Project: Wine recognition

Task

Tell if a particular wine (represented by a set of measurements) comes from one of three cultivars (all three types are grown in the same region in Italy).

Data

There are 178 observations representing different wine bottles. You are given 130 of these from training, and 48 are witheld for out-of-sample testing. There are 13 variables:

(a) Alcohol

(b) Malic acid

(c) Ash

(d) Alcalinity of ash

(e) Magnesium

(f) Total phenols

(g) Flavanoids

(h) Nonflavanoid phenols

(i) Proanthocyanins

(j) Color intensity

(k) Hue

(l) OD280/OD315 of diluted wines

(m) Proline

You are given the file *wineITrain.mat*, which contains the matrices *wineTrainX* (13×130), *wineTrainY* (1×130), and *wineTestX* (13×48). The first matrix, *wineTrainX*, contains the input patterns for the training data. The second matrix, *wineTrainY*, contains the output category $\in \{1, 2, 3\}$. The third matrix, *wineTestX*, contains the inputs for the test data. You are supposed to use the latter file to produce outputs that are handed in to me.

Steps and subgoals

1. Get acquainted with the data. Plot it and try to get a feel for the possible relationships between input and output.

2. Try to transform (e.g. Box-Cox transformation) the variables and see if this changes the information content (this can be measured using e.g. a "Fisher Index")..

3. Construct a Gaussian classifier (you can choose either quadratic or linear) for the problem. Try different combinations of variables too see if you can improve your classification.

4. Construct a linear logistic regression classifier for the problem (e.g. three linear logistic regressors, one for each class).

5. Construct a k-nearest neighbor (kNN) classifier for the problem, using all the variables, and estimate the generalization error (use $k \in \{1, 3, 5\}$).

6. Prune the kNN classifier by successively removing the variable that results in the least degradation of the generalization error, until the degradation is significant. Note the classification error (generalization).

7. Construct a multilayer perceptron (MLP) model using the remaining inputs. Optimize the number of hidden units (one hidden layer) with respect to the generalization error.

8. Try to prune the MLP model by successively removing the variable that results in the least degradation of the generalization performance, until the degradation is significant. Optimize the number of hidden units for the final model. Note the classification error.

9. Train a few networks with your optimal number of parameters. Combine these into a committee.

10. Hand in the test results for your best logistic regression model, your best kNN classifier, and your best MLP committee together with your estimate of the generalization classification error.

    Note: These test results must be handed in no later than 48 hours before your oral presentation.

10. Write a report.

Report and presentation of results

You will present the results from your project in two ways: (1) A written report where the main conlusions are presented together with figures and tables supporting your conclusions. (2) An oral presentation, of about 20 minutes, to your course colleagues.

The report should be about 10 pages, including figures and tables, and should contain the elementary report constituents:

- Introduction (brief presentation of problem, 1 page)
- Methodology (brief listing of methods, 1 page)
- Data (presentation of your data set with important observations, 1-2 pages)
- Results (4-5 pages)
- Discussion (your results and comparison to other researchers' results, 1 page)

The report writing should not take much more than one full day, since you are three persons sharing the work.

When you are finished with your report, and it has been accepted, then you should produce a postscript file with it, and pack it together with your dataset and other important parts of your project (like MATLAB M-files). The idea being that someone else could unpack it and repeat the main steps in your analysis without rewriting everything.