# Problem 2 Nearest Neighbors (25 points)

A local machining company, Self-adjusting Widgets (SAW), has contracted with you to help them analyze some problems in machine learning. SAW produces widgets on a completely automated assembly-line. At some point widgets must be sorted into two grades: "GREAT" and "GOOD". 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**).

In order to classify these widgets SAW intends to take a two measurement of each widget. SAW has two problems. First they need to pick suitable sensors to make these measurements. Once they have settled on sensors, they need to select an algorithm to perform the automatic classification. They want you to help them with both of these problems.

SAW sends you several files of data that have been generated on an assembly-line. Each file contains measurements of widgets taken with two sensors which are being evaluated. Each file of data is displayed as a set of points on a two dimensional graph: GREAT widgets are displayed as a 'diamond'; GOOD widgets are displayed as a 'plus'. For each file of data you are asked to determine if Nearest Neighbor classification can be used to automatically classify widgets on the assembly line. If you determine that nearest neighbor is a possible approach you must also specify which type of nearest neighbor algorithm is most appropriate.

The data that SAW sends you is to be considered *training data*. It is meant to show the natural types of variation in both the GREAT and GOOD widgets. You are being asked to comment on whether a nearest neighbor system trained with this data (has this data in memory) can be used to correctly classify widgets which are *not* in the *training data*. You can assume that this new data, which is often called the *test* data, has the same types of variations as the training data, but is not precisely the same.

Note: In the questions below you will be asked about Nearest Neighbor using a Euclidean distance between points. The Euclidean distance between two dimensional points $X$ and $Y$ is:

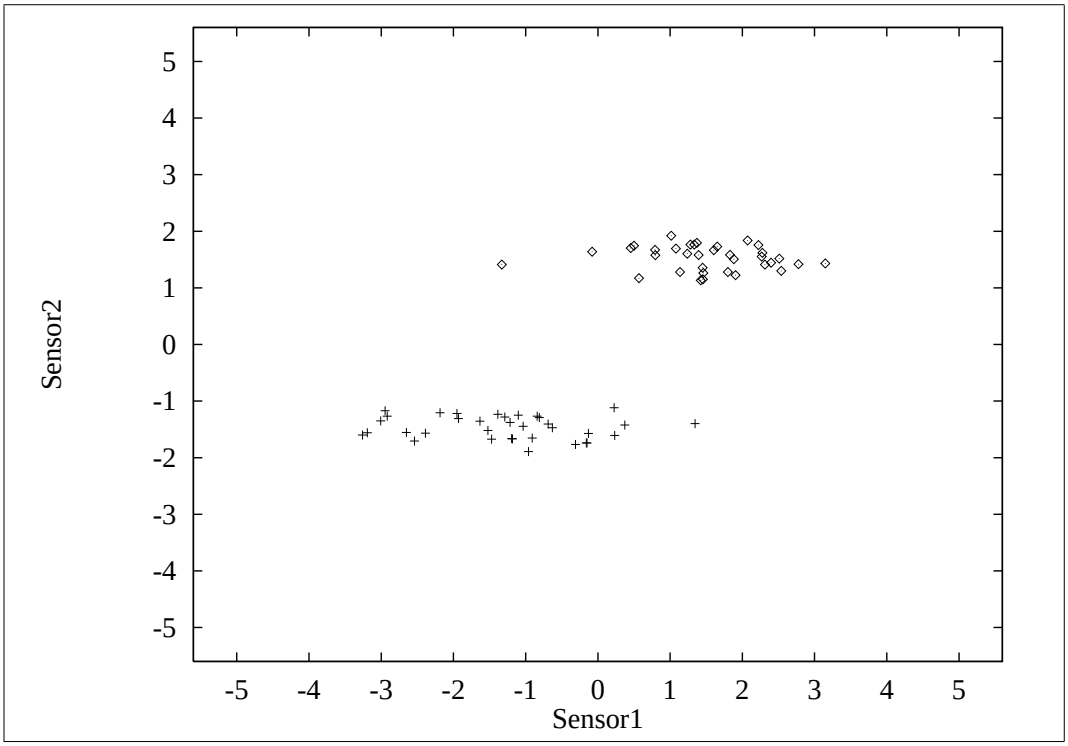
$$d(X,Y) = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2}$$

where $X_1$ is the value of dimension 1 and $X_2$ is the value of dimension 2. You will also be asked about Nearest Neighbor using weighted Euclidean distance. The weighted Euclidean distance between two points $X$ and $Y$ is:

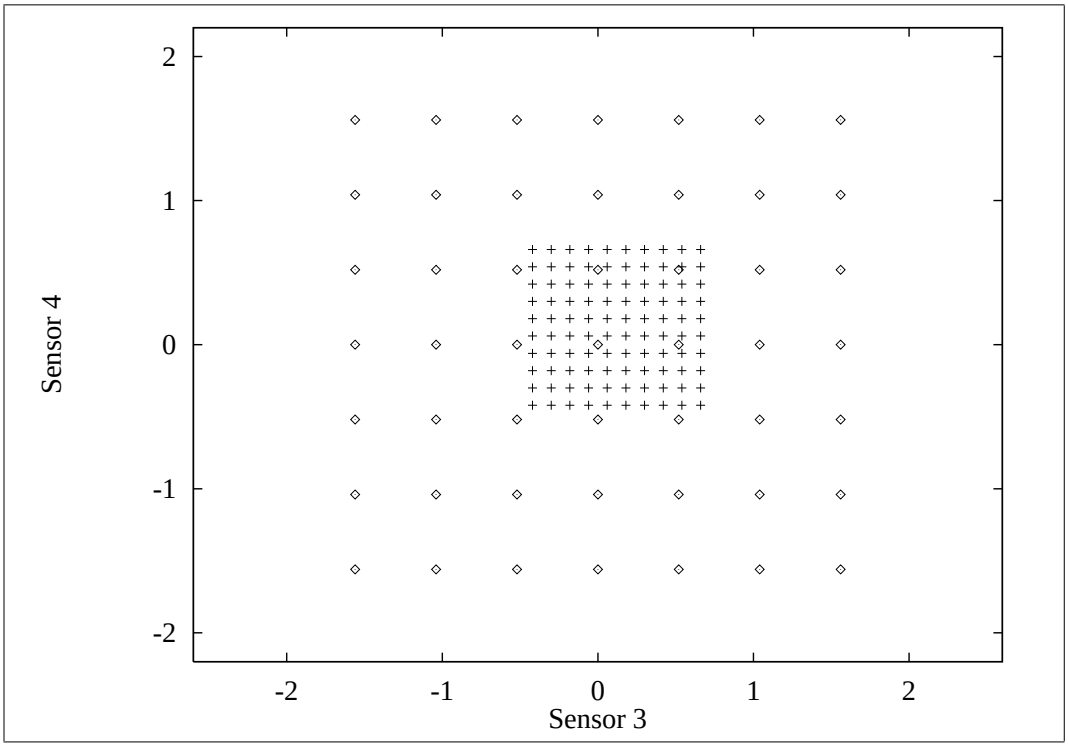
$$d(X,Y) = \sqrt{w_1(X_1 - Y_1)^2 + w_2(X_2 - Y_2)^2}$$

where $w_1$ and $w_2$ are used to differentially weight the two components of the distance.

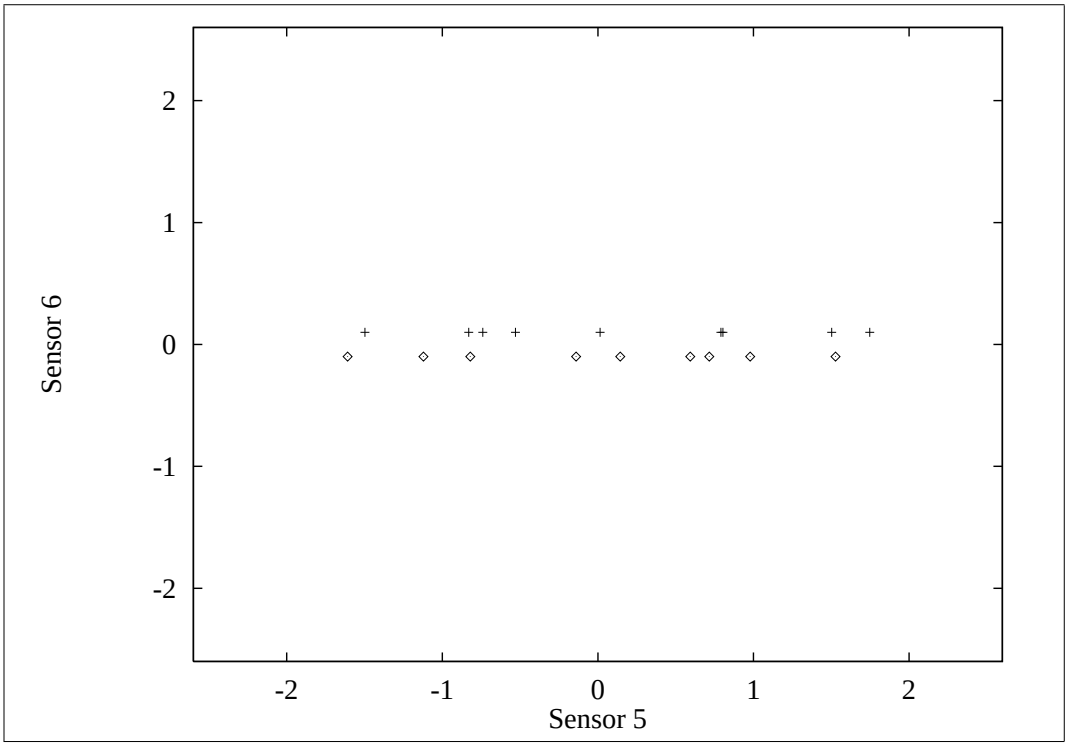
The graphs representing the five data files are on the following pages.



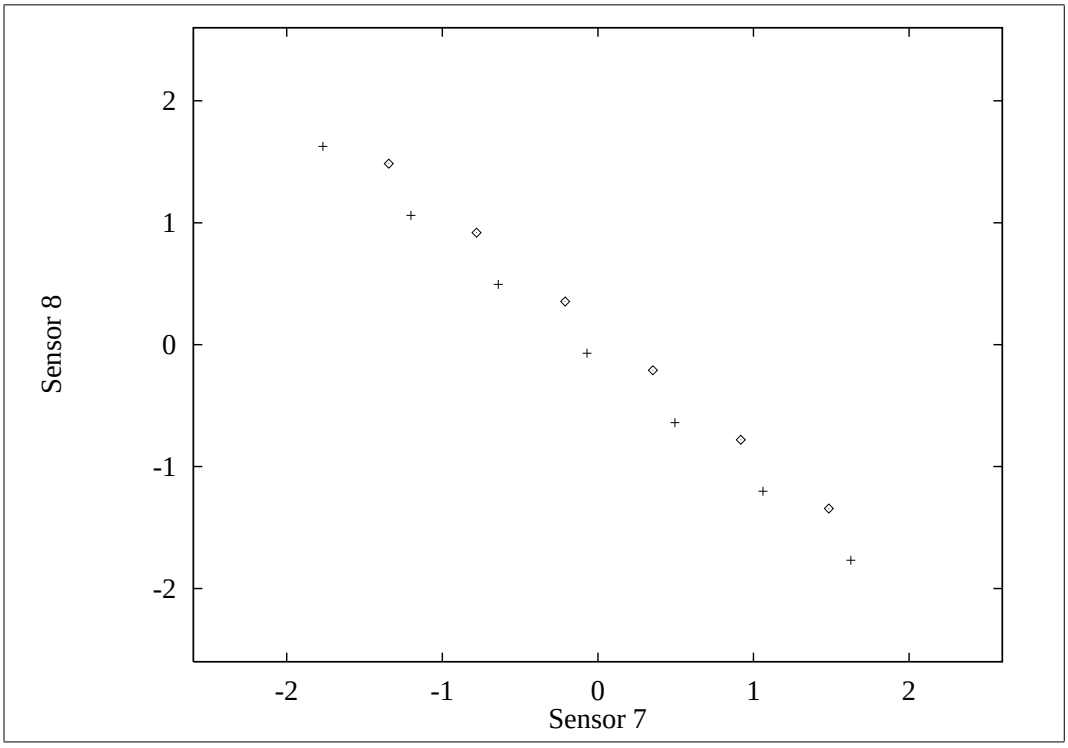

**Data from File A.**



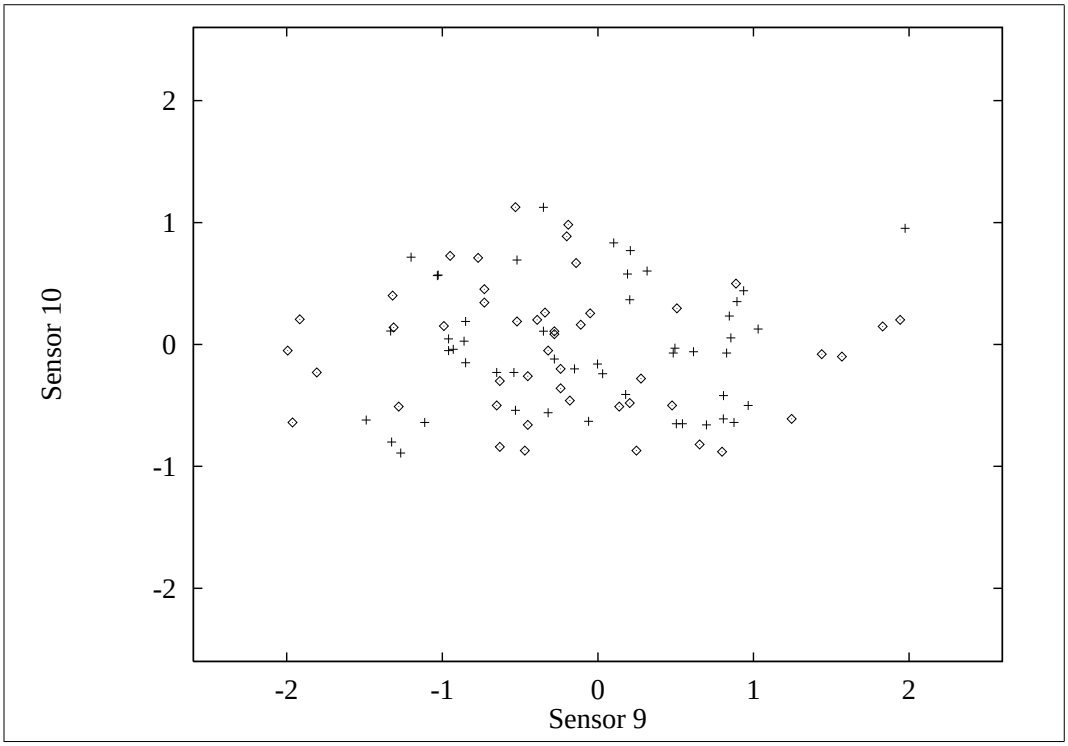

**Data from File B.**




**Data from File C.**




**Data from File D.**




**Data from File E.**

# Questions

For each of the five data files you must first determine if nearest neighbor classification is appropriate. You have one of two options:

- **Yes**: Nearest neighbor classification should work well. The probability of error will be low.
- **No**: This type of problem is too hard. Using nearest neighbor classification will result in too many errors.

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.

1. One Nearest Neighbor using Euclidean distance between points.
2. One Nearest Neighbor using a weighted Euclidean distance (the difference between each sensor value is multiplied by a sensor specific weight).
3. Five Nearest Neighbors (i.e. find the 5 nearest neighbor and take the majority vote).
4. Five Nearest Neighbors *plus* weighted Euclidean distance.

If you chose one of the weighted Eucldean distance techniques, give a set of weights for each sensor that would work (note: $w_i$ is the weight for sensor $i$).

Write your answers in the table below:

| Data File | Appropriate? | Which Approach? | Weights (If necessary) | |
|---|---|---|---|---|
| File A | | | $w1 =$ | $w2 =$ |
| File B | | | $w3 =$ | $w4 =$ |
| File C | | | $w5 =$ | $w6 =$ |
| File D | | | $w7 =$ | $w8 =$ |
| File E | | | $w9 =$ | $w10 =$ |

# Problem 3 Classy Objects d'Art (25 Points)

Arriving home after a particularly gruelling 6.034 recitation, you find a letter from a lawyer telling you that you have inherited some valuable works of art. These artworks, worth millions of dollars, have mistakenly been residing in art galleries around the world, but now they are yours.

It is only a matter of time before journalists start phoning you, asking all sorts of dumb questions about the objects you now own, and wanting background information about the artists who created them

Naturally, you decide to use the frame system in `frames.scm` (part of Problem Set 6) to represent knowledge about your new property. This way, journalists can quiz the frame system leaving you to appreciate the finer things in life.

A 6.034 professor drops by and tells you that the primary object classes in your frames systems should be `artwork` and `artist`; that artworks are either 2-dimensional or 3-dimensional; and that 2-d artworks are either paintings, drawings, or prints. As a consultancy fee, the professor takes away a Cezanne painting.

The lawyer's letter tells you that most of the art you've inherited is by Spanish artists. The famous 2-d artworks you've inherited include Picasso's *"Guernica"*, Dali's *"Autumnal Cannibalism"*, Miro's *"Birth of the World"*, all of which are paintings, and *"The Knight, Death, and the Devil"*, a print by Dürer. The only 3-d artworks mentioned by name are Michaelangelo's *"David"* and Rodin's *"The Kiss"*.

All you remember from high school art classes is that Michaelangelo was Italian, and Dürer was German.