Machine Learning to Inform Breast Cancer Post-Recovery Surveillance

Maxwell Allman (mallman@stanford.edu), Lin Fan (linfan@stanford.edu), Jamie Kang (kangjh@stanford.edu)
CS 229 Autumn Quarter 2017
Stanford University

Introduction

Motivation:
- Women who recover from breast cancer are often given surveillance care with the goal of early detection of relapses
- Currently in the US, there are no clear guidelines for surveillance care, and it is not known whether intense surveillance care is cost effective

Goal:
- To assess cost-effectiveness of surveillance care by using patient characteristics and the intensity and type of surveillance care undertaken to predict medical costs in the event of a relapse

Methods:
- Linear regression on polynomial maps of basic patient features with relevant polynomial feature selection via forward search, regularization and feature selection via ridge and LASSO regression, and dimensionality reduction via PCA/PCR

Result summary:
- No clear relationship between surveillance intensity/type and cost of care in the event of a relapse
- Regularization methods (ridge and LASSO) and PCA/PCR contributed to improving the model's test and training errors, but the poor quality of the data made high bias and variance unavoidable

Data

- Obtained the data for this work from the Stanford Prevention Research Center's Oncoshare data resource through Tina Seto.
- Database contains patient information for over 11,000 patients diagnosed with breast cancer between 2000 and 2014.
- Of these over 11,000 patients, 362 were recorded to have recovered and relapsed, and had sufficient information. We randomly selected 80% of the data as the training set and 20% as the test set.
- The relapse medical costs were computed by using the CPT codes of procedures done during each patient's relapse period, and a reference for the cost of each code.

Features

- The raw data patient features that we decided to be relevant include age at diagnosis, stage of cancer at diagnosis, several different biomarkers of the cancer, and cost/day of imaging and outpatient surveillance. There were 12 of these features of both continuous and categorical type in total.
- We also used second order polynomial terms of the raw input data, to capture interactions between the features and relapse cost of care.

Model Selection / Regularization

Linear regression + forward search:
- We fit a linear model with the polynomial features, using forward search and leave-one-out cross validation to find the most predictive subset of features, since the total number of features was large relative to the number of data points we had.

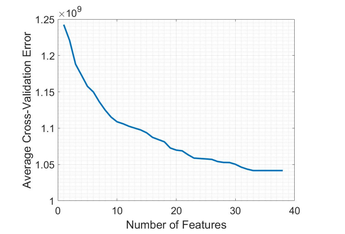

Figure 1: The average leave-one-out cross-validation error as the forward search proceeds.

Ridge regression:
- Minimize convex function with L^2 penalty on coefficients
- Shrinkage with L^2 penalty gives good predictive power

$$\hat{\beta}_{\text{ridge}} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{m} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

LASSO regression:
- Minimize convex function with L^1 penalty on coefficients
- Shrinkage with L^1 penalty has feature selection property

$$\hat{\beta}_{\text{lasso}} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{m} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

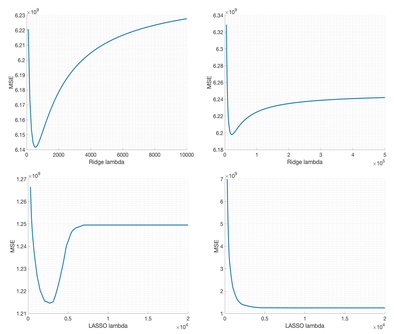

Figure 2: 5-fold CV to select penalty parameter λ.
Top / bottom: ridge / LASSO regression.
Left / right: 27 base set / 624 complete set of features

Principal Component Analysis:
- In order to reduce dimensionality, PCA selects dimensions U with the highest variances by solving: $\max_U \operatorname{Var}(Z^T U) \text{ s.t. } U^T U = 1$
- To prevent overfitting, we choose the first 141 PCs (out of 624) and retrain our model

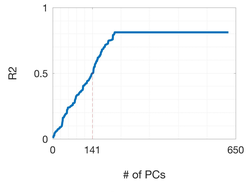

Figure 3: R^2 of PCA regression v.s. Number of PCs used

Results

Table 1: Training and test errors of different models/methods

	Training error	Test error
Linear Regression (45 features)	1.108e9	1.950e9
Ridge	1.123e9	1.855e9
LASSO	1.242e9	1.844e9
PCA/PCR (141 PCs)	6.202e8	2.657e9

This table shows the mean squared error for several different methods, where we randomly partitioned 80% of the data into a training set, and 20% of the data into a test set.

The test error when the predicted relapse cost is just set to be the average relapse cost over the training set, was 1.844e9.

Discussion

- Our dataset presented several difficulties: the number of patients who relapsed and had sufficiently complete data was quite small, and the patients' data we could use had some incomplete values.
- The linear regression with forward search found a set of 30 features, 5 of which depended on the two surveillance cost-per-day values. Taking an average over all of the patients, the partial derivative of the predicted relapse cost with respect to imaging surveillance cost-per-day was -2.12, and with respect to outpatient surveillance cost-per-day was 0.56.
- Overall, our model suffered from high bias and variance. In an attempt to resolve this, we utilized different regularization methods: ridge and LASSO regression.
- Ridge and LASSO regression gave similar predictive power on the test dataset. LASSO regression (starting with the 27 base set of features) found three features significant at the cross-validated optimal value of the λ parameter (imaging cost/day, ER biomarker, and HER2 biomarker).
- PCA successfully reduced the dimensionality from 624 features to 141 principal components, with R-squared value greater than 0.50. This resulted in a lower bias than in our original linear regression model, but could not avoid the overfitting issue.
- None of our methods gave a smaller test error than just setting the predicted relapse cost for every patient to be the average relapse cost of the training set. The underlying problem may be too noisy for us to be able to be able to find a significant predictive relationship with the amount and quality of the data we have.
- We found no clear relationship between intensity of surveillance and and relapse cost.

Future

- Our methods would very likely have better predictive power if we had access to more data, with less missing values.
- Missing data is an inevitability in medical research. We could investigate better methods of data imputation, as well as clean/screen the data in other ways.

References

• Stokes, Michael E., et al. *Ten-Year Survival and Cost Following Breast Cancer Recurrence: Estimates from SEER-Medicare Data*. Value in Health 11.2 (2008): 213-220.
•Hiramanek, N. *Breast cancer recurrence: follow up after treatment for primary breast cancer*. Postgraduate Medical Journal 80.941 (2004): 172-176.
•Cismondi, Federico, et al. *Missing data in medical databases: Impute, delete or classify?* Artificial Intelligence in Medicine 58.1 (2013): 63-72.