1. Consensus sequence: T G A|G C G C  
     
   For each combination of the extended alphabet, I counted the number of times the pattern occurred in the positive as well as negative sequences, in addition to their z-scores (compared to random occurrences). Sorting the results by z-scores for the positive sequences, I found the following possible candidates for the consensus sequence:

PATTERN POS\_COUNT POS\_ZSCORE NEG\_COUNT NEG\_ZSCORE

G C G C G C 6.0000 11.9770 16.0000 16.3454

C G C G C G 6.0000 11.9770 16.0000 16.3454

AT A A A A A 8.0000 11.0650 11.0000 7.0864

**T G AG C G C 8.0000 11.0650 1.0000 -0.5468**

T G AG CT G C 11.0000 10.4572 2.0000 -0.7732

While there are other sequences with higher z-scores for the positive sequences, they also have high z-scores for the negative sequences as well. One assumption I’m making is that the consensus sequence occurs more often in the positive sequences, and rarely in the negative sequences. Thus the sequence with the highest z-score that satisfies these properties is T G A|G C G C.

If we only count how many of the positive/negative sequences match the consensus sequence (i.e., multiple matches in one sequence only count once), we will get the same result since the same pattern now has the highest z-score for the positive sequences.

PATTERN POS\_COUNT POS\_ZSCORE NEG\_COUNT NEG\_ZSCORE

**T G AG C G C 7.0000 9.5967 1.0000 -0.5468**

T G AG CT G C 10.0000 9.4190 2.0000 -0.7732

T G A CT G C 6.0000 8.1285 1.0000 -0.5468

T G G C G C 4.0000 7.8241 1.0000 0.1531

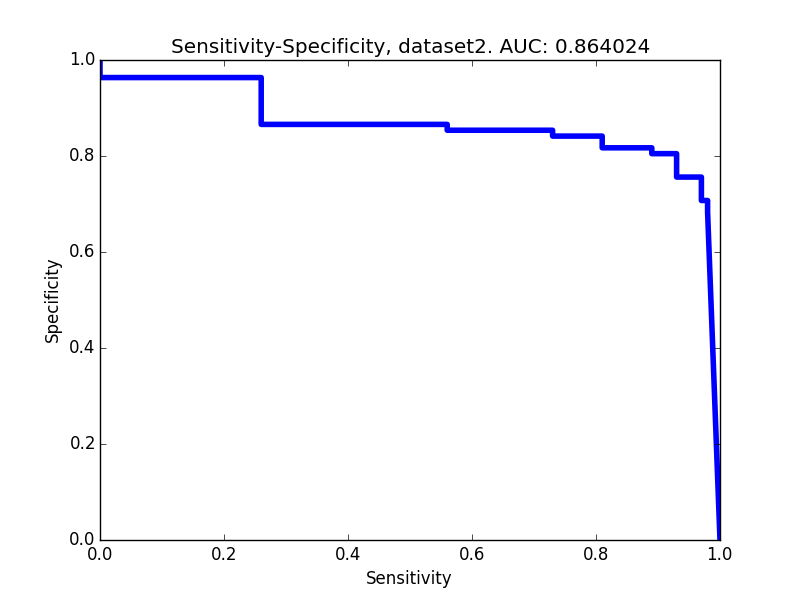
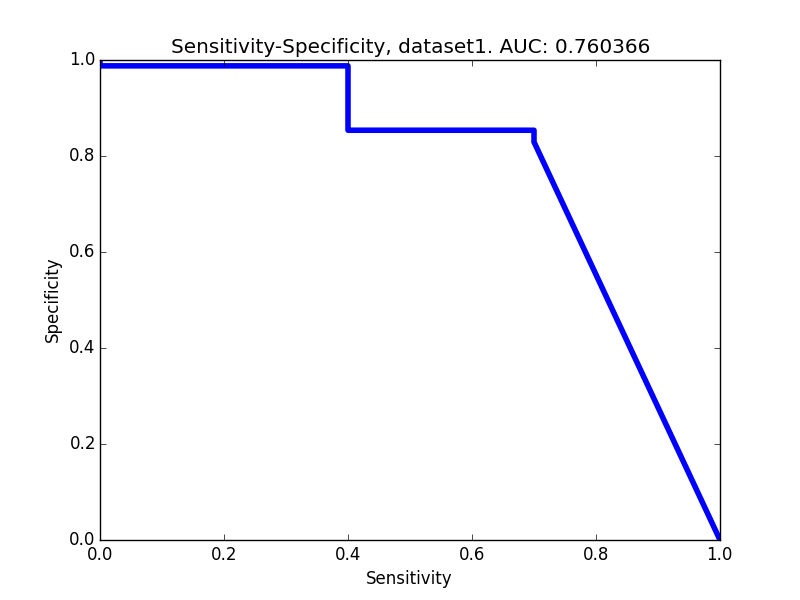
1. A.

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

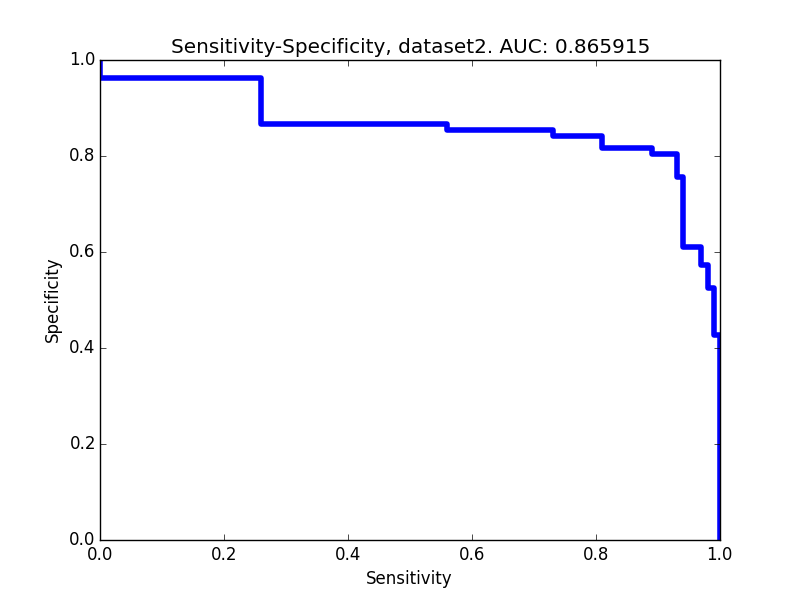
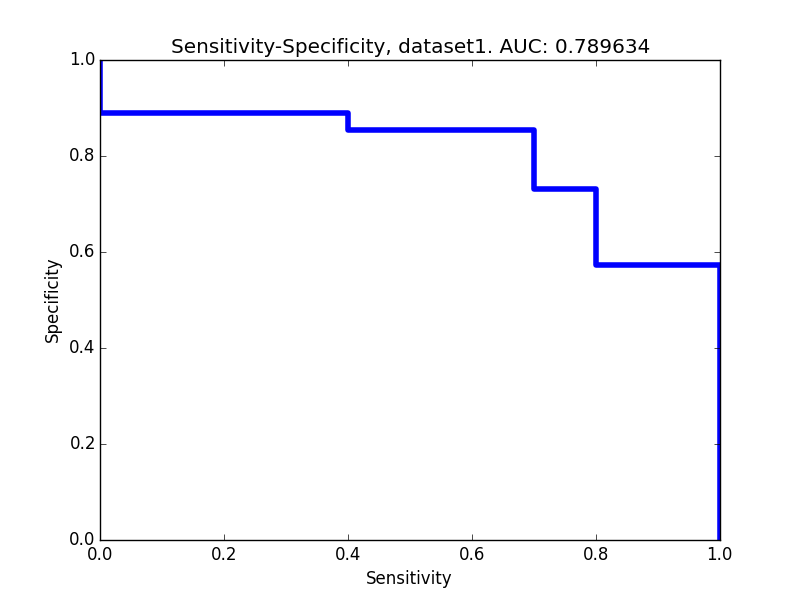
B.

|  |  |  |
| --- | --- | --- |
|  | Dataset1 | Dataset2 |
| Threshold | 0.096342 | 0.001031 |
| Sensitivity | 0.400000 | 0.970000 |
| Specificity | 0.987805 | 0.756098 |

C.



D.



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).