Group Projects on Optimization.

P8160 Advanced Statistical Computing

Project: Breast Cancer Diagnosis

The data *breast-cancer.csv* have 569 row and 33 columns. The first column **ID** lables individual breast tissue images; The second column **Diagnonsis** indentifies if the image is coming from cancer tissue or benign cases (M=malignant, B = benign). There are 357 benign and 212 malignant cases. The other 30 columns correspond to mean, standard deviation and the largest values (points on the tails) of the distributions of the following 10 features computed for the cellnuclei;

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter$\hat{2}$ / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The goal of the exercise is to build a predictive model based on logistic regression to facilitate cancer diagnosis;

Your to-do-list is

1. Build a logistic model to classify the images into malignant/benign, and write down your likelihood function, its gradient and Hessian matrix.
2. Develop a Newton-Raphson algorithm to estimate your model;
3. Build a logistic-LASSO model to select features, and implement a path-wise coordinate-wise optimization algorithm to obtain a path of solutions with a sequence of descending λ's.
4. Use 5-fold cross-validation to select the best λ. Compare the prediction performance between the 'optimal' model and 'full' model
5. Write a report to summarize your findings.