EAEEE 4257 Environmental Data Modeling and Analysis

Homework 0

Upmanu Lall

Due Monday, January 22, 2018

Prerequisites for this course include some exposure to probability or statistics and linear algebra. This homework will give you an opportunity to test your background. All questions have been designed so that you can solve them *without using a computer*.

If you need some review, you may find the following resources helpful:

- Zico Kolter's brief review of Linear Algebra (`http://cs229.stanford.edu/section/cs229-linalg.pdf`) is useful for anyone who has seen linear algebra but needs some review
- Joe Blitzstein's Harvard course (`http://projects.iq.harvard.edu/stat110`) has lecture notes and practice problems for an in-depth look at introductory probability
- Naresh Devineni's Data Analysis classroom blog (`http://www.dataanalysisclassroom.com/`) has short lessons (blog posts) on a variety of topics related to statistics and data analysis. It's recommended to start from the beginning.

Although you will be allowed to work in groups for some later homeworks, **work alone for this assignment**.

Grading and Submission Instructions

Please turn in your answers to this course as a `.pdf` file on Courseworks using the `Assignments` tab. You can type your answers (for example using LaTeX) or write clearly by hand and scan them (please make sure the scanned document is legible!)

You will receive full credit for this homework if you submit complete answers, by the due date, and following the submission instructions, regardless of how many questions you answer correctly.

The purpose of this assignment is for you to evaluate whether this class is a good fit for you. Use the following rubric:

- If you find that you are able to answer all questions easily, the first several weeks of this class may be easy for you.
- If you are able to answer most questions, but need to look up some material and/or have some questions that you are unable to answer, you are well-prepared for this class.
- If you are able to answer some of the questions easily, need to look up information for some others, and ar eunable to answer some questions, this class is a good fit for you but some self-study may be helpful during the first weeks.
- If you are unable to answer most or all of these questions, this class will be challenging for you and you should not enroll without talking to me or the TA.

1 Open-Ended

1. Have you ever taken a linear algebra course before? If so, what course?
2. Have you ever taken a probability course before? If so, what course?
3. Have you ever taken a statistics course before? If so, what course?
4. Have you ever taken a multi-variate calculus before? If so, what course?
5. Have you ever written computer code before? If so, when and what language?
6. Have you ever used `R` before? What is your level of experience?
7. Why are you taking this course? What do you hope to learn?
8. A key component of this course is a final project, in which you will use the tools you learn in this class to analyze a data set of interest to you. Do you have any idea yet what you would like to do for your final project?

2 Linear Algebra

Let us define

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} e & f \\ g & h \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} k \\ \ell \end{bmatrix} \tag{2.1}$$

We will also use the notation $(\cdot)^T$ to mean the transpose of $(\cdot)$. The transpose is sometimes written as $(\cdot)'$ but this can lead to confusion and we will avoid writing it this way.

9. What is $A_{2,1}$?
10. What is A^T?
11. What is AB?
12. What is x^TAx?
13. What is x^Tx?
14. What is xx^T?

Now let C be a matrix of shape 10×2.

15. Is AC defined? If so what shape is the resulting matrix?
16. Is CA defined? If so what shape is the resulting matrix?

3 Probability and Random Variables

Define y and w as random variables

$$y \sim \mathcal{N}(\mu_y, \sigma_y) \tag{3.1}$$
$$w \sim \mathcal{N}(\mu_w, \sigma_w) \tag{3.2}$$

where $\mathcal{N}(\mu, \sigma)$ denotes a normal variable with mean μ and standard deviation σ.

17. What is the expected value of y^2, written $\mathbb{E}\left[y^2\right]$?
18. If y and w are independent, what is the distribution of $y + w$?

Now let us consider a random 5-card poker hand delt from a standard 52-card deck[1] What is the probability that the 5-card hand is:

19. A pair (two cards of the same number but different suit)
20. A flush[2] (all cards of the same suit).

Also consider the following longer problems

[1] This is a deck with 4 suits, each of which has 13 cards: the numbers 2-10, plus the jack (J), queen (Q), king (K), and ace (A). See https://en.wikipedia.org/wiki/Standard_52-card_deck.

[2] Do not exclude a royal flush

21. A spam filter is designed based on common words and phrases.[3] Suppose that 75% of email is spam (and so 25% of email is not spam). In 25% of spam emails, the phrase "congratulations" is used, whereas this phrase is used in only 10% of non-spam emails. A new email has just arrived, which includes the word "congratulations". What is the probability that this email is spam?

22. A coin is tossed repeatedly until it lands Tails for the first time. Let Y be the number of tosses that are required (including the final one that lands Tails). If it is a fair coin (equal probability of heads or tails), what is the probability that:

    a) $Y = 1$

    b) $Y = 2$

    c) $Y = 100$ (write the expression you would solve, but do not attempt to write it as a decimal)

4 Statistics and Data Analysis

The following question is more open-ended than the previous questions. Please note that we do not expect everyone who enrolls in the course to be able to answer all of these questions. Please answer to the best of your abilities and be as clear as possible about any assumptions you have made. If there are questions that you don't know how to answer, please indicate that. If you would like to use a computer or calculator for this problem, you may do so but please indicate on your answer that you used a computer or calculator (you will not be penalized!)

Water quality data has been collected at a lake for several years. Answer the following questions given the data in Table 4.1:

23. What are the average and the median of the Dissolved Oxygen (DO) concentration?

24. If one of the observations (which one?) is likely to be a typographical error, which estimate, median or average do you think is more typical of the data? Why?

25. What is the standard deviation of the DO concentration? How does it change if you skip the observation that may be wrong? What do you learn from this?

26. Based on just the data, what do you think is the chance (probability) that DO can be below $8\,\mathrm{mg\,L^{-1}}$ – the level that is critical for the fish in the lake?

27. What is the chance that you can have 2 consecutive events with DO $< 8\,\mathrm{mg\,L^{-1}}$?

28. What is the chance that in 10 events you will have exactly 2 events with DO $< 8\,\mathrm{mg\,L^{-1}}$?

[3] This would be a very primitive spam filter!

Event	DO [$\mathrm{mg\,L^{-1}}$]	Dead Fish Present
1	7	Y
2	6	N
3	9	Y
4	11	N
5	120	N
6	11	N
7	9	N
8	7	N
9	5	Y
10	9	N

Table 4.1: Lake water quality data

29. What is the chance that in 10 events you will have at least 1 event with DO < $8\,\mathrm{mg\,L^{-1}}$?

30. What did you assume about the data in the calculations above? Can you name the process or probability distribution you used?

31. A new data set was collected by a citizen group. This data is shown in Table 4.2. Unfortunately they did not report the actual DO values because their sensor can only report whether the value is above or below $8\,\mathrm{mg\,L^{-1}}$. Based on what you estimated above about the probability or likelihood of new data, is this sample consistent with the data that was collected previously? Specifically what is the chance that 4 out of 5 events are Y (DO < $8\,\mathrm{mg\,L^{-1}}$)? Discuss how you approached the problem. Note that you can think of this as 4 Y in a row followed by a N, or as 4 Y out of 5. What is the difference? Which one did you choose?

Event	DO < $8\,\mathrm{mg\,L^{-1}}$
1	Y
2	Y
3	Y
4	Y
5	N

Table 4.2: Lake water quality data collected by citizen scientists

32. Now let us say that you wanted to use the Normal distribution to fit the original DO data. This distribution has 2 parameters: mean and standard deviation. Discuss

which values you would use for this data, and then use the Normal distribution to compute the probability that DO $< 8\,\mathrm{mg\,L^{-1}}$.

33. Do you think it is a good idea to use the Normal distribution with this data? Why or why not? Would you take logs of the data first? If you say yes, then use that procedure to estimate the probability that DO $< 8\,\mathrm{mg\,L^{-1}}$ from that data. Which one agrees better with your original analysis? How would you test whether the Normal distribution applied to the original data or to the logs of the data is a better choice in this case? You don't need to actually do the test, just name it and discuss its application.

34. In Table 4.1 we see that whether or not dead fish were found was also recorded for each event. What is the probability or chance of finding dead fish in the lake? What is the probability of finding dead fish if DO $< 8\,\mathrm{mg\,L^{-1}}$? What is the probability of finding dead fish if DO $> 8\,\mathrm{mg\,L^{-1}}$?

35. Now let us introduce some notation: let

$$y_i = \begin{cases} 1 & \text{if dead fish found} \\ 0 & \text{else} \end{cases} \tag{4.1}$$

and

$$x_i = \begin{cases} 1 & \text{if DO} < 8\,\mathrm{mg\,L^{-1}} \\ 0 & \text{else} \end{cases} \tag{4.2}$$

Now if we are interested in the joint probability $P(y = 1, y = 1)$ we could estimate this as all the events were estimated and DO $< 8\,\mathrm{mg\,L^{-1}}$, divided by the total number events – similarly for the other joint probabilities. Looking at Table 4.1, we would estimate $P(y = 1, x = 1) = \frac{2}{10}$. What are:

a) $P(y = 1, x = 0)$?

b) $P(y = 0, x = 1)$?

c) $P(y = 0, x = 0)$?

Now define $P(y = 1|x = 1)$ as the conditional probability that $y = 1$ if we observe that $x = 1$. Events 1, 2, 8, and 9 correspond to $x = 1$. So we have 4 observations that $x = 1$, of which 2 have $y = 1$. Thus $P(y = 1|x = 1) = \frac{2}{4} = \frac{1}{2}$. What are all the other conditional probabilities that we can estimate? Please estimate them.

Now that you have estimated these conditional probabilities, what is the probability that DO $< 8\,\mathrm{mg\,L^{-1}}$ if a couple taking a walk spotted a dead fish? How did you compute this? Suppose that you had been given only the joint probabilities that you computed, and also the marginal probabilities $P(x = 1)$ and $P(y = 1)$: could you compute $P(x = 1|y = 1)$? Lay out your argument and if you decided to apply Bayes rule, explain what you did in plain English.

5 Slack and Syllabus

36. Please confirm that you have joined the Slack page

37. Please confirm that you have joined the `math-stats` and `r-computing` channels on the Slack channel (you may need to join them manually)

38. Please confirm that you have read the syllabus at https://github.com/jdossgollin/EnvDataS18/. *Please note that that the syllabus may be updated before the beginning of class.*

39. Read the academic honesty policy at http://www.cs.columbia.edu/education/honesty/ and confirm that you have read it. If you have any questions about the academic honor policy in this course, please write them here so that we can clarify. Please write the following:

    > I, `Your Name`, will abide by the academic honesty policy presented above and on the course syllabus. If I have any questions about the academic honesty policy I will discuss them with the professor or TA.