# Breast cancer

## Task:

The task is a classification task: To tell if a patient has a benign or malign breast cancer, based on image features from a Fine Needle Aspiration (FNA).

The diagnosis test is done in the following way: (description taken from the original source, which I will not tell here so that it is not too easy to find the original data)

- An FNA is taken from the breast mass. This material is then mounted on a microscope slide and stained to highlight the cellular nuclei. A portion of the slide in which the cells are well-differentiated is then scanned using a digital camera and a frame-grabber board.
- The user then isolates the individual nuclei using an image processing software.
- When all (or most) of the nuclei have been isolated, values for each of ten characteristics of each nuclei are computed, measuring size, shape and texture. The mean, standard error and extreme values of these features are computed, resulting in a total of 30 nuclear features for each sample.

The ten nuclei characteristics are:

(a) radius (mean of distances from center to points on the perimeter)

(b) texture (standard deviation of gray-scale values)

(c) perimeter

(d) area

(e) smoothness (local variation in radius lengths)

(f) compactness (perimeter$^2$/area - 1.0)

(g) concavity (severity of concave portions of the contour)

(h) concave points (number of concave portions of the contour)

(i) symmetry

(j) fractal dimension ("coastline approximation" - 1)

## Data

There are 569 observations, of which 400 are provided to you for training. You are given a file, *cancerWTrain.mat*, which contains the matrices *cancerTrainX* ($30 \times 400$), *cancerTrainY* ($1 \times 400$), and *cancerTestX* ($30 \times 169$). The output is coded as $0 =$ benign and $1 =$ malign.

## 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. Compute the Fisher index for each variable and rank your variables according to the Fisher index. Estimate how many of your variables that contain significant information about the problem (when considered one by one).
3. Construct a linear logistic regression classifier, trying different feature subsets in forward selection (follow the rankings by the Fisher index).
4. Construct a multilayer perceptron (MLP) model using all the inputs, and using the best inputs for the linear classifier above. Optimize the number of hidden units (one hidden layer) with respect to the generalization error.
5. Estimate the generalization errors for your models. Try constructing a committee model from the MLP models.
6. Produce the classifications for the test inputs and mail to me (for your best linear and MLP models), together with your estimates for how well you will do on the test samples.

## 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 two 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.