Problem 1:

1. The Bayes assumption probably does not hold, because GC content and complexity might have some dependence. This is because if GC content is 0 or 1, then the complexity will be low, but if GC content is around 0.5, the complexity is likely higher. Another example is that given that you have a coding region in a gene, you know that length is directly correlated with high GC content.
2. P(Y): P(Gene)=3/10, P(Repeat)=3/10, P(Motif)=4/10

P(GC=low|Gene) = 2/3, P(GC=medium|Gene)=1/3, P(GC=high|Gene)=0

P(GC=low|Repeat) = 0, P(GC=medium| Repeat)=0, P(GC=high| Repeat)=1

P(GC=low|Motif) = 0, P(GC=medium| Motif)=1/2, P(GC=high| Motif)=1/2

P(Length=long|Gene)=1, P(Length=short|Gene)=0

P(Length=long| Repeat)=1, P(Length=short| Repeat)=0

P(Length=long| Motif)=0, P(Length=short| Motif)=1

P(Complexity=high|Gene)=2/3, P(Complexity|low |Gene)=1/3

P(Complexity=high| Repeat)=1/3, P(Complexity|low | Repeat)=2/3

P(Complexity=high| Motif)=3/4, P(Complexity|low | Motif)=1/4

1. P(gene) = p(gene(\*pg\_gc

Problem 2:

1. See notebook
2. P(S|C) > P(S|N) = ~12%
3. P(S|N) > P(S|C) = ~87%
4. We need to pick a pair of the scores that is helpful for discriminating
   1. 6 and 1
   2. 2 and 3

0 is twice as more likely to be in N than C. If we include 0, and the model is from N, then it is more likely for the model to be correct. If we include 0 and the model is from C, then it is less likely that the model is correct.

1. It is more likely for a rare event to happen, so we could not consider scores of 2 and 3 since they don’t add much to the model. We could also make the model more complex by looking at nucleotide pairs, triplets, or even kmers.

Problem 3:

1. Code Submitted
2. Plots Submitted
3. The algorithm didn’t pick up and separate two very distinct groups. To fix this problem, I would initialize random start points and run the algorithm a few times. I’d then pick the graph that has the best plot
4. To implement fuzzy k-means, instead of just assigning a single variable to each point assigning it to a center, I would add a vector, where each slot has the probability that the point belongs to the corresponding center.