Exercise: Medical Diagnosis with Costs

(7/7 points)

In this exercise, we continue off our running medical diagnosis example, now looking at a decision variant of it where we need to decide whether to treat a patient, and there are associated costs.

We will be using both conditional expectation and the expectation of the function of a random variable. In fact, we need to combine the two! For random variables X and Y for which we know (or have computed) the conditional distribution $p_{X|Y}(\cdot \mid y)$, and for any function f such that $f(x)$ is a real number for every x in the alphabet of X, then

$$\mathbb{E}[f(X) \mid Y = y] = \sum_x f(x) p_{X|Y}(x \mid y).$$

(See if you can figure out why this is true from our earlier coverage of expectation!)

We previously computed the probability that someone has a infection given that a test for the infection turned out positive. But how do we translate this into a decision of whether or not we should treat the person? Ideally, we should always choose to treat people who, in truth, have the infection and choose “don't treat" for people without the infection. However, it is possible that we may treat actually healthy people, a *false positive* that leads to unnecessary medical costs. As well, we may fail to treat people with the infection, a *false negative* that could require more expensive treatment later on.

We can encode these costs into the following “cost table" (or cost function) $C(\cdot, \cdot)$, where $C(x, d)$ is the cost when the patient's true infection state is x (“healthy" or “infected"), and we make decision d (“treat" or “don't treat"). For our problem, here is the cost function:

Truth x	Decision d	$C(x, d)$
healthy	don't treat	\$0
healthy	treat	\$20000
infected	don't treat	\$50000
infected	treat	\$20000

Exercises due Oct 12, 2016 at 21:00 UTC

Measuring Randomness (Week 4)
Exercises due Oct 12, 2016 at 21:00 UTC

Towards Infinity in Modeling Uncertainty (Week 4)
Exercises due Oct 12, 2016 at 21:00 UTC

Homework 3 (Week 4)
Homework due Oct 12, 2016 at 21:00 UTC

This cost table corresponds to a situation where the treatment costs $20000, regardless of whether the patient's true state is healthy or infected. If the “don't treat" choice is made, then there are two potential costs: $0 if the patient is indeed healthy, or $50000 if the patient actually has the infection, perhaps corresponding to a more expensive future medical procedure needed if the infection is not treated now.

Recall the medical diagnosis setup from before, summarized in these tables:

		Prob.
X	healthy	0.999
	infected	0.001

$p_{Y\mid X}$		X	
		healthy	infected
Y	positive	0.01	0.99
	negative	0.99	0.01

Let's suppose we've observed $Y = \textbf{positive}$. To use the cost table to help us make a decision, note that there are two possible realizations of the world given that $Y = \textbf{positive}$:

- First, it could be that $X = \textbf{infected}$ given $Y = \textbf{positive}$. This realization has probability $p_{X|Y}(\textbf{infected} \mid \textbf{positive})$. Under this realization of the world, by looking at the last two rows in the above cost table:
  - If we decide $d = \textbf{treat}$, the cost is $20000
  - If we decide $d = \textbf{don't treat}$, the cost is $50000
- Second, it could be that $X = \textbf{healthy}$ given $Y = \textbf{positive}$. This realization has probability $p_{X|Y}(\textbf{healthy} \mid \textbf{positive}) = 1 - p_{X|Y}(\textbf{infected} \mid \textbf{positive})$. Under this realization of the world, by looking at the two first rows in the cost table:
  - If we decide $d = \textbf{treat}$, the cost is $20000

- If we decide $d = \textbf{don't treat}$, the cost is $0

Thus, putting together the pieces:

- If we decide $d = \textbf{don't treat}$, then the cost is a random variable and is given by the probability table $p_{C(X,\textbf{don't treat})|Y}(\cdot|\textbf{positive})$ below:

		Probability
Cost	$50000	$p_{X\mid Y}(\text{infected} \mid \text{positive})$
	$0	$p_{X\mid Y}(\text{healthy} \mid \text{positive})$

What is $\mathbb{E}[C(X, \textbf{don't treat}) \mid Y = \textbf{positive}]$?

Express your answer as a function of $a \triangleq p_{X|Y}(\textbf{infected} \mid \textbf{positive})$ (unless a does not appear in the answer). Do *not* plug in the numerical value for a that you determined in a previous exercise, and also do *not* put a dollar sign in your answer.

In this part, please provide your answer as a mathematical formula (and not as Python code). Use ^ for exponentiation, e.g., x^2 denotes x^2. Explicitly include multiplication using *, e.g. x*y is xy.

50000 * a

✔ **Answer:** 50000*a

$50000 \cdot a$

- If we decide $d = \textbf{treat}$, then the cost is a random variable and is given by the probability table $p_{C(X,\textbf{treat})|Y}(\cdot|\textbf{positive})$.

What is $\mathbb{E}[C(X, \textbf{treat}) \mid Y = \textbf{positive}]$? (Hint: Figure out what the probability table $p_{C(X,\textbf{treat})|Y}(\cdot|\textbf{positive})$ is first. See if you can figure out how we obtained the probability table $p_{C(X,\textbf{don't treat})|Y}(\cdot|\textbf{positive})$ in the previous part.)

Express your answer as a function of $a \triangleq p_{X|Y}(\textbf{infected} \mid \textbf{positive})$ (unless a does not appear in the answer). Do *not* plug in the numerical value for a that you determined in a previous exercise. Express your answer as a function of a, where $a \triangleq p_{X|Y}(\textbf{infected} \mid \textbf{positive})$. Do *not* plug in the numerical value for a that you determined in a previous exercise, and also do *not* put a dollar sign in your answer.

20000 * a + 20000 * (1-a)

✔ **Answer:** 20000

$20000 \cdot a + 20000 \cdot (1 - a)$

- Given that $Y = \textbf{positive}$, to choose between which of the two decisions $d = \textbf{healthy}$ and $d = \textbf{infected}$ to make, we can choose whichever decision d that has the lower expected cost $\mathbb{E}[C(X, d) \mid Y = \textbf{positive}]$ (note that it could be that the expected costs are the same, in which case we could just break the tie arbitrarily).

Look at your answers to the previous two parts. You should notice that "don't treat" has the lowest possible cost when a is at most some threshold t, which you will now determine.

Phrased in terms of an equation, with variable a defined as before ($a \triangleq p_{X|Y}(\text{infected} \mid \text{positive})$), there is a threshold t for which

$$\arg \min_{d \in \{\text{treat}, \text{don't treat}\}} \mathbb{E}[C(X, d) \mid Y = \text{positive}] = \begin{cases} \text{don't treat} & \text{if } a \le t, \\ \text{treat} & \text{if } a > t, \end{cases}$$

where we are breaking the tie in favor of not treating when $a = t$.

Determine what the threshold t is. (Please provide an **exact** answer.)

$t =$ 0.4 ✔ **Answer:** 2/5

Let's look at what happens if instead we had observed $y = \text{negative}$.

- Determine what variable b and threshold s should be in the below equation:

$$\arg \min_{d \in \{\text{treat}, \text{don't treat}\}} \mathbb{E}[C(X, d) \mid Y = \text{negative}] = \begin{cases} \text{don't treat} & \text{if } b \le s, \\ \text{treat} & \text{if } b > s. \end{cases}$$

$b =$

- ○ $p_{X|Y}(\text{healthy} \mid \text{positive})$
- ○ $p_{X|Y}(\text{infected} \mid \text{positive})$
- ○ $p_{X|Y}(\text{healthy} \mid \text{negative})$
- ◉ $p_{X|Y}(\text{infected} \mid \text{negative})$ ✔

$s =$ 0.4 ✔ **Answer:** 2/5

Putting together the pieces, the decision rule that achieves the lowest cost can be written as:

$$\widehat{x}(y) = \begin{cases} \text{don't treat} & y = \text{positive and } a \le t, \\ \text{treat} & y = \text{positive and } a > t, \\ \text{don't treat} & y = \text{negative and } b \le s, \\ \text{treat} & y = \text{negative and } b > s, \end{cases}$$

where a, b, t, and s are from the previous parts.

Now finally consider when $a = 0.09016$ and $b = 0.00001$ (these correspond to values from the previous exercises on this medical diagnosis problem setup). What is the optimal treatment plan given the cost table that we have introduced?

- If we observe that $Y = \text{positive}$, then the decision with lowest expected cost is:

  - ○ treat
  - ◉ don't treat ✔

- If we observe that $Y = \text{negative}$, then the decision with lowest expected cost is:

  - ○ treat
  - ◉ don't treat ✔

Of course, if a and b had been different, then the decisions we would make to minimize expected cost could change! In fact, this problem shows just how good the medical test needs to be designed before we can even justify treating the patient. A bad test would be one where the optimal decision for minimizing expected cost does not use the result of the test at all — we might as well not do the test!

Solution:

- What is $\mathbb{E}[C(X, \text{don't treat}) \mid Y = \text{positive}]$?

  Solution: By looking at the probability table for $C(X, \text{don't treat})$ given that $Y = \text{positive}$, we see that \$50000 has probability a and \$0 has probability $1 - a$. Thus, the conditional expectation

  $$\mathbb{E}[C(X, \text{don't treat}) \mid Y = \text{positive}] = 50000 \cdot a + 0 \cdot (1 - a) = \boxed{50000 \cdot a}$$

- What is $\mathbb{E}[C(X, \text{treat}) \mid Y = \text{positive}]$?

  Solution: If you look at the cost table specifically restricted to when we decide to treat, i.e., $C(\cdot, \text{treat})$, the cost is always \$20000 for treating. Thus, it does not even matter what the conditional probability of X given $Y = \text{positive}$ is here: the cost is always \$20000 so the expected cost is likewise **20000** dollars.

- Determine what the threshold t is.

  Note that “don't treat" has cost $50000 \cdot a$ and “treat" has cost 20000. Thus, “don't treat" has the lowest cost when $50000 \cdot a \leq 20000$, i.e., $a \leq 2/5$, so $t = 2/5$.

- Determine what variable b and threshold s should be in the below equation:

$$\arg \min_{d \in \{\text{treat}, \text{don't treat}\}} \mathbb{E}[C(X,d) \mid Y = \text{negative}] = \begin{cases} \text{don't treat} & \text{if } b \le s, \\ \text{treat} & \text{if } b > s. \end{cases}$$

Everything is still the same as before where the only change is that instead of writing an answer in terms of $a = p_{X|Y}(\text{infected} \mid \text{positive})$, we now just need to condition on $Y = \text{negative}$ instead, so $\boxed{b = p_{X|Y}(\text{infected} \mid \text{negative})}$. In particular, the threshold s is still the same as before, i.e., $\boxed{s = t = 2/5}$.

Now finally consider when $a = 0.09016$ and $b = 0.00001$ (these correspond to values from the previous exercises on this medical diagnosis problem setup).

- **Solution:** If we observe that $Y = \text{positive}$, then the decision with lowest expected cost is **don't treat** because $a = 0.09016$ is less than $t = 2/5$.
- **Solution:** If we observe that $Y = \text{negative}$, then the decision with lowest expected cost is **don't treat** because $b = 0.00001$ is less than $s = 2/5$.

You have used 3 of 5 submissions