

(c) Based on Monty's pre-show behavior, Elan thinks that Monty is sober with probability 0.7 and drunk with probability 0.3. Repeat the analysis from parts (a) and

(b) in this situation.

Problem 3. (20 pts.) **Posterior probability and posterior predictive probability**
For this problem assume there are 2 four-side, 3 six-sided, 4 eight-sided, 5 twelve-sided and 6 twenty-sided dice in a bag. One of the dice is chosen at random and rolled repeatedly.

You will find it easiest to use R to do the computations for this problem.

(a) Suppose the first roll is a 6. Find the posterior probabilities that the chosen die is 4, 6, 8, 12, or 20 sided.

(b) Find the prior (to the first roll) predictive probability that the first roll is a 6.

Find the posterior (to the first roll) predictive probability that the next roll is a 6.

(c) Suppose that by some extraordinary chance the first n rolls are all 6. Find the posterior probabilities that the chosen die is each type of die. Your answer will need to depend on n.

What is the limit of these probabilities as n goes to infinity? This can be done algebraically, but if you prefer you can use R to do this numerically. Explain in a few sentences why this makes sense.

(d) Assuming the first 10 rolls were all 6, find the predictive probabilities for the 11th roll. That is, find $P(x_{11}|x_1 = 6, x_2 = 6, \cdots, x_{10} = 6)$ for $x_{11} = 1, 2, \cdots, 20$.

Problem 4. (15 pts.) **What are the odds?**
A screening test for a disease is reasonably accurate: it has a 10% false positive rate and a 2% false negative rate. The base rate of the disease in the population is 0.1%.

(a) What are the prior odds a random person has the disease?

(b) Suppose this person tests positive. Find the Bayes factor and use it and your answer to part (a) to find the posterior odds they have the disease.

(c) A second more accurate test is going to be given. This test has a 1% false positve rate and a 1% false negative rate. What are the posterior (to the first test) predictive odds that the second test will be positive.

You should assume that the tests are conditionally independent. That is, if a person has the disease then the two tests are independent, and likewise, if a person doesn't have the disease the tests are independent.

Problem 5. (10 pts.) **Legal Trickery**
(a) [Mackay, *Information Theory, Inference, and Learning Algorithms*, and OJ Simpson trial] Mrs S is found stabbed in her family garden. Mr S behaves strangely after her death and is considered as a suspect. On investigation of police and social records it is found that Mr S had beaten up his wife on at least nine previous occasions. The prosecution advances this data as evidence in favor of the hypothesis that Mr S is

guilty of the murder. 'Ah no,' says Mr S's highly paid lawyer, '*statistically*, only one in a thousand wife-beaters actually goes on to murder his wife. So the wife-beating is not strong evidence at all. In fact, given the wife beating evidence alone, it's extremely *unlikely* that he would be the murderer of his wife – only a 1/1000 chance. You should therefore find him innocent.'

Is the lawyer right to imply that the history of wife-beating does not point to Mr S's being the murderer? Or is this a legal trick? If the latter, what is wrong with his argument?

Use the following scaffolding to reason precisely:
Hypothesis: M = 'Mr S murdered Mrs S'
Data: K = 'Mrs S was killed', B = 'Mr S had a history of beating Mrs S'
How is the above probability 1/1000 expressed in these terms? How is the (posterior) probability of guilt expressed in these terms? How are these two probabilities related? Hint: Bayes' theorem, conditioning on B throughout.

(b) [True story] In 1999 in Great Britain, Sally Clark was convicted of murdering her two sons after each child died weeks after birth (the first in 1996, the second in 1998). Her conviction was largely based on the testimony of the pediatrician Professor Sir Roy Meadow. He claimed that, for an affluent non-smoking family like the Clarks, the probability of a single cot death (SIDS) was 1 in 8543, so the probability of two cot deaths in the same family was around "1 in 73 million." Given that there are around 700,000 live births in Britain each year, Meadow argued that a double cot death would be expected to occur once every hundred years. Finally, he reasoned that given this vanishingly small rate, the far more likely scenario is that Sally Clark murdered her children.

Carefully explain at least two errors in Meadow's argument.