Homework 2

Yuan Li, N19728558, NetID: yl6606

Part I: Written Exercises

1. (a) Answer:

Set x as the method 1's result, y as the method 2's result, and m as the person's true result. $+$ represents the positive result, and $-$ represents the negative result. Based on the MAP hypothesis, we can use $arg\ maxp(m|x^+, y^+) = arg\ maxp(x^+, y^+|m)p(m)$ to get the answer. Now calculate these respectively.

$$\begin{aligned} p(x^+, y^+|m^+)p(m^+) &= (1 - 20\%) * (1 - 7\%) * 0.01\% \\ &= 0.00744\% \end{aligned}$$

$$\begin{aligned} p(x^+, y^+|m^-)p(m^-) &= 20\% * 7\% * (1 - 0.01\%) \\ &= 1.39986\% \end{aligned}$$

$p(x^+, y^+|m)p(m^+) < p(x^+, y^+|m)p(m^-)$, so the MAP hypothesis is **"The person doesn't have the disease".**

1. (b) Answer:

For this question, we only need to care about $arg\ maxp(x^+, y^+|m)$.

$$\begin{aligned} p(x^+, y^+|m^+) &= (1 - 20\%) * (1 - 7\%) \\ &= 74.4\% \end{aligned}$$

$$\begin{aligned} p(x^+, y^+|m^-)p(m^-) &= 20\% * 7\% \\ &= 1.4\% \end{aligned}$$

$p(x^+, y^+|m)p(m^+) > p(x^+, y^+|m)p(m^-)$, so the ML hypothesis is **"The person has the disease".**

1. (c) Answer:

Because the results of the two screening methods are independent, we can use chain rule to calculate the probability:

$$\begin{aligned} p(pos1, pos2, disease) &= p(pos1|pos2, disease) * p(pos2|disease) * p(disease) \\ &= p(pos1|disease) * p(pos2|disease) * p(disease) \\ &= (1 - 20\%) * (1 - 7\%) * 0.01\% \\ &= 0.00776\% \end{aligned}$$

2. (a) Answer:

We can use Python to calculate these easily.

```
x2 = [40, 51, -1, 2, 26, 41]
new_x2 = [(i - min(x2))/(max(x2)-min(x2)) for i in x2]
```

Python code

So the result is $[0.79, 1.0, 0.0, 0.06, 0.52, 0.81]$.

2. (b) Answer:

Firstly, scale the new example $x = \begin{bmatrix} 3.9 \\ 4 \end{bmatrix}$ to $x = \begin{bmatrix} 1.02 \\ 0.10 \end{bmatrix}$.
Then calculate distance between x and each point in dataset. Set point in dataset as $d_i (i = 1, 2, ..., 6)$.

$$\begin{aligned} dis &= ||x - d_i|| \\ &= \sqrt{(x'_1 - x_{i1})^2 + (x'_2 - x_{i2})^2} \\ &= \begin{bmatrix} 0.767 \\ 0.900 \\ 1.025 \\ 0.777 \\ 0.697 \\ 0.806 \end{bmatrix} \end{aligned}$$

x is closest to the 5th point in dataset. So the label for the example is **"-"**

3. (a) Answer:

3. (b) Answer:

3. (c) Answer:

4. (a) Answer:

$$\begin{aligned} P(x_1 = Low|+) &= \frac{2 + 0.2}{3 + 0.2 * 3} = \frac{11}{18} \\ P(x_2 = Yes|+) &= \frac{0.2}{3 + 0.2 * 3} = \frac{1}{18} \\ P(x_3 = Green|+) &= \frac{2 + 0.2}{3 + 0.2 * 3} = \frac{11}{18} \\ P(x_1 = Low|-) &= \frac{1 + 0.2}{4 + 0.2 * 4} = \frac{1}{4} \\ P(x_2 = Yes|-) &= \frac{3 + 0.2}{4 + 0.2 * 4} = \frac{2}{3} \\ P(x_3 = Green|-) &= \frac{3 + 0.2}{4 + 0.2 * 4} = \frac{2}{3} \end{aligned}$$

4. (b) Answer:

$$\begin{aligned} P(x_1 = Low, Yes, Green|+) &= P(x_1 = Low|+) * P(x_2 = Yes|+) * P(x_3 = Green|+) \\ &= \frac{11}{18} * \frac{1}{18} * \frac{11}{18} \\ &\approx 0.0209 \end{aligned}$$

$$\begin{aligned} P(x_1 = Low, Yes, Green|-) &= P(x_1 = Low|-) * P(x_2 = Yes|-) * P(x_3 = Green|-) \\ &= \frac{1}{4} * \frac{2}{3} * \frac{2}{3} \\ &\approx 0.1112 \end{aligned}$$

4. (c) Answer:

Because $P(x_1 = Low, Yes, Green|+) < P(x_1 = Low, Yes, Green|-)$, the ML label should be **"-"**.

4. (d) Answer:

$$P(x_1 = Low, Yes, Green|+) * P(+) \approx 0.0089$$
$$P(x_1 = Low, Yes, Green|-) * P(-) \approx 0.0635$$

Because $P(x_1 = Low, Yes, Green|+) * P(+) < P(x_1 = Low, Yes, Green|-) * P(-)$, the MAP label should be **"-"**.