ECE 532 Final Project Proposal
Vivian Hsiao

Title: Comparison of Machine Learning Methods for : a Classification Problem

Project Dataset: NIH Surveillance, Epidemiology and End Results (SEER) Program
Available: https://seer.cancer.gov

The SEER dataset is a publicly available dataset of cancers in the United States. Data collection began in 1973 and has since expanded to include more areas and demographics. There are 8,131,919 records total in the full database from 1973-2017; however, I intend to examine only papillary and follicular thyroid cancer data, of which there are around 60,000 entries of patients who were either alive at the time the dataset was finalized or perished as a direct result of thyroid cancer. Every row corresponds to a tumor. Variables include patient age, race, sex, tumor size (by various metrics) and extension, regional and distant lymph node status, staging (by various standards), and survival. It also includes the treatments received, including surgery, chemotherapy and radiation.

The classification will be 5-year survival versus death secondary to thyroid cancer, with a '1' label corresponding to death and '0' corresponding to survival greater than 5 years after initial diagnosis.

Algorithms that will be Investigated

1. ElasticNet Regularized Logistic Regression
   a. *Parameters: coefficients, regularization amount (lambda); LASSO/Ridge tradeoff parameter(alpha)*
   b. *Validation: 10-fold cross-validation on small grid (~10) of alpha values, 100 logarithmically spaced values of lambda; minimize classification error and select a sparse classification model within 1 mean-squared-error of minimum.*
2. Principal Components Analysis
   a. *Parameters: number of components to keep*
   b. *Validation: 10-fold cross-validation leaving out 10% of the data rows with every fold and minimizing residual error; evaluation on held-out test set.*
3. Artificial Neural Networks
   a. *Parameters: number of layers, number of neurons per layer*
   b. *Validation: 10-fold cross validation on parameters.*

Comparison: classification accuracy and confidence intervals will be compared for the three methods described above.

Project GitHub: https://github.com/vhhsiao/ece532seer.git

Project Timeline

Task	Description	Completion Date
Research	Review of literature of similar problems and algorithms used to address these	11/10/2020

Implementation Plan	High-level plan for implementation of algorithms and cross-validation, and comparision between algorithms	11/15
Implementation	Coding, debugging and testing implementation of algorithms	12/1/2020
Final Report	Writing of final report	12/15/2020