19SP-IS590DT Assignment #7

This assignment deals with another type of OCR problems (digits).
Go to the repository: http://abel.lis.illinois.edu/data/ . The data directory contains 5 files related to a type of optical character recognition (OCR) problem. The filenames begin with optdigits: two csv files, a .names file, and a Python program. Your goal is to develop an understanding of how lusters/classifiers work in this particular domain.

a) Familiarize yourself with the documentation (the .names file) so you understand the problem domain and how the data was constructed.

b) Visualization. Use the Python program to get a feel for how the numerical tuple representation corresponds to the actual image. Frequently, you will see a tuple of five numbers, what are they? How are they related to the image? Provide a sample of image and justify your results.

c) Try to replicate the results of the original authors (i.e., use the instance-based classifier they reference and see if you get a similar accuracy). Hint: Remember to try different parameter values for the classifier and use the test set as provided for evaluation. Are there certain digits that are more difficult or easier to distinguish than others? If so, which ones? Optional: Identify an instance that was misclassified and display the corresponding image by modifying the provided Python program.

d) What are instance-based classifiers and what are non-instance based classifiers?
Use NB, BN and one instance-based classifier to do the digit classification. Compare the performance. Do you think instance-based classifiers will generally perform better in this problem domain? Why?

e) Suppose you expanded the dataset to include additional characters such as lower-cased English alphabetical characters, or any ASCII character, or any UTF-8 character. To what degree would you expect the accuracy to be reduced?
Hint: This question discusses the relationship between number of classes and the accuracy. You can try using the current database (but with filtering), test the classifier accuracy for the given dataset restricted to only 2, 3, 4, etc. of the digits and report the decrease in accuracy as a function of number of classes.

f) Suppose you were given a handwritten note and you scanned it into a pixel-based image. What kind of pre-processing steps do you imagine need to be performed before you can apply a trained character classifier?

g) find a reference paper / article / book / webpage / … on structural prediction. Summarize it in no more than two paragraphs.
Hint: use right reference formats. And it's always good to express your own thinking.