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 $arg\ maxp(m|x^+)$ to get the answer. Now calculate these respectively.

$$\begin{aligned}
p(m^+|x^+) &= \frac{p(x^+|m^+)p(m^+)}{p(x^+)} \\
&= \frac{p(x^+|m^+)p(m^+)}{p(x^+|m^+)p(m^+) + p(x^+|m^-)p(m^-)} \\
&= \frac{(1-15\%)*0.01\%}{(1-15\%)*0.01\% + 20\%*(1-0.01\%)} \\
&= 0.00042486
\end{aligned}$$

$$\begin{aligned}
p(m^-|x^+) &= \frac{p(x^+|m^-)p(m^-)}{p(x^+)} \\
&= \frac{p(x^+|m^-)p(m^-)}{p(x^+|m^+)p(m^+) + p(x^+|m^-)p(m^-)} \\
&= \frac{20\%*(1-0.01\%)}{(1-15\%)*0.01\% + 20\%*(1-0.01\%)} \\
&= 0.999575
\end{aligned}$$

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

1. (b) Answer:

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

$$\begin{aligned}
p(x^+|m^+) &= (1-15\%) \\
&= 0.85 \\
p(x^+|m^-) &= 0.2
\end{aligned}$$

$p(x^+|m^+) > p(x^+|m^-)$, 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:

$$\begin{aligned}
p(m^+|x^+,y^+) &= \frac{p(x^+,y^+|m^+)p(m^+)}{p(x^+,y^+)} \\
&= \frac{p(x^+,y^+|m^+)p(m^+)}{p(x^+,y^+|m^+)p(m^+) + p(x^+,y^+|m^-)p(m^-)} \\
&= \frac{(1-0.15)*(1-0.04)*0.0001}{(1-0.15)*(1-0.04)*0.0001 + 0.2*0.07*(1-0.0001)} \\
&= 0.00579537
\end{aligned}$$

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:

Set μ_+, Σ_+ as the mean and covariance matrix for label "+", μ_-, Σ_- as the mean and covariance matrix for label "-".

$$
\begin{aligned}
\mu_+ &= \frac{\sum_{t=1}^{N} x^t}{N} \\
&= [\frac{\sum_{t=1}^{3} x_1^t}{3}, \frac{\sum_{t=1}^{3} x_2^t}{3}]^T \\
&= \begin{bmatrix} 1.83 \\ 3.20 \end{bmatrix} \\
\mu_- &= \frac{\sum_{t=1}^{N} x^t}{N} \\
&= [\frac{\sum_{t=1}^{4} x_1^t}{4}, \frac{\sum_{t=1}^{4} x_2^t}{4}]^T \\
&= \begin{bmatrix} 1.50 \\ 2.53 \end{bmatrix} \\
\Sigma_+ &= \frac{\sum_{t=1}^{N} (x^t - \mu_+)(x^t - \mu_+)^T}{N} \\
&= \begin{bmatrix} 2.555 & 3.933 \\ 3.933 & 6.140 \end{bmatrix} \\
\Sigma_- &= \frac{\sum_{t=1}^{N} (x^t - \mu_-)(x^t - \mu_-)^T}{N} \\
&= \begin{bmatrix} 0.420 & 0.990 \\ 0.990 & 2.442 \end{bmatrix}
\end{aligned}
$$

3. (b) Answer:

Using QDA, we can calculate $p(x|\mu,\Sigma)$ to determinate the class of the example x. To simplify the process, try $log(p(x|\mu,\Sigma))$.

$$\begin{aligned}
log(p(x|\mu_+,\Sigma_+)) &= log(\frac{1}{2\pi|\Sigma_+|^{1/2}}exp(-\frac{1}{2}(x-\mu_+)^T\Sigma_+^{-1}(x-\mu_+))) \\
&= -log2\pi - \frac{1}{2}log(|\Sigma_+|) - \frac{1}{2}(x-\mu_+)^T\Sigma_+^{-1}(x-\mu_+) \\
&= -157.5806 \\
log(p(x|\mu_-,\Sigma_-)) &= log(\frac{1}{2\pi|\Sigma_-|^{1/2}}exp(-\frac{1}{2}(x-\mu_-)^T\Sigma_+^{-1}(x-\mu_-))) \\
&= -100.9198
\end{aligned}$$

$p(x|\mu_+,\Sigma_+) < p(x|\mu_-,\Sigma_-)$, so the class should be **"-"**.

3. (c) Answer:

If we estimate a single covariance matrix using the entire dataset, it becomes a Linear Discriminant Analysis. The decision boundary should be a linear. It's easier to calculate, however, this answer is not accurate enough. We may think the distribution should be different between each class, so the covariance matrix should be different.

4. (a) Answer:

$$\begin{aligned}
P(x_1 = Low|+) &= \frac{2+0.2}{3+0.2*3} = \frac{11}{18} \\
P(x_2 = Yes|+) &= \frac{0.2}{3+0.2*2} = \frac{1}{17} \\
P(x_3 = Green|+) &= \frac{2+0.2}{3+0.2*2} = \frac{11}{17} \\
P(x_1 = Low|-) &= \frac{1+0.2}{4+0.2*3} = \frac{6}{23} \\
P(x_2 = Yes|-) &= \frac{3+0.2}{4+0.2*2} = \frac{8}{11} \\
P(x_3 = Green|-) &= \frac{3+0.2}{4+0.2*2} = \frac{8}{11}
\end{aligned}$$

4. (b) Answer:

$$\begin{aligned}
P(x_1 = Low, Yes, Green|+) &= P(x_1 = Low|+) * P(x_2 = Yes|+) * P(x_3 = Green|+) \\
&= \frac{11}{18} * \frac{1}{17} * \frac{11}{17} \\
&\approx 0.02326
\end{aligned}$$

$$\begin{aligned}
P(x_1 = Low, Yes, Green|-) &= P(x_1 = Low|-) * P(x_2 = Yes|-) * P(x_3 = Green|-) \\
&= \frac{6}{23} * \frac{8}{11} * \frac{8}{11} \\
&\approx 0.13798
\end{aligned}$$

4. (c) Answer:

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

4. (d) Answer:

$$\begin{aligned}
P(x_1 = Low, Yes, Green|+) * P(+) \approx 0.00997 \\
P(x_1 = Low, Yes, Green|-) * P(-) \approx 0.07885
\end{aligned}$$

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