Ast3: Death's End

Due mar 27 at 11:59pm on moodle. No email submissions.

The assignment can easily be done individually but pairs/groups are *encouraged*. Only one group member should submit the assignment. Include all group member names. Other group members will have a moodle warning they didn't do an assignment; ignore that. I manually enter grades for everyone.

You're welcome (actually, you're encouraged) to use online resources to help you find facts/insights for this assignment. This includes using Generative AI. However, you are ultimately responsible for content, and you must *document* which knowledge/tool you use and where. Any content without explicit citation is assumed to be your own work.

This assignment probably not require citations but if you have any, they should be properly formatted bibliography/references. See assignment 1 requirements regarding citations.

Assignment objectives

- Code up your own non-parametric smoothers and compare performances
- Decide which information is most relevent to include in a constrained report

Your assignment will be a very brief report in an unconventional structure.

1 Questions

Using the palmer penguin data, subset your data so that you only model males for this assignment.

I want you to create kernel smoother to estimate whether the continuous variables tell you whether the species is a male Gentoo (vs a male of another Species) using all of the other continuous variables.

Create a new variable, `isGentoo` that 1 when the species is Gentoo, 0 otherwise. This will be your response.

1. Because your response is binary (0 or 1), and you are modelling its expected value, what is your kernel smoother actually modelling? Hint: one of two words, both of which start with p.
2. You will create a kernel smoother using all of the other continuous measurements and evaluate each of the following two approaches (four combinations in total)

- Different ways of standarizing the scales
  - Using z-scores (univariately) on each of the variables before applying distances
  - Turning all (univariate) variables into quantiles
- Choice of distance metric
  - The L_1 (Manhattan) distance between $\mathbf{a} = (a_1, a_2, \dots)$ and $\mathbf{b} = (b_1, b_2, \dots)$ is defined by

$$dist(\mathbf{a}, \mathbf{b}) = |a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3| + \dots$$

  - The L_2 (Euclidean) distance between $\mathbf{a}$ and $\mathbf{b}$ is defined by

$$dist(\mathbf{a}, \mathbf{b}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + \dots}$$

Choose some simple method of correctness regarding whether your estimated kernel smoother guesses if it s a Gentoo correctly. Note

- you don't have to distinguish false positives/negatives, but you can if you really want to
- You do not have to create a training/test partition, just report the training errors

As part of your analysis, I want to see which combination of the four seems to perform best (or whether several perform approximately the same). I also want you to explain any other decisions that you had to make in determining this. Your answer should make sense to anyone in the class who has not seen the assignment, but attended class all the way up and including to the kernel smoother class.

Write your answer to this question in a maximum 2-page report (excluding figures/tables/code). You are allowed a maximum of one figure and one table (you can use only one, or none, as well), which can be at the end of the report. Include code in your appendix. The formatting should follow previous assignments. For your report, do not describe algorithmic or coding details, just explain what's being assessed and the results.

Ensure that you bold/underline a single investigative question that your report answers, and at the end bold/underline the result. Each is a single question.