

CS-473: Pattern Recognition

Prof. Panos Trahanias, trahania@csd.uoc.gr
T.A. Emmanouil Sylligardos, sylligardos@csd.uoc.gr

Note: This assignment should be implemented entirely in Google Colaboratory. Google's notebook allows you to combine executable Python scripts with rich text in a single document. Your deliverable should be a single .ipynb file along with its corresponding .py file (both can be easily exported from Google Colaboratory). **Every single question should be implemented in a single code block or text block. Code blocks should be clearly and shortly explained (you may use the text boxes for that goal). For questions that only require a written answer, write it in a text block. Use text boxes with big letters to clearly distinguish between exercises, questions, steps, etc. Before you submit the exercise make sure that you carefully read and follow the notes/instructions at the end of this pdf.**

Exercise 1 (40%): Medical Test Paradox

You are a medicine student working at your local hospital. A few days ago a person took a test for a rare disease which turned out to be positive and your boss asked you to work out the probability of that person actually having the disease. You quickly went through the description of the aforementioned test and found out that its sensitivity is 98% and its specificity is 95%. Before notifying the doctor about the terrible results you were reminded of the Bayes Theorem from your Pattern Recognition course and decided to do some extra research.

According to the Bayes theorem, the probability of that person having the disease given that it tested positive is the following:

$$P(\text{disease} \mid \text{test} +) = \frac{P(\text{test}+ \mid \text{disease}) * P(\text{disease})}{P(\text{test}+)}$$

In simple words: the probability of having the disease given that you tested positive is the probability of testing positive given that you have the disease (the likelihood),

multiplied by the overall probability of having the disease (the prior), divided by the general frequency of positives tests (the evidence), which itself is given by:

$$P(\text{test} +) = P(\text{disease}) * P(\text{test} + | \text{disease}) + P(! \text{disease}) * P(\text{test} + | ! \text{disease})$$

You go through some research regarding the disease and you find out the following numbers:

1. The sensitivity of the test is 98%.
2. The specificity of the test is 95%.
3. The known total cases in 2021 were approximately 350,000 worldwide.
4. The world population in 2021 is estimated to be $7,9 * 10^9$ people.

(Hint: What you need to calculate the evidence is the True-Positive rate and the False-Positive rate. Both numbers are given to you directly or indirectly. You can read this wikipedia article for further understanding [Sensitivity and specificity - Wikipedia](#))

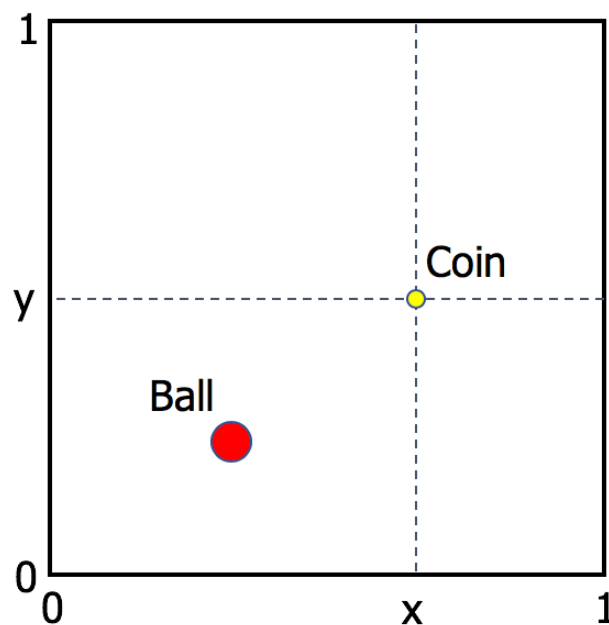
Questions:

1. Write a function in python that given the sensitivity, the specificity of a test, and the prior probability of having a disease it returns the posterior probability of having that disease given that you tested positive.
2. Use the function you implemented to calculate that given the statistics above, what is the probability of that person having the disease? Print the result as a percentage.
3. Having calculated the probability of that person having the rare disease you decided to tell the doctor that another test should be carried out since there is not enough evidence to come to a conclusion. Another test is being done to that person and again it comes out positive. After another positive test calculate and print (as a percentage) the posterior probability of that person having the disease.
4. Calculate step 2 again given that the total cases in 2021 were approximately:
 - a. 1 million people
 - b. 10 million people
 - c. 1 billion people
5. In 1-2 sentences write your thoughts on how and why the prior affects the posterior probability of actually having the disease given that you tested positive.

Exercise 2 (60%): “In the footsteps of Bayes” (Visualizing Bayesian updating)

Bayes’ theorem was only discovered a decade after his death by his friend Richard Price who had been asked to go through his estate looking for publishable material. Among his essays was a now famous thought experiment:

Bayes would sit with his back to a table. His assistant would place a coin randomly on the table without revealing its position to Bayes. He would then uniformly at random throw balls on the table, notifying Bayes only of whether the ball landed left of/right of/in front of/behind the coin:



With each new piece of evidence, Bayes was able to update his belief about the coin’s position and estimate it with greater and greater accuracy.

- Assume a 1m x 1m table where the x position extends horizontally and the y position extends vertically.
- A ball is said to land left of the coin if its x position is less than that of the coin, right of it otherwise.
- A ball is said to land in front of the coin if its y position is less than that of the coin, behind it otherwise.
- Assume Bayes’ assistant threw $N = 30$ balls and noted the following final observations:

- L = 12 balls landed left of the coin (and thus 18 landed right of it)
- F = 8 balls landed in front of the coin (and thus 22 landed behind it)

We would like to calculate the posterior probability of the coin lying at each possible point on the table.

Questions:

1. Write a python function to calculate the posterior probability of the coin being located at (x, y) given the observation {L, F, N}.

$$P(x, y | L, F, N) = P(x | L, F, N) * P(y | L, F, N) = P(x | L, N) * P(y | F, N) = ?$$





(Hint: The last equality comes from the fact that the coin's x-coordinate is independent of F and vice versa.)

- a. Given a single random toss, what is the probability of the ball landing left of/ right of/ in front of/ behind the coin? ($0 \leq x \leq 1$ Hint: The greater the coin's x-coordinate, the greater the probability of the ball landing left of the coin - Try it yourself but highlight to reveal the first solution if you need to).
 - i. $P(\text{left} | x, y) = P(\text{left} | x) =$
 - ii. $P(\text{right} | x, y) = P(\text{right} | x) = ?$
 - iii. $P(\text{front} | x, y) = P(\text{front} | y) = ?$
 - iv. $P(\text{behind} | x, y) = P(\text{behind} | y) = ?$
 - b. What is the probability of the ball landing left of the coin 5 times in a row? (Hint: individual ball tosses are completely independent events)
 - i. $P(L = 5, N = 5 | x, y) = P(L = 5, N = 5 | x) = ?$
 - c. What is the probability that 3 out of 10 balls land left of the coin? (Hint: It is the probability of the ball landing left 3 times multiplied by the probability landing right 7 times multiplied by the total number of possible orders in which 3 'lefts' and 7 'rights' can occur - for the last term *use the binomial coefficient*).
 - i. $P(L = 3, N = 10 | x, y) = P(L = 3, N = 10 | x) = ?$
 - d. If you have completed all the previous small steps, you should be equipped with all the knowledge required to calculate $P(x, y | L, F, N)$ using Bayes Theorem. (Hint: The terms $P(x)$ and $P(y)$, that is the prior probabilities of the points of the table are the same among the points and can thus be ignored, since they would only scale the overall distribution. Additionally, $P(L, N)$ and $P(F, N)$, namely the evidence, are constants and can thus be ignored).
2. Using the values of L, F, N given to you and calculate the posterior probabilities of the points of the table using the function you implemented. Visualize the results on a 3D plot. You could make use of the following functions:

- a. Similarly to the first assignment, the points that represent the table should be made using the “np.linspace” function. The table should be a 100 x 100 matrix, and you should feed that matrix to the function you implemented (be careful, the shape of the matrix should change to 10000 x 2 prior feeding it to the function and back to 100 x 100 prior to plotting the results).
 - b. In order to create the 3D plot you could use the functions:
 - i. “ plt.gca(projection='3d') ” which will return the axis of a 3D figure
 - ii. “ ax.plot_surface ” using the axis the previous function returned
3. Lastly, we would like to visualize and understand how our belief about the coin's position changes with each new observation, similar to Bayes' original thought experiment. For this purpose you are provided with the files fronts.csv and lefts.csv (Bonus 10%).
- a. Load the csv files given to you.
 - b. Use the following libraries and commands to create an animated 3D plot of 30 frames, that is 1 frame for each new observation (the first elements of the files are the first observation, the second elements are the second observation, etc...). Each frame should be a 3D plot (as in the previous step) of the posterior probabilities of each point of the table after each observation.

```
c. from matplotlib import rc
d. rc('animation', html='jshtml')
e.
f. from mpl_toolkits import mplot3d
g. import matplotlib.animation as animation
```

Additional reading:

-  How To Update Your Beliefs Systematically - Bayes' Theorem
-  Bayes theorem, the geometry of changing beliefs
-  The quick proof of Bayes' theorem
-  The medical test paradox, and redesigning Bayes' rule

Important notes:

- To ask questions send an email to hy473-list@csd.uoc.gr with the **subject** “[CS473]: Assignment 1 question”.
- To submit your implementation send an email to sylligardos@csd.uoc.gr with the **subject** “[CS473]: Assignment 1 submission”.
- You should submit only the files .ipynb and .py in a zipped folder (.zip) with the **name** “hw2_<am>” where am is your university identification number. The folder should contain **nothing else** than the two aforementioned files.
- The **names** of the submitted .ipynb and .py should be “hw2_<am>.ipynb” and “hw2_<am>.py” respectively.
- The whole .ipynb file should run when selecting ‘Runtime -> Run all’ without any problem. Meaning that any unnecessary code blocks should be removed prior to submitting. **Code blocks that can not run will not be graded. If they prevent the smooth execution of the file they will have a negative impact on the grade of the assignment.**

Good luck!