# Project: 2DE gel spot classification

## Task

Tell if a particular pixel on a 2DE gel comes from a true protein spot instead of from some contamination or artefact in the gel.

## Data

You are given the file *trainingSpot.mat*, which contains the matrices *training_bg_pixels2* ($21087 \times 8$) and *training_spot_pixels* ($9449 \times 8$). These represent pixels from the background (bg) and from true protein spots (spot). Each pixel is represented with a set of eight variables:

1-2. Strength and type of local spiral/circular symmetry.

3-4. Measures of the variation in the local spiral structure.

5-6. Strength and direction of parabolic structure.

7. Local variation of parabolic structure.

8. Local convex/concave curvature.

## 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 the variables and see if this changes the information content (this can be measured using e.g. a "Fisher Index")..

3. Construct a $k$-nearest neighbor ($k$-NN) classifier for the problem, using all the variables, and estimate the generalization error (use e.g. $k = 5$).

4. Construct a Gaussian classifier for the problem, using all eight variables, and estimate the generalization error.

7. Construct an artificial neural network (ANN) model. Optimize the number of hidden units (one hidden layer) with respect to the generalization error.

8. Try to prune the ANN 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. (Optional) Build a nonlinear SVM (Gaussian kernel) classifier. The classifier should be optimized with respect to the set of input variables (use backward elimination, like in step 8 above), the value of $C$, and the value of $\gamma$. Use 10-fold cross-validation to do this.

10. When you're done, you will be provided with the inputs of a test set of pixels. Hand in the test results for your best $k$-NN classifier, best Gaussian classifier, your best ANN model, and your best SVM model together with your estimate of the generalization classification errors for each classifier.

## 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

The report writing should not take more than one full day.

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.