CEE690-05 HW1
Zilin Yin

P1.
For all of the following answers, let 1 denotes patient being positive, whereas 0 denotes patient being negative. And P denotes test being positive, and N denotes test being negative. And lastly, T denotes test outcome being true.

a. Given that p(P|1) = 0.95, p(N|1) = 0.05, p(N|0) = 0.95, p(P|0) = 0.05, p(1) = 0.0007, p(0) = 0.9993, we have:
   p(P) = p(P|1) * p(1) + p(P|0) * p(0) = 0.95 * 0.0007 + 0.05 * 0.9993 = 0.05063
   Hence, p(1|P) = p(P|1) * p(1) / p(P) = 0.95 * 0.0007 / 0.05063 = 0.01313450523

b. Now that p(1) = 0.1, p(0) = 0.9, we have:
   p(P) = p(P|1) * p(1) + p(P|0) * p(0) = 0.95 * 0.1 + 0.05 * 0.9 = 0.14
   Hence, p(1|P) = p(P|1) * p(1) / p(P) = 0.95 * 0.1 / 0.14 = 0.67857142857

c. The accuracy of the first test is 0.95
   In detail: p(T) = p(T,1) + p(T,0) = p(P|1) * p(1) + p(N|0) * p(0) = 0.95 * 0.0007 + 0.95 * 0.9993 = 0.95

d. Given that p(P|1) = 0.9, p(N|1) = 0.1, p(N|0) = 0.999, p(P|0) = 0.001, p(1) = 0.0007, p(0) = 0.9993, we have:
   p(P) = p(P|1) * p(1) + p(P|0) * p(0) = 0.9 * 0.0007 + 0.001 * 0.9993 = 0.0016293
   Hence, p(1|P) = p(P|1) * p(1) / p(P) = 0.9 * 0.0007 / 0.0016293 = 0.3866691217

   Now that p(1) = 0.1, p(0) = 0.9, we have:
   p(P) = p(P|1) * p(1) + p(P|0) * p(0) = 0.9 * 0.1 + 0.001 * 0.9 = 0.0909
   Hence, p(1|P) = p(P|1) * p(1) / p(P) = 0.9 * 0.1 / 0.0909 = 0.9900990099

   The accuracy of the second test is p(T) = 0.9 * 0.0007 + 0.999 * 0.9993 = 0.9989307

e. When the overall rate of lung cancer is below certain value such that the computed overall accuracy score of the second test is higher than that of the first test, the second test is more preferable.

f. Given that p(P|1) = 0.1, p(N|1) = 0.9, p(N|0) = 0.999, p(P|0) = 0.001, p(1) = 0.0007, p(0) = 0.9993, we have:
   The accuracy of the test is p(T) = 0.1 * 0.0007 + 0.999 * 0.9993 = 0.9983707

g. Now that p(1) = 0.05, p(0) = 0.95, we have:

The accuracy of the test is p(T) = 0.1 * 0.05 + 0.999 * 0.95 = 0.95405

h. Accuracy is a good metric for these tests, as the test outcomes and patience's conditions are all binary (either positive or negative). One thing to pay attention to is that since we have multiple tests with different properties (e.g. the probability of a patient being positive weights more in the first test than in the second test), in practical, we would want to use different tests based on the overall rate of lung cancer, such that the overall accuracy score could be maximized.

P2.

a. Given $a^* = argmin \sum_{i=1}^{N}(x_i - a)^2$, taking the expansion of the $\sum_{i=1}^{N}(x_i - a)^2$ part yields:

$$(x_1 - a)^2 + (x_2 - a)^2 + \cdots + (x_N - a)^2$$
$$= {x_1}^2 - 2x_1a + a^2 + {x_2}^2 - 2x_2a + a^2 + \cdots + {x_N}^2 - 2x_Na + a^2$$

Taking the derivative of it yields:

$$-2x_1 + 2a - 2x_2 + 2a - \cdots - 2x_N + 2a$$
$$= 2(Na - x_1 - x_2 - \cdots - x_N)$$

Set it equal to 0 we have:

$$Na = x_1 + x_2 + \cdots + x_N$$
$$a = \frac{1}{N}(x_1 + x_2 + \cdots + x_{N)}$$
$$= \frac{1}{N}\sum_{i=1}^{N} x_i$$

b. Given model $p(x_i|a) = N(a, \sigma^2)$, we have the likelihood function of it as:

$$a^* = argmax \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - a)^2}{2\sigma^2}\right)$$

Taking the negative log of the function yields:

$$a^* = argmin\ log(\frac{1}{\sqrt{2\pi\sigma^2}}) \sum_{i=1}^{N}(x_i - a)^2 - 2\sigma^2$$

where $log(\frac{1}{\sqrt{2\pi\sigma^2}})$ and $\sum_{i=1}^{N} -2\sigma^2$ are just some constants, hence we have:

$$a^* = argmin \sum_{i=1}^{N} (x_i - a)^2$$

c. Given $Bernoulli(y; p) = p^y(1-p)^{1-y}$, we have:

$$\begin{aligned} &-\log Bernoulli(y; \sigma(z)) \\ &= -\log\left(\sigma(z)^y(1-\sigma(z))^{1-y}\right) \\ &= -ylog\big(\sigma(z)\big) - (1-y)\log\left(1-\sigma(z)\right) \end{aligned}$$

d. Given $Bernoulli(y_i; \sigma(b \odot x_i))$, we have the likelihood function of it as:

$$argmax \prod_i \sigma(b \odot x_i)^{y_i}(1-\sigma(b \odot x_i))^{1-y_i}$$

Then taking the negative log of the function yields:

$$\begin{aligned} &argmin \sum_{i=1}^{N} -y_i \log(b \odot x_i) - (1-y_i)\log\big(1-\sigma(b \odot x_i)\big) \\ &= argmin \sum_{i=1}^{N} l(y_i, \sigma(b \odot x_i)) \end{aligned}$$