12/1 Final Project Interim Update

Predicting 10-year Thyroid Cancer Mortality using the SEER Database

Vivian Hsiao

Preprocessing:

I have completed the data cleaning process and have converted the original variables for age, sex, race, tumor grade, extension and size, and lymph node status into numeric variables, or converted to one-hot encoding for categorical variables. After excluding rare tumors and patients who had mortality attributable to other causes, there are 187,668 data points. For now, all missing data is assumed to be missing at random.

Next steps: implemen regularized regression, or principal components analysis (PCA)tation of the SMOTE algorithm to generate synthetic samples of the minority class; *k*-nearest neighbors to impute missing values, with *k* determined by 10-fold cross-validation, regularized regression, or principal components analysis (PCA).

Framework for cross-validation:

The data is split into a 70% training set and 30% holdout set at random (in a future iteration will consider splitting by data year – training on older years and validating prospectively) and the off-the-shelf ten-fold cross validation capabilities of the MATLAB *fitclinear* functions for logistic regression and SVM have been used to perform cross-validation.

Next Steps: transition from using off-the-shelf cross-validation option to coding framework that will allow parameter selection for regularized regression and artificial neural networks.

Classification:

Implemented off-the shelf logistic regression and support vector machines on the training data. These algorithms achieved average classification AUROCs of 0.8692 and 0.8204, respectively, using 10-fold cross-validation.

Next Steps: transition from using off-the-shelf implementation of logistic regression to implementing my own logistic regression function. Implement ElasticNet regularized regression, using the 10-fold cross-validation framework above to choose *lambda* and *alpha* (the tradeoff between Ridge and LASSO penalty terms) and perform feature selection. Implement artificial neural network.