# Project: Laryngeal diseases

## 1. Task

Tell if a particular set of voice parameters comes from a person who is normal, or suffers from developed larynx mass lesions of two types: diffuse or nodular (i.e. 3 output categories).

## 2. Data—pitch and amplitude perturbation measures

There are 777 observations representing patients. You are given 501 of these, and 276 are withheld for out-of-sample testing. There are 24 variables (there is no information on what these represent). You are given the file *larynPertubTrain.mat*, which contains the matrices *trainLarynPertubInput* (501 × 24), *trainLarynPertubOutput* (501 × 3), and *testLarynPertubInput* (276 × 24). The first matrix, *trainLarynPertubInput*, contains the input patterns for the training data. The second matrix, *trainLarynPertubOutput*, contains the outputs coded in a 1-out-of-3 fashion. That is, the outputs are coded as (1,0,0), (0,1,0), or (0,0,1). The third matrix, *testLarynPertubInput*, contains the inputs for the test data. You are supposed to use the latter file to produce outputs that are handed in to me.

## 3. Steps and subgoals

1. Get acquainted with the data. Plot the data and try to get a feel for the possible relationships between input and output.
2. Construct a k-nearest neighbor (k-NN) classifier for the problem, using all the variables, and estimate the generalization error (use e.g. k = 5).
3. Prune the k-NN 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).
4. Construct a multilayer perceptron (MLP) using the remaining inputs. Optimize the number of hidden units (one hidden layer) with respect to the generalization error.
5. 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.

6. Train a few networks with your optimal number of parameters. Combine these into a committee.
7. Hand in the test results for your best k-NN classifier and your best MLP committee together with your estimate of the generalization classification error.
   **These results must be handed in no later than 48 hours before your oral presentation.**
8. Write a report.

## 4. Report and presentation of results

You will present the results from your project in two ways:

i. A written report, where the main conclusions are presented together with figures and tables supporting your conclusions.
ii. An oral presentation, of about 15 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)

When you are finished with your report, and it has been accepted, then you should produce a pdf file with it, and pack it together with your data set 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.