10. The performance of a nearest-neighbor recognition system will, in general, go down if you add measurements in dimensions that are unrelated to identity.

    True. If something to be identified accidently lies close to the wrong sample in irrelevant dimensions, it may seem closer to that wrong sample than to the sample that provides the proper identity.

Problem 2 Nearest Neighbors (25 points)

One of the key components of the question was the following statement:

> If your answer to the above question was **Yes**, SAW is curious about the "correct" nearest neighbor approach for this file. They ask you to recommend the *simplest* possible nearest neighbor classifier that will work for this data. The different possible approaches are listed below. The simplest technique is first and the most complex last. You need only give the corresponding number.

So in general when there are two approaches which might work, you should pick the *simplest*.

Another key piece of information was the following:

> The customers of SAW are picky about their widgets, if they pay extra for a GREAT widget they want to get a GREAT widget (i.e. **it is important to never classify a GOOD widget as a GREAT widget**). SAW loses the additional profit they might make by charging more for a GREAT widget, but they are not too worried about it (i.e. **you can classify some of the GREAT widgets as GOOD widgets, but fewer is better**).

From this you should conclude that classifying a "PLUS" as a "DIAMOND" is much worse that classifying a "DIAMOND" as a "PLUS".

Given this information the correct answers are:

Data File	Appropriate?	Which Approach?	Weights (If necessary)	
File A	Yes	1	$w1 =$	$w2 =$
File B	Yes	3	$w3 =$	$w4 =$
File C	Yes	2	$w5 = 1$	$w6 = 10000$
File D	No		$w7 =$	$w8 =$
File E	No		$w9 =$	$w10 =$

File A is "Yes" because the data is well separated; any of the nearest neighbor approaches will work.

File B is "Yes" because the GOOD widgets (the PLUSes) are clustered together tightly. In the region containing GOOD widgets there are only a few GREAT ones. In general if you find the 5 nearest neighbors

of any PLUS, at least 3 of them are PLUSes. So a 5-Nearest-Neighbor approach will almost always correctly classify the GOOD widgets. The 5-Nearest-Neighbor algorithm will incorrectly classify a few of the GREAT widgets, but we don't care too much about that.

File C is "Yes" because the data is separated, but only by sensor 6. Sensor 5 is ineffective. The range of values that sensor 6 takes on is much smaller than sensor 5, so the Euclidean distance places the PLUSes very near to the DIAMONDs. The solution is to stretch out the distances in the vertical direction. This is done by setting the weight for sensor 6, $w6$, to a very large value (with respect to $w5$). In general anything where $\frac{w6}{w5} > 3$ was accepted.

File D is "No". Let's examine cases: 1) Nearest neighbor places some DIAMONDS near PLUSES; 2) 5-Nearest-Neighbor does not help, the DIAMONDs are about equally far from a set of DIAMONDS as they are from a set of PLUSes; 3) There are no positive weights that bring DIAMONDS closer together, while keeping them away from the PLUSes – weights allow you to increase the sensitivity to variations in one of the sensors, but in this case neither sensor does the trick; 4) The combination of 2 and 3 does not help either.

File E is "No". The data are completely intermingled. PLUSes and DIAMONDS are not separated in any way