NAME: Kevin Motia
DATE: 2024-02-08
Homework: Assignment 1
Data Science II

Problem 1

In a data-driven context, the statement "the map is not the territory" conveys the message that a model of a system does not represent the system in its totality. This is exemplified in the ways in which these models fail in representing the complexity found in their real-world counterparts. There are numerous examples in life to show this.

The 2019-2020 Boeing 737 MAX groundings are representative of a huge decision-theoretic analytical failure. Spanning from 2018 to 2019, Lion Air Flight 610 and Ethiopian Airlines Flight 302 crashed, killing 346 people. Boeing relied on simulations to assess the safety of the planes' Maneuvering Characteristics Augmentation system, but these simulations did not account for the risks associated with when this system malfunctioned. When even a single sensor failed, resulting in bad data, the system was allowed to autonomously initiate a nose dive. These 2 crashes lead to the grounding of all Boeing 737 MAX planes, an investigation into the system certification process, and more rigorous safety analyses.

Another (but much less frightening) example of a failure of a decision-theoretic analysis occurred to me as an undergraduate when I was choosing a course load. I was majoring in physics, and attempting to optimize course selection in order to reduce the number of difficult courses I had to take together while still choosing those that would allow me to graduate on time. I surveyed those who had taken the courses in previous years to get an idea of each course's workload. My analysis failed when I neglected to account for changes to curriculum across years, which meant that my analysis of course work per class was flawed. It led to me taking on significantly more work than I was prepared for. The experience taught me to rely more on the syllabus presented by the course rather than word-of-mouth.

Overall, these instances show that the "map is not the territory" by providing examples where a model is unable to account for various factors that could happen in the actual system. Such a limitation is important to recognize especially in the field of data science due to the fact that decisions in the field are data-driven through things like models and simulations. Computational tools like these often aim to simplify complex systems, which potentially means that variables that are critical to realism are overlooked. Understanding the limitations of models is a key aspect of data-driven decision-making. Through the acknowledgement of the uncertainties and assumptions that are inherent to the model, we can seek to validate its accuracy using real-world data. Additionally, it involves considering alternative assumptions and perspectives to avoid the risk of decision-theoretic failures like the one I encountered when picking my course load.

In conclusion, the statement "the map is not the territory" emphasizes the importance in recognizing a model's limitations in a data-driven context. While the impact of a decision-theoretic failure can vary greatly, being attentive of the potential pitfalls in modeling and simulations can lead us to avoid those that compromise their functions.

Problem 2

2.1 Bayes Theorem

$$Pr(cancer|+B) = \frac{(.75)(.9)}{(.75)(.9)+(.04)(.1)} = .9941$$

$$Pr(cancer|-B) = \frac{(.25)(.9)}{(.25)(.9)+(.96)(.1)} = .7009$$

$$Pr(benign|+B) = 1 - .9941 = .0059$$

$$Pr(benign|-B) = 1 - .7009 = .2991$$

2.2 Biopsy with Malignant Tumor

$$U(R)_{cancer} = (.9941)(16.7) + (.0059)(34.8) - 1 = 15.8 months$$

$$U(S)_{cancer} = .35(0) + .65[(.9941)(20.3) + (.0059)(34.8) - 1] = 12.6 months$$

$$U(N)_{cancer} = (.9941)(5.6) + (.0059)(34.8) = 5.77 months$$

2.3 Biopsy with Benign Tumor

$$U(R)_{benign} = (.7009)(16.7) + (.2991)(34.8) - 1 = 21.114 months$$

$$U(S)_{benign} = .35(0) + .65[(.7009)(20.3) + (.2991)(34.8) - 1] = 15.36 months$$

$$U(N)_{benign} = (.7009)(5.6) + (.2991)(34.8) - 1 = 14.33 months$$

2.4 Combining Biopsy Scenarios

$$U(R)_{biopsy} = .75(15.8) + .04(21.114) - 1 + .06(0) = 11.69 months$$

$$U(S)_{biopsy} = .75(12.6) + .04(15.36) - 1 + .06(0) = 9.06 months$$

$$U(N)_{biopsy} = .75(5.77) + .04(14.33) - 1 + .06(0) = 3.90 months$$

From class, we know that the U(R), U(S), and U(N) are 17.5 months, 13.5 months, and 8.5 months respectively. Since getting a biopsy leads to a lower life expectancy for each treatment option, we don't suggest that this man undergoes a biopsy.

Problem 3

3.1 Bayes Estimator Under Absolute Error Loss

$$L(\theta|\hat{\theta}) = |\theta - \hat{\theta}|$$

$$r(\hat{\theta}|x) = \int |\theta - \hat{\theta}| f(\theta|x) d\theta$$

$$= \int_{-\infty}^{\infty} |\theta - \hat{\theta}| f(\theta|x) d\theta$$

$$= \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta}) f(\theta|x) d\theta + \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta) f(\theta|x) d\theta$$

$$\int_{\hat{\theta}}^{\infty} \theta f(\theta|x) d\theta - \int_{\hat{\theta}}^{\infty} \hat{\theta} f(\theta|x) d\theta + \int_{-\infty}^{\hat{\theta}} \hat{\theta} f(\theta|x) d\theta - \int_{-\infty}^{\hat{\theta}} \theta f(\theta|x) d\theta$$

$$\int_{\hat{\theta}}^{\infty} \theta f(\theta|x)d\theta - \hat{\theta}[1 - F(\hat{\theta})] + \hat{\theta}F(\theta) - \int_{-\infty}^{\hat{\theta}} \theta f(\theta|x)d\theta$$

$$\frac{dr(\theta|x)}{d\hat{\theta}} = -\hat{\theta}f(\hat{\theta}|x) - 1 + F(\hat{\theta}|x) + \hat{\theta}f(\hat{\theta}|x) + F(\hat{\theta}|x) + \hat{\theta}f(\hat{\theta}|x) - \hat{\theta}f(\hat{\theta}|x)$$

$$= 0 = -1 + F(\hat{\theta}|x) + F(\hat{\theta}|x)$$

$$\frac{1}{2} = F(\hat{\theta}|x)$$

$$\therefore \hat{\theta} = median(f(\theta|x))$$

3.2 Bayes Estimator Under Zero-One Loss

$$L(\theta|\hat{\theta}) = \begin{cases} 0 \text{ if } \theta = \hat{\theta} \\ 1 \text{ if } \theta \neq \hat{\theta} \end{cases}$$

$$r(\hat{\theta}|x) = \int_{-\infty}^{\hat{\theta}} 1 * f(\theta|x) + \int_{\hat{\theta}}^{\infty} 1 * f(\theta|x)$$

$$= \lim_{z \to 0} [\int_{-\infty}^{\hat{\theta}-z} 1 * f(\theta|x) + \int_{\hat{\theta}+z}^{\infty} 1 * f(\theta|x)]$$

$$= \lim_{z \to 0} [1 - \int_{\hat{\theta}-z}^{\hat{\theta}+z} f(\theta|x)d\theta]$$

$$= 1 - \lim_{z \to 0} [\int_{\hat{\theta}-z}^{\hat{\theta}+z} f(\theta|x)d\theta]$$

$$= 1 - f(\hat{\theta}|x)$$

We know that when $f(\hat{\theta}|x)$ is maximized, $1 - f(\hat{\theta}|x)$ is minimized.

$$\therefore \hat{\theta} = Mode(f(\theta|x))$$