Use the demo program eigenpens.sas to add the output from PROC PRINCOMP to your enterprise miner data sets. Pull the data into Enterprise Miner and set DIGIT as the target, PRIN1, PRIN2, and PRIN3 as the inputs, and set everything else to rejected.

Split the data into 70% training and 30% validation.

Run a default decision tree. How many leaves (terminal nodes) does it have? For the most accurate node, what digit does it choose and what percentage of the training data consists of that digit? Does the validation data agree on the choice of digit? What is the validation misclassification rate? Are there other leaves with as good performance? In the three dimensional space of principal components, what do the regions defined by the leaves look like? The region containing, for example, all the 2’s might be oval or kidney bean shaped or spiral, etc. From this information, explain why Enterprise Miner might not give us only 10 leaves, one for each digit, and still be pretty accurate.

Look at the “Leaf Statistics” plot in the results panel. What’s going on there? Is the information shown as complete as one might like?

We know that there are only 10 digits. Try fitting a decision tree with 10 leaves. For each leaf, list the digit chosen by the training data, the proportion of the chosen digit in the training and in the validation data, and the misclassification rate. Are there any digits that fail to be chosen by any leaf?

For the 10 leaf tree, suppose I have these values: PRIN1=-3, PRIN2=1, PRIN3=-1. What digit does the 10 leaf tree predict for this data point? Do the training and validation data sets agree on this choice? What proportion of the variation in the original 16 coordinates is picked up by the 3 principal components?

Explain how you would compute PRIN1 as a weighted average of your 16 observed coordinates X1 through X16 by listing the 16 weights that would be used (hint: go back to the demo program).

Write up any other analytics you try on the data. You might for example be reporting to a bank manager thinking about electronic scanning of handwritten deposit slips. For example, suppose you want to compare using the 3 principal components versus using X1-X16. You can use the metadata node (utilities subtab) to change X1-X16 to inputs and reject the principal components. Once done you have the two trees above and maybe a couple more. You can use the assess subtab to find a comparison node, connect all of your trees to it, and run it. This will compare the trees on several of the criteria we previously discussed.