KNN Classification Exercise

In this seminar you'll build a KNN classifier to predict the classes of cancer tumours based on the expression levels of particular genes (*ie* how much each gene is 'switched on' in the tumour cells). In the Excel spreadsheet you will find a subset of data taken from the following paper:

Khan,J., Wei,J.S., Ringnér,M., Saal,L.H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C., and Meltzer,P.S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*. 7(6): 673--679.
http://www.nature.com/nm/journal/v7/n6/abs/nm0601_673.html

Each instance gives the expression of a number of particular genes taken from a tumour sample, along with the known type of cancer. The classes are labeled as follows:

BL: Burkitt Lymphoma
EWS: Ewing Sarcoma
NB: Neuroblastoma
RMS: Rhabdomyosarcoma

The first worksheet gives the data for the expression values of just two genes for each sample, along with a 2D scatter plot of the data.

1. Looking at the plot, how successful would you expect KNN classification to be on this data set? Why?

2. The first task is to create a table that gives the distance between each pair of instances. The outline of the table has already been created, with the attributes (and classes) of each instance given down the side and repeated along the top. The first distance value has already been filled in. Check that you understand the formula that calculates the distance.
    a. Why is the first value 0?
    b. Fill in the rest of the table.
    c. Why is the resulting table symmetrical about the diagonal?

3. We now want to use the distance values to test the performance of a KNN classifier using 'hold-one-out' on a sample of instances (say one per class). That is, for each of the chosen instances we need to create a list of the classes of its nearest neighbours, sorted by distance. So copy the column of distances to each of the other instances to a new area, and copy the column of classes of those instances next to it. Then use the Data>Sort function to sort the classes by distance.
    a. If *K=1*, then what class would KNN classify this instance as?
    b. How about *K=2*, *3*, or *4*?
    c. How reliable is KNN classification on this data set?

4. The Excel file also contains another worksheet containing the same 20 instances, but with the gene expression levels of 5 genes rather than just 2. Repeat steps 2-3 for this data set – though this time you will have to write the equation for calculating the Euclidean distance between instances yourself.
    a. Is KNN any more reliable on this expanded data set?