Appendix A

Final assignment

A.1 Case

For the purpose of this final assignment, a case is introduced. Assume your group plays the role of a pattern recognition consultancy company. Your company is faced with an assignment by a client company working on automatic bank cheque processing applications. This client wants you to research the possibility of using pattern recognition techniques to classify individual digits in bank account numbers and the monetary amount.

They are interested in two scenarios:

1. the pattern recognition system is trained once, and then applied in the field;
2. the pattern recognition system is trained for each batch of cheques to be processed.

In pattern recognition terms, Scenario 1 means the amount of training data available is large (in this case, at least 200 and at most 1000 objects per class), whereas for Scenario 2 it is much smaller (at most 10 objects per class).

A.2 Input: the NIST dataset

To help you to construct your system, the client has supplied you with a standard dataset of handwritten digits put together by the US National Institute of Standards & Technology, NIST (see `http://www.nist.gov/`). This set consists of 2800 images of handwritten digits for each of the 10 classes, "0","1", ..., "9". They were scanned from forms filled out by volunteers, thresholded in black-and-white and automatically segmented in images of 128 x 128 pixels. Sometimes parts of other digits are visible in an image. In order to save diskspace, we determined for each digit a bounding box, thereby causing images to have different sizes.

The data can be loaded and displayed as:

```
>> a = prnist([0:9],[1:40:1000])
>> show(a)
```

This loads, for each of the 10 digits, 25 objects out of the available 1000 into a `prdatafile` `a`. The result is shown in figure A.1.

Figure A.1: Subset of the NIST digit dataset

A.3 Expected output: deliverables

The client requires two deliverables:

- a report detailing the design choices you made *for both scenarios*, including detailed motivation. This report should discuss the following:
  1. the choice of representation (by pixels, features or dissimilarities);
  2. the actual features or dissimilarity measure used;
  3. for each representation, whether feature reduction is possible and, if so, what number of features should be used;
  4. for each representation, the optimal classifier and its type, i.e. parametric (`nmc`, `ldc`, `qdc`, `fisherc`, `loglc`), non-parametric (`knnc`, `parzenc`) or advanced (neural networks, support vector classifiers, one-class classifiers, combining classifiers);
  5. the estimated performance of the optimal classifier on novel data.
- a test of your system on a set of benchmark data withheld from you by the client (see below). For Scenario 1, your system should have lower than 5% test error; for Scenario 2, your target is 25% test error.

Note that each choice is expected to be backed up by either evidence or solid reasoning.

In the last exercise(s) for each week you have already experimented somewhat with the handwritten digit dataset. However, it is important to question each step as you construct your final digit recognition system, for two reasons:

- for Scenario 1, you should use as much training data as possible for the construction of your system (but *at least* 200 objects per class), rather than the smaller sets you used earlier; this may invalidate some of your earlier conclusions;
- you are supposed to optimise the overall performance of the system, rather than just the individual steps.

A.4 The benchmark

It is common for clients of your company to withhold data from you, which they can then use to test your system and verify whether the performance you claim to obtain is actually valid. In this case, the client has supplied you with a function you can use to benchmark your system yourself, without having access to the actual data used. It goes without saying that you may not use the test error to optimise your system in any way.

The procedure agreed upon is as follows:

- write a function `a = my_rep(m)` in which `m` is a NIST `prdatafile` set and `a` the resulting dataset;
- compute your classifier as PRTools classifier `w` in which dimension reduction and classification are incorporated, so `a*w*labeld` should generate the proper labels;
- call `e = nist_eval(filename,w,n)` in which `filename` is the filename for the `my_rep` routine, `w` is your classifier and `n` is the desired number of digits per class to be tested (which you can leave empty, if you want to test on all data). A typical call therefore looks like: `e = nist_eval('my_rep',w);`. The maximum value of `n` is 100. Evaluation may take a long time. The fraction of classification errors is returned in `e`.

In your report, give the performance obtained using the classifiers selected for both scenarios and compare these to the performance predicted using the training data. Are there large differences; if so, what could be their cause?

A.5 Grading

Your report will be graded by the supervisors, acting as the client company. If the minimum performance requirements (see section A.3 are not met, a passing grade will not be given. The report is graded on the results obtained, the choices based on their results and the motivation behind these choices. The maximum grade that can be obtained for the report is an 8.

The final two points can be earned as follows:

- Include a section "Live test" in your report, in which you report on the performance obtained on actual handwritten digits and compare it to the expected performance and the benchmark performance. To this end, you should:
  - use a scanner to read in an image of a sheet of paper on which you have written a test set of digits;

  - segment the individual digits out of the resulting image;
  - classify the digit images.

  You will have to organize the scanner yourself.

- Include a section "Recommendations" to your report, in which you advise your client in detail on possible steps to improve performance in light of the overall system they are constructing. For example:
  - will having more training data help?
  - would other features help?
  - does the application require reject options?
  - does a single classifier suffice for all digits on a bank cheque?
  - what is the role of time constraints?
  - etc.

A.6 Feedback

While you work on the final assignment during the last weeks, the supervisors will of course be available to answer questions and comment on your results.