Machine Learning Project Proposal: Breast Cancer Tumor Classification

Vicky Bilbily and Selena Wang

Dataset Description

Breast Cancer Wisconsin (Diagnostic) Data Set from UCI Machine Learning

- 569 examples of tumours
- Binary classification: malignant vs. benign
- 10 attributes, all numbers with 4 significant digits
    - Radius
    - Texture
    - Perimeter
    - Area
    - Smoothness
    - Compactness
    - Concavity
    - Concave points
    - Symmetry
    - Fractal dimension
- Mean, standard error, and "worst" calculated for each feature → 30 features total
- No missing data
- Class distribution: 357 benign, 212 malignant

Problem

The problem is to correctly classify tumours as benign or malignant.

Algorithms

- K-Nearest Neighbour
- Logistic Regression
- Support Vector Machines
- Neural Networks

Evaluation

To learn our model-variables we will use k-fold cross-validation. In this situation we want to avoid false negatives more than anything, so to compare our models, we must evaluate the recall rate, but we will also consider %error as normal.

Progress

So far we have made the above decisions and attempted to load our data into Jupyter notebook.