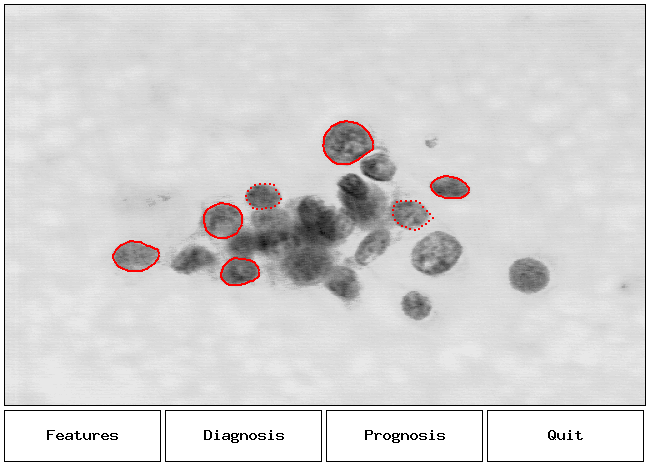
Breast Tumor Diagnosis Using Artificial Neural Networks.

In this project, my goal is to create an artificial neural network (ANN) for diagnosis of breast tumors. I am using a dataset from University of California Irvine Machine Learning Repository that consists of 569 labeled samples with 32 attributes. “Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.” There are no missing values for the attributes.

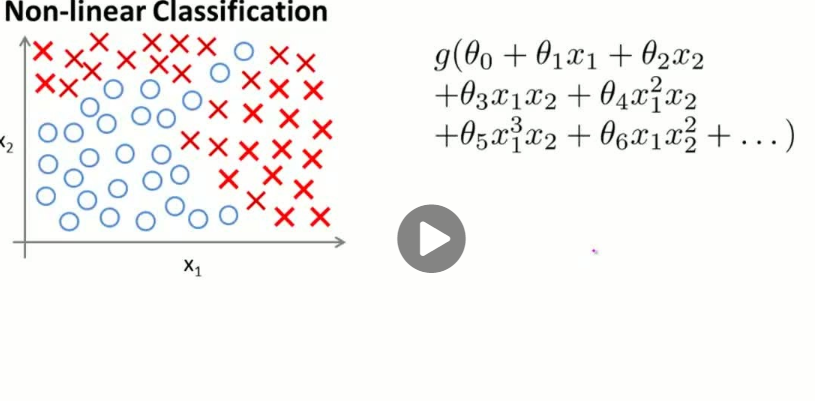
Information on attributes.

1. ID number
2. Diagnosis (M = malignant, B = benign)

“All 10 features are real-valued, all of them are numerically modeled such that larger values will typically indicate a higher likelihood of malignancy”. For each attribute the mean, standard error (SE), and “worst” (mean of 3 largest values) were computed. So, in my dataset field 4 corresponds to Mean texture, field 14 – Texture SE, field 24- Worst Texture.

1. Radius
2. Texture
3. Perimeter
4. Area
5. Smoothness
6. Compactness
7. Concavity
8. Concave points
9. Symmetry
10. Fractal dimension

Representation.

I decided my classifier to be a ANN without even visualizing my data, because they can compute any function in principle, unlike logistic or linear regression. Logistic regression works by learning the coefficients for attributes to best separate classes, in other words, the goal is to minimize the number of misclassified examples. However, it will not work if the hypothesis space is rather complex than linear. Consider the example below:

To learn this hypothesis function, we will need more than just a linear model, because there is no straight line that can separate the classes. If we decided to use 6 features to create a hypothesis using all the quadratic terms, we would have new features. Now, imagine having 30 features. So, using a ANN would be a much more computationally effective way.

Update: However, right now I tried to use just Logistic Regression and it ran really fast, because my data set is fairly small. Using 10-fold CV, I got an accuracy of 95%, however, authors of the research achieved %97.5 accuracy of predictions.