Ariana Aimani 260501657
COMP 462 ASSIGNMENT 3:

1. **The consensus sequence for the GATA2 transcription factor is [A/T]GATAA**. For each combination of the alphabet and extended alphabet given, I counted the number of times the pattern occurred in the positive as well as the negative sequences while taking into account their z-scores (compared to the random occurrences). I then sorted my results by the z-scores for the positive sequences and found that TTATC[AT] has the highest z-score but it also has high z-scores in the negative sequences as well. Just based on the biology of the transcription factor, the TF is supposed to occur more in the positive sequences and rarely in the negative sequence. Therefore, the consensus sequence that satisfies these properties is [A/T]GATAA.
2.

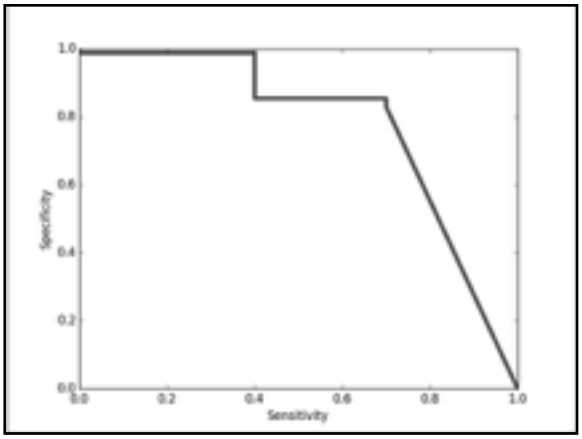
A.

| | Sensitivity | Specificity |
|---|---:|---:|
| Dataset1 | 0.700000 | 0.853659 |
| Dataset2 | 0.960000 | 0.804878 |

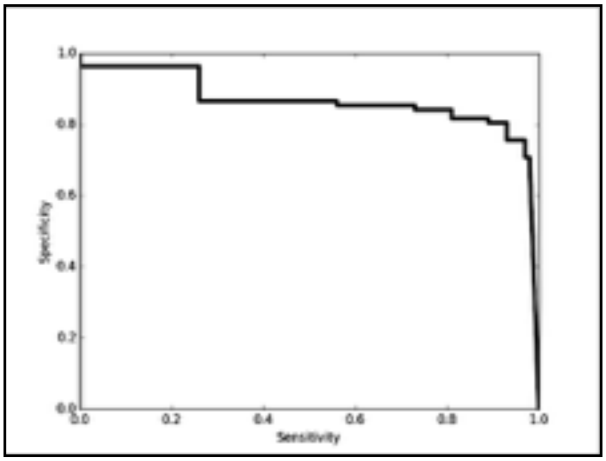
B.

| | Sensitivity | Specificity | Threshold |
|---|---:|---:|---:|
| Dataset1 | 0.4000000 | 0.987805 | 0.096342 |
| Dataset2 | 0.970000 | 0.756098 | 0.001031 |

C.

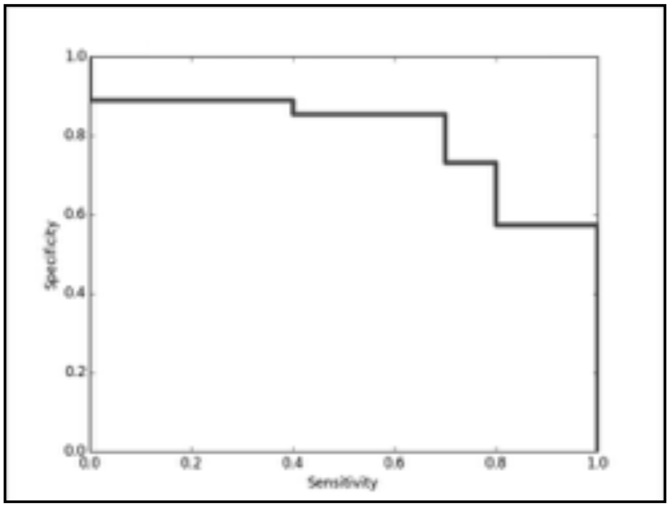


Dataset 1: Sensitivity-Specificity
AUC 0.760366

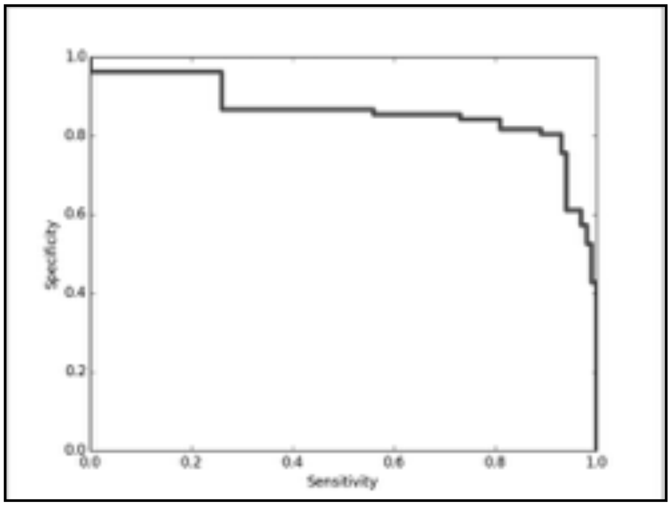


Dataset 2: Sensitivity- Specificity
AUC 0.864024

D.



Dataset 1: Sensitivity- Specificity
AUC 0.789634



Dataset 2: Sensitivity- Specificity
AUC 0.865915

The curves are better than those in (C), since the area under the curve (AUC) is larger. We can also see this visually since the curves for x=1 are farther away from the diagonals (i.e., closer to the top-right) than those in (c).