MSCR 509: High Dimensional Analysis

Homework 7

Anish Shah

March 23, 2020

Question 1

The dataset (lasso_homework_prostate_tumor) contains prostate cancer tumor gene data. The original data contains 6033 genes. For this exercise we have selected 29 genes. The sample size contains 102 observations (men), whereof 52 have prostate tumor and 50 do not have prostate tumor. The response variable Y is a binary variable that indicates the presence or absence of prostate tumor (cancer or healthy). The goal of the study is to build the model that correctly classifies the cancer and healthy observations.

Part A

Use logistic regression procedures (forward, backward) to develop a model for predicting cancer status. Read the LOG window and decide if it is successful, explain the reasoning.

% latex table generated in R 3.5.3 by xtable 1.8-4 package % Sun Mar 22 21:16:13 2020

	term	estimate	std.error	statistic	p.value
1	(Intercept)	2.7e-52	4.4e+04	-0.0027	1
2	X610	3.7e-50	2.5e+04	-0.0046	1
3	X739	2e+30	3.9e+04	0.0018	1
4	X921	7.9e+34	1.8e+04	0.0045	1
5	X1068	2.6e-54	3.2e+04	-0.0038	1
6	X1130	2.1e-57	2.8e+04	-0.0046	1
7	X1589	1.6e+25	1.7e+04	0.0034	1
8	X3647	5.2e-20	2e+04	-0.0022	1
9	X4316	4.4e+40	2.5e+04	0.0038	1
10	X4546	7.5e+49	2.4e+04	0.0049	1

This is not a successful model. There are several variables that have an OR of almost 0, or an OR that is approximating infinity. The p-value for the "best" variables by stepwise regression are all very close to ~1.

Part B

Using Lasso method develop a model to predict cancer vs normal subjects (Use correct coding in model statement to predict cancer... 1 = prostate cancer, 2 = no cancer).

The following covariates were found to be the most predictive using a LASSO method with a 50/50 test/training split. The optimal lambda was found to be 0.0305. The data as you can see has several variables that were shrunk to zero: *x332, x363, x914, x1113, x1130, x3665, x3991, x4546*.

% latex table generated in R 3.5.3 by xtable 1.8-4 package % Sun Mar 22 22:06:21 2020

	s0
(Intercept)	0.59
X332	0.04
X364	-0.05
X579	0.04
X610	0.00
X735	0.00
X739	-0.02
X914	0.05
X921	-0.00
X1068	0.07
X1077	0.03
X1089	0.08
X1113	0.00
X1130	0.06
X1346	-0.06
X1557	0.03
X1589	-0.02
X1720	0.02
X3375	0.04
X3647	0.00
X3665	-0.10
X3940	-0.02
X3991	0.00
X4073	0.00
X4088	-0.01
X4316	-0.01
X4331	-0.02
X4518	0.05
X4546	-0.08
X4549	0.00

Question 2

Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography

Part A

Briefly describe the goal of this article.

The goal of this article was to compare two different models in the evaluation of clinical characteristics into image analysis to improve breast cancer diagnosis.

Part B

Describe the statistical methods and results.

The researchers used a stepwise LR and a least absolute shrinkage and selection operator (LASSO) regression. The BIRADS score and teh clinical factors were used as the covariates. The models were fit with training data only, and then used for prediction on the test data. They used cross-validation for lambda selection. The BIRADS score, generated by radiologists, were pooled, with a majority agreement leading to a positive diagnosis.

The LASSO approach resulted in a misclassification rate of 0.234 versus 0.253 by stepwise regression. Using clinical covariates improved misclassification errors (from 0.194 to 0.234). Their overall LASSO approach was similar to that of the AUC of radiologists with BIRADS alone.

Part C

Based on this article, what are the advantages of the LASSO approach?

The advantages of LASSO are that it allows the inclusion of multiple covariates of unknown importance, and with its shrinkage methods it allows for the retention of only the important covariates. It outperforms stepwise regression, which can get “stuck” on non-informative parameters.