Bayes Tutorial - BIOL 470/670 (Dr. G.A. Babbitt – Rochester Institute of Technology)

The 'Bayesian paradigm' in statistics offers a more natural way of relating and adjusting models to observational evidence than the traditional 'frequentist' approaches you are already probably familiar with from Stats 101. Fundamentally, it assumes that we need to go beyond the consideration of the likelihood of an observation or set of data (i.e. evidence) under a given model and consider two more additional things; the prior likelihood of the hypothesis and the total or marginal likelihood of observing the evidence. The following video by Ian Olasov (CUNY) steps out the mechanics of Bayes rule and introduces the concept of base rate fallacy.

https://www.youtube.com/watch?v=OqmJhPQYRc8

1. Mathematically and verbally define the false assumption that is often made when people fall into the base rate fallacy.

The following video by Julia Galef takes Bayesian philosophy to a higher level and offers practical advice for ensuring rational critical thinking.

https://www.youtube.com/watch?v=BrK7X_XlGB8

2. Following Julia's examples, describe a potential application of Bayesian thinking to your own life or to a problem you are working on.

The next video by Derek Muller (at Veritasium) describes how base rate fallacy can relate to false positive detection rates in a medical test.

https://www.youtube.com/watch?v=R13BD8qKeTg

3. This video is a wonderful example of base rate fallacy, but makes an unrealistic assumption that an entire population is always tested for a given disease. In reality, a doctor will often use clinical evidence (i.e. educated opinion or lab test) to assign a diagnosis. However, this does not eliminate the problem of false positives and base rate fallacy. Explain, how you might adjust the mathematic terms of Bayes Rule to adjust for this reality that clinical observational evidence may correlate with the truth of the hypothesis.

In the next video Dr. Kristin Lennox gives a fun and lucid description of both frequentist and Bayesian probability. It's a longer video but will be crucial to our understanding of these concepts.

https://www.youtube.com/watch?v=eDMGDhyDxuY

4. Use the Bayes Calculator (I provided in MS Excel) to consider the simple case of two independent events A and B, both with probability thresholds of 0.5 and level of causation = 0. A physical model for this would be the flipping of two fair coins. What are the conditional probabilities of the outcomes of each coin given the other (i.e. p(A|B) and p(B|A)? Are they the same in both directions? What if both coins are similarly unfair (e.g. favoring heads)? What if both are dissimilar and unfair (e.g. one favors heads while other favors tails)?

5. Use the Bayes Calculator (I provided in MS Excel) to now consider a single fair coin (i.e. set the level of causation to 1). Now we can more abstractly apply Bayes in relation to a common frequentist scenario. Here, we hypothesize that the coin is fair by setting p(H) = 0.5 and the evidence we observed or p(E) will exactly match given the long-range outcome of random events or coin flips. What do you notice about the conditional probability or likelihood of observing E given that the coin is fair? Does the Bayesian perspective/calculation here inform us any more than that of a frequentist perspective? What two values on the calculator should you compare in order to answer this?

6. Use the Bayes Calculator (I provided in MS Excel) to determine if the curved coins or nontransitive dice I provided in class are truly fair. Flip or roll them many times and calculate a new frequency to enter as probability of evidence. Does your evidence provide much information regarding your null hypothesis that the coin or dice is fair? (see Bayes factor which is a ratio that represents how similar your evidence and hypothesis are)

7. Now use the calculator to represent a situation similar to that of question #3. Here, we will consider more 'real life' situation that will challenge the frequentist perspective (…a perspective that always requires a very well-defined physical model for determining probable outcomes). Let's assume that we have an imperfect relationship between a given hypothesis H and observed evidence E and that this creates uncertainty that is subjective rather than objective in its fundamental nature. This is where a Bayesian approach can truly inform us. The hypothesis will be that we have a specific disease and the evidence could take the form of a well-trained doctor's opinion or clinical lab test with a known rate of false positive detection. In a first example, let's assume that 9/1000 people in the general population will have the disease (set probability threshold of the hypothesis H = 0.009) and that the clinical evidence observed is higher, but still precise. So let's also assume that roughly 5% of patients present symptoms clinically that could be interpreted as indicating the disease among a variety of other problems (i.e. set probability threshold of evidence E = 0.05) and let's assume that the doctor is very good at spotting the disease correctly …let's say 70% correct (i.e. set level of causation to 0.7). Does the likelihood of the doctor observing symptoms of the disease given that you really do have the disease or p(E|H) match what you really care about…the probability that you have the disease given that the doctor has noticed your symptoms or p(H|E)? i.e. Does a base rate fallacy appear? What does this tell you mathematically? Be able to describe each term of the Bayes Rule equation in terms of this example.

8. Now let's consider the same scenario as above in a second example, but instead of a doctor's diagnosis with 70% accuracy, we have a clinical lab test that is 95% accurate (i.e. set probability threshold for evidence to 0.05 and level of causation to 0.95). Does the greater accuracy of the test eliminate base rate fallacy? i.e. is the probability of your having the disease given the positive test result or p(H|E) still much lower than the probability of a positive test result given that you have the disease or p(E|H)? Do you now see the value of the Bayesian paradigm? Explain in your own words.

9. Go to the Naïve Bayes Classifier (I provide in MS Excel). Here, we present a small dataset that records weather conditions and the subsequent behavior of a golfer. The Bayes classifier will predict whether the golfer will play or not play based upon the state of four weather conditions present in the data. Describe verbally and mathematically, in your own words, how the Bayes Rule is used for the problem of classification. Why is this procedure called 'naïve'? What assumption does it make that is not likely to be realistic, given what we already know from multivariate statistical analyses like PCA, multiple regression and factor analysis.

10. Which of the four weather conditions seem to influence the golfer the most…the least? Report the how the specific posterior probabilities change when this condition is altered. Now change the raw data to make the golfer more responsive to rainy days. Now report the how the specific posterior probabilities change when conditions are changed from sunny to rainy on a day with MILD temperature, normal humidity, and no wind.

11. Explain the general advantages and disadvantages of this Bayesian approach to machine learning classification, when compared to other more 'black box' methods like support vector machine and deep learning. What sorts of applications might it be ideal for?