DS 100: Principles and Techniques of Data Science **Date: February 2, 2018**

Discussion #2 Solutions

Name:

Bayes' Rule

1. **Are you smarter than a doctor? ONLY 46% OF DOCTORS GOT THIS QUESTION RIGHT.**

   100 out of 10,000 women at age forty who participate in routine screening have breast cancer. 80 of those 100 women with breast cancer test positive. 950 out of 9,900 women without breast cancer also test positive. If 10,000 women in this age group undergo a routine screening, about what fraction of these women with positive tests will actually have breast cancer?

   Solution: Always begin by figuring out what you want to know. In this case, we want to know what fraction (or percentage) of the women with positive tests actually have breast cancer.

   First, lets figure out how many women have positive tests. Thats the denominator of our fraction.

   The story above says that 950 of the 9,900 that do not have breast cancer will test positive. So thats 950 women with a positive test result right there.

   The story also says that 80 out of the 100 women who do have breast cancer will get a positive test result. So thats another 80 women, and 950 + 80 = 1,030 women with a positive test result.

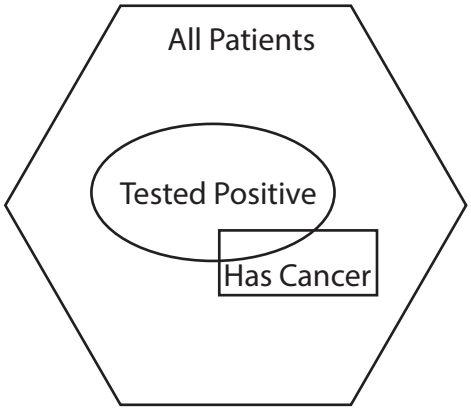
   Good. We've got half our fraction. Now, how do we find the numerator? How many of those 1,030 women with a positive test result actually have breast cancer?

   Well, the story says that 80 of the 100 women with breast cancer will get a positive test result, so 80 is our numerator.

   The fraction of women with positive test results who actually have breast cancer is 80/1,030, which is a probability of .078, which is 7.8%.

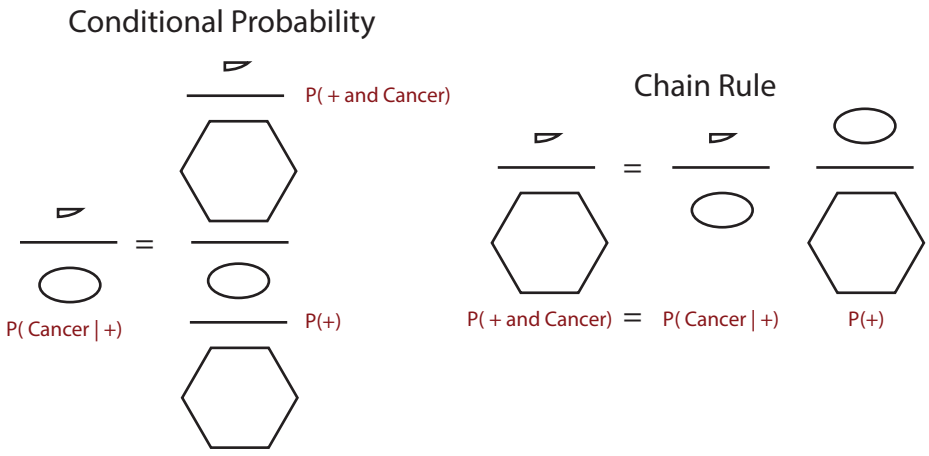
   So if one of these 40-year-old women tested positive, and the doctor knew the above statistics, then the doctor should tell the woman she has only a 7.8% chance of having breast cancer, even though she had a positive mammography. Thats much less stressful for the woman than if the doctor had told her she had a 70%-80% chance of having breast cancer like most doctors from 1976 apparently would!

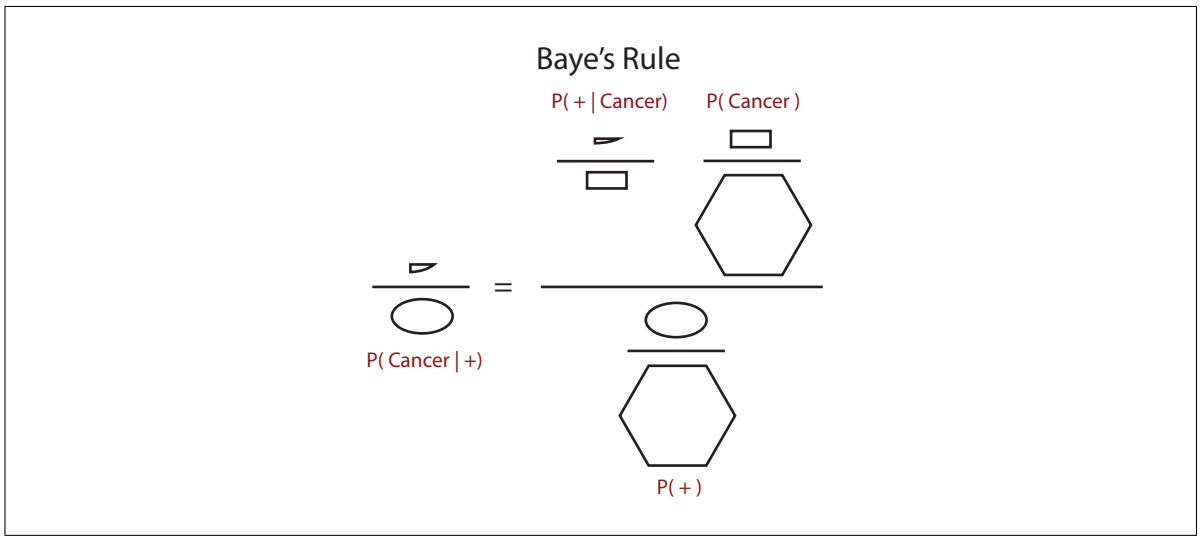
2. Consider the following Venn Diagram of the same problem.



Discuss as a class how we might represent conditional probabilities, chain rule, and Bayes' rule using the Venn Diagram. Use the space below for your notes.

Solution:




Baye's Rule
P(+ | Cancer)
P(Cancer)
=
P(Cancer | +)
P(+)

3. **HEALTH PROFESSIONALS HATE THIS.**

Only 1% of women at age forty who participate in a routine mammography test have breast cancer. 80% of women who have breast cancer will test positive, but 9.6% of women of women who dont have breast cancer will also get positive tests. A woman of this age tested positive in a routine screening. What is the probability that she actually has breast cancer?

Solution: Again, we want to know what percentage of women with positive tests actually have breast cancer.

We can directly apply Baye's rule:

$$\mathbb{P}\left(\text{Has Cancer} \mid \text{Tested Positive}\right) = \frac{\mathbb{P}\left(\text{Tested Positive} \mid \text{Has Cancer}\right) \times \mathbb{P}\left(\text{Has Cancer}\right)}{\mathbb{P}\left(\text{Tested Positive}\right)}$$

We'll split the denominator into two mutually exclusive cases: a) testing positive with cancer and b) testing positive with no cancer. Then we'll use some basic probability rules to get the values we actually have.

$$\begin{aligned}
\mathbb{P}\left(\text{Tested Positive}\right) &= \mathbb{P}\left(\text{Tested Positive AND Cancer}\right) + \mathbb{P}\left(\text{Tested Positive AND No Cancer}\right) \\
&= \mathbb{P}\left(\text{Tested Positive} \mid \text{Cancer}\right)\mathbb{P}\left(\text{Cancer}\right) \\
&\quad + \mathbb{P}\left(\text{Tested Positive} \mid \text{No Cancer}\right)\mathbb{P}\left(\text{No Cancer}\right) \\
&= \mathbb{P}\left(\text{Tested Positive} \mid \text{Cancer}\right)\mathbb{P}\left(\text{Cancer}\right) \\
&\quad + \mathbb{P}\left(\text{Tested Positive} \mid \text{No Cancer}\right)\left(1 - \mathbb{P}\left(\text{Cancer}\right)\right)
\end{aligned}$$

So ultimately we have the form:

$$\frac{\mathbb{P}\left(\text{Tested Positive} \mid \text{Has Cancer}\right) \times \mathbb{P}\left(\text{Has Cancer}\right)}{\mathbb{P}\left(\text{Tested Positive} \mid \text{Cancer}\right)\mathbb{P}\left(\text{Cancer}\right) + \mathbb{P}\left(\text{Tested Positive} \mid \text{No Cancer}\right)\left(1 - \mathbb{P}\left(\text{Cancer}\right)\right)}$$

Into which we can plug in the values to get our answer:

$$\mathbb{P}\left(\text{Has Cancer} \mid \text{Tested Positive}\right) = \frac{0.80 \times 0.01}{0.80 \times 0.01 + 0.096 \times (1 - 0.01)} \approx 0.078$$

So the patient has a 7.8% chance of having cancer.

Data Visualization and Scope

This part of the discussion will be centered on this video

https://tinyurl.com/data100-rosling

4. Answer the following questions about the quality of the visualization in the video

   (a) What variables are being represented in the graphic?

   Solution: Income, Life Expectancy, Time, Population, Country region, Country

   (b) How are the variables being represented visually?

   Solution: In order to prepare our mindsets for programmatically creating visualizations of the data, we should begin thinking of plots as a mapping from data onto a visual property. In this particular example, we have:

   1. Income $\rightarrow$ horizontal location (x)
   2. Life expectancy $\rightarrow$ vertical location (y)
   3. Time $\rightarrow$ Text/Plot frame
   4. Population $\rightarrow$ Circle size
   5. Country region $\rightarrow$ circle color
   6. Country $\rightarrow$ Text label

   (c) How do we interpret the visual qualities? In other words, how can we look at the image and know how to interpret the properties of the plot into data?

   Solution:

   1. Axes with scales are given for GDP per capita and life expectancy.
   2. There is background text for each year as the video plays and labels for certain countries of interest.
   3. We are told by Rosling how to interpret the color and sizes of the circles

   (d) Does it look like the raw values of the data were plotted or were they (numerically) transformed before plotting?

Solution:

1. Income has been hit by a log-transformation. We see the absolute values on the axes, though.
2. Life expectancy has been centered around the global average (axis does not start at 0)
3. The year progression is slowed/sped up to emphasize certain points in history
4. Population was scaled so that the radius of the circle is the square root of the population count (hard to tell!)

(e) Is there any information present that is not represented visually?

Solution: Rosling's narration! He gives a selective account of the historical **context** of the data. Remember that each and every plot you present to others should tell a story of some sort.

(f) How good do you think the data are for this visualization?

Solution: It paints the general trend fairly well, but we should expect some issues in scope.

The data are compiled from a variety of sources. Data from high-income countries are mainly derived from registers, whereas surveys are a common source in low and middle-income countries. Such surveys are based on interviews with a representative sample of the population. Data for the 19th century is often based on various types of estimates.

Many countries had different borders or did not exist at all in the past. The data concerns the area of the present day borders of the country.

Go to 1:30. Point out the some countries move a lot more in this period than others. This is because those countries' were actually collecting finer data at the time. For every other country, we have to make estimates!