**Question 1**:

The VC dimension of an interval in R is 2. We can easily show such an interval shatters two points by choosing points with coordinates a, b respectively, such that a < b. Then when a is the positive example, we choose interval [a, c] such that a < c < b. Likewise, when b is the positive example, we choose [c, b]. When both are positive, we choose [a, b]. When both are negative, we can choose [a-2, a-1].

The reason we can't shatter three points is that no matter how we arrange the points, we'll always be able to set the outer points to be positive and the inner point to be negative. To prove this rigorously would be challenging, but I'm operating under the assumption that we only need to provide intuition for why an interval in R can't shatter three points.

Therefore, the VC dimension of an interval in R is 2.

**Question 2**:

a)

We begin with the probability, which is the product of each individual term:

$$p(\theta|x) = \prod_{i=1}^{n} \frac{1}{\theta - 1} e^{\frac{-x_i}{\theta - 1}}$$

However, it is more useful to deal with the log likelihood, so we take the log of the probability to get:

$$L(\theta|x) = \sum_{i=1}^{n} \ln \frac{1}{\theta - 1} - \frac{x_i}{\theta - 1} = \sum_{i=1}^{n} -\ln(\theta - 1) - \frac{x_i}{\theta - 1}$$

Taking the derivate with respect to our parameter:

$$\frac{\partial L}{\partial \theta} = \sum_{i=1}^{n} -\frac{1}{\theta - 1} + \frac{x_i}{(\theta - 1)^2} = \sum_{i=1}^{n} \frac{x_i - \theta + 1}{(\theta - 1)^2}$$

Setting the derivate to 0 yields:

$$\sum_{i=1}^{n} \frac{x_i - \theta + 1}{(\theta - 1)^2} = 0 => \sum_{i=1}^{n} x_i - \theta + 1 = 0 => \sum_{i=1}^{n} x_i - n\theta + n = 0$$

And solving for the parameter gives:

$$\theta = \frac{\sum_{i=1}^{n} x_i}{n} + 1$$

b)

The likelihood can be expressed as the following product:

$$p(\theta|x) = \prod_{i=1}^{n} (\theta - 1) x_i^{\theta - 2}$$

Taking the log yields:

$$L(\theta|x) = \sum_{i=1}^{n} \ln(\theta - 1) + (\theta - 2) \ln x_i$$

Now taking the derivative:

$$\frac{\partial L}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{\theta - 1} + \ln x_i = \frac{n}{\theta - 1} + \sum_{i=1}^{n} \ln x_i$$

Setting the derivative to 0 and solving:

$$\frac{n}{\theta - 1} + \sum_{i=1}^{n} \ln x_i = 0$$

$$-n = (\theta - 1) \sum_{i=1}^{n} \ln x_i$$

Which eventually gives us:

$$\theta = -\frac{n}{\sum_{i=1}^{n} \ln x_i} + 1$$

c)

We can see that the probability can be expressed as:

$$p(\theta|x) = \frac{1}{\theta^n}$$

However, this function is monotonically decreasing, so we want to choose the smallest theta we can. We know that 0 <= x <= theta, and we also know that theta > 0. Therefore, we select theta to be the largest x, since this will be the minimum value which satisfies all constraints. Note that this will not work if all values of x are 0. In that case, we actually have a problem, since there is no smallest value greater than zero (reals are uncountable); I assume this is not intended, and we are to assume at least one x is not 0.

**Question 3**:

a)

For a given sample, a natural classification rule would be to take the probability of the sample being in C1 minus the probability of it being in C2. If this subtraction yields a number greater than 0, we classify the sample as being in C1; otherwise, we say the sample is in C2. The expression in question can be expressed as follows:

$$P(C_1|x) - P(C_2|x) \geq 0$$

Using Bayes's Rule, we can rewrite this expression in terms we may know:

$$\frac{P(x|C_1)P(C_1)}{P(x)} - \frac{P(x|C_2)P(C_2)}{P(x)} \geq 0$$

And multiplying out the denominators gives:

$$P(x|C_1)P(C_1) - P(x|C_2)P(C_2) \geq 0$$

Now we'd like to take this general formula and write it in terms of the specific values we're interested in. In particular, here is the expression when we plug in x = 0:

$$P(x = 0|C_1)P(C_1) - P(x = 0|C_2)P(C_2) \geq 0$$

And since we know P(x=0|C1) and we know P(x=0|C2), we can write this in terms of totally known quantities, giving us a classification rule for x=0:

$$p_1 P(C_1) - p_2 P(C_2) \geq 0$$

Likewise, we can write the classification rule for x=1, since these are Bernoulli density functions:

$$(1 - p_1)P(C_1) - (1 - p_2)P(C_2) \geq 0$$

Thus we have a classification rule for both x=0 and x=1.

b)

We start with the general form of Bayes's rule for C1:

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x)}$$

Expanding P(x|C1) for the multidimensional case gives us:

$$P(C_1|x) = \frac{P(C_1)\prod_{j=1}^{D} P\left(x_j = 0\middle|C_1\right)^{(1-x_j)}\left(1 - P\left(x_j = 0\middle|C_1\right)\right)^{x_j}}{P(x)}$$

We now change our classification rule. Instead of subtracting like before, we divide and see if the resulting value is greater than 1. If it is, we classify the value as C1. Otherwise, we classify the value as C2. The reason for choosing division rather than subtraction will become apparent as we attempt to simplify this expression using logarithms:

$$\frac{P(C_1)\prod_{j=1}^{D} P_{1j}^{1-x_j}\left(1 - p_{1j}\right)^{x_j}}{P(C_2)\prod_{j=1}^{D} P_{2j}^{1-x_j}\left(1 - p_{2j}\right)^{x_j}} \geq 1$$

Taking the log of this expression and simplifying yields:

$$\ln\left(\frac{P(C_1)}{P(C_2)}\right) + \sum_{j=1}^{D}\left(1 - x_j\right)\ln\frac{p_{1j}}{p_{2j}} + x_j \ln\frac{1 - p_{1j}}{1 - p_{2j}} \geq 0$$

When this expression is greater than 0, we choose C1; otherwise, we choose C2. Thus we have a multidimensional Bayesian classifier. It is better to use the log version of this classifier because the product version may struggle due to floating-point precision. Note that when the Bernoulli distributions are 0 or 1, we need to adjust them slightly, as 0s and 1s will cause infinite values.

c)

To compute the posteriors, we need the more general version of Bayes Rule:

$$P(C_1|x) = \frac{P(C_1)\prod_{j=1}^{D} P_{1j}^{1-x_j}\left(1-p_{1j}\right)^{x_j}}{P(C_1)\prod_{j=1}^{D} P_{1j}^{1-x_j}\left(1-p_{1j}\right)^{x_j} + P(C_2)\prod_{j=1}^{D} P_{2j}^{1-x_j}\left(1-p_{2j}\right)^{x_j}}$$

Below I will compute all posteriors using the given priors:

When P(C1) = 0.2

$$P\left(C_1|(0,0)\right) = \frac{0.2 * 0.6 * 0.1}{0.2 * 0.6 * 0.1 + 0.8 * 0.6 * 0.9} = 0.027$$

$$P\left(C_1|(0,1)\right) = \frac{0.2 * 0.6 * 0.9}{0.2 * 0.6 * 0.9 + 0.8 * 0.6 * 0.1} = 0.69$$

$$P\left(C_1|(1,0)\right) = \frac{0.2 * 0.4 * 0.1}{0.2 * 0.4 * 0.1 + 0.8 * 0.4 * 0.9} = 0.027$$

$$P\left(C_1|(1,1)\right) = \frac{0.2 * 0.4 * 0.9}{0.2 * 0.4 * 0.9 + 0.8 * 0.4 * 0.1} = 0.69$$

$$P\left(C_2|(0,0)\right) = \frac{0.8 * 0.6 * 0.9}{0.2 * 0.6 * 0.1 + 0.8 * 0.6 * 0.9} = 0.97$$

$$P\left(C_2|(0,1)\right) = \frac{0.8 * 0.6 * 0.1}{0.2 * 0.6 * 0.9 + 0.8 * 0.6 * 0.1} = 0.31$$

$$P\left(C_2|(1,0)\right) = \frac{0.8 * 0.4 * 0.9}{0.2 * 0.4 * 0.1 + 0.8 * 0.4 * 0.9} = 0.97$$

$$P\left(C_2|(1,1)\right) = \frac{0.8 * 0.4 * 0.1}{0.2 * 0.4 * 0.9 + 0.8 * 0.4 * 0.1} = 0.31$$

When P(C1) = 0.6

$$P\left(C_1|(0,0)\right) = \frac{0.6 * 0.6 * 0.1}{0.6 * 0.6 * 0.1 + 0.4 * 0.6 * 0.9} = 0.14$$

$$P(C_1|(0,1)) = \frac{0.6 * 0.6 * 0.9}{0.6 * 0.6 * 0.9 + 0.4 * 0.6 * 0.1} = 0.93$$

$$P(C_1|(1,0)) = \frac{0.6 * 0.4 * 0.1}{0.6 * 0.4 * 0.1 + 0.4 * 0.4 * 0.9} = 0.14$$

$$P(C_1|(1,1)) = \frac{0.6 * 0.4 * 0.9}{0.6 * 0.4 * 0.9 + 0.4 * 0.4 * 0.1} = 0.93$$

$$P(C_2|(0,0)) = \frac{0.4 * 0.6 * 0.9}{0.6 * 0.6 * 0.1 + 0.4 * 0.6 * 0.9} = 0.86$$

$$P(C_2|(0,1)) = \frac{0.4 * 0.6 * 0.1}{0.6 * 0.6 * 0.9 + 0.4 * 0.6 * 0.1} = 0.069$$

$$P(C_2|(1,0)) = \frac{0.4 * 0.4 * 0.9}{0.6 * 0.4 * 0.1 + 0.4 * 0.4 * 0.9} = 0.86$$

$$P(C_2|(1,1)) = \frac{0.4 * 0.4 * 0.1}{0.6 * 0.4 * 0.9 + 0.4 * 0.4 * 0.1} = 0.069$$

When P(C1) = 0.8

$$P(C_1|(0,0)) = \frac{0.8 * 0.6 * 0.1}{0.8 * 0.6 * 0.1 + 0.2 * 0.6 * 0.9} = 0.31$$

$$P(C_1|(0,1)) = \frac{0.8 * 0.6 * 0.9}{0.8 * 0.6 * 0.9 + 0.2 * 0.6 * 0.1} = 0.97$$

$$P(C_1|(1,0)) = \frac{0.8 * 0.4 * 0.1}{0.8 * 0.4 * 0.1 + 0.2 * 0.4 * 0.9} = 0.31$$

$$P(C_1|(1,1)) = \frac{0.8 * 0.4 * 0.9}{0.8 * 0.4 * 0.9 + 0.2 * 0.4 * 0.1} = 0.97$$

$$P(C_2|(0,0)) = \frac{0.2 * 0.6 * 0.9}{0.8 * 0.6 * 0.1 + 0.2 * 0.6 * 0.9} = 0.69$$

$$P(C_2|(0,1)) = \frac{0.2 * 0.6 * 0.1}{0.8 * 0.6 * 0.9 + 0.2 * 0.6 * 0.1} = 0.027$$

$$P(C_2|(1,0)) = \frac{0.2 * 0.4 * 0.9}{0.8 * 0.4 * 0.1 + 0.2 * 0.4 * 0.9} = 0.69$$

$$P(C_2|(1,1)) = \frac{0.2 * 0.4 * 0.1}{0.8 * 0.4 * 0.9 + 0.2 * 0.4 * 0.1} = 0.027$$

**Question 4**:

Below is a table of error values corresponding to different prior parameters. Note that I chose to make the table with the parameter rather than the actual prior, because I was not sure if the prior for C1 or the prior for C2 or both were supposed to be included. Using the parameter, we can easily compute the priors, so I think this shouldn't be a problem.

| Parameter | Error |
|---|---|
| -5.000000 | 0.235955 |
| -4.000000 | 0.202247 |
| -3.000000 | 0.224719 |
| -2.000000 | 0.213483 |
| -1.000000 | 0.235955 |
| 0.000000 | 0.280899 |
| 1.000000 | 0.280899 |
| 2.000000 | 0.325843 |
| 3.000000 | 0.325843 |
| 4.000000 | 0.325843 |
| 5.000000 | 0.314607 |

The error rate obtained for the test set using the best priors and the learned distributions is:

0.1461

This appears to be a pretty good classification rate, since it's significantly better than just always guessing C2 (which is the more frequent class).