Overview of regularization techniques in machine learning

Debadrita Banerjee, Statistical Programmer, Novartis

IASCT Indian Association for Statistics in Clinical Trials

Introduction

Many clinical studies involve data on several prognostic factors and in many such cases we face the challenge of understanding how one factor/multiple factors impact the variable of our interest.

Regression is one of the simplest and most commonly used statistical modelling technique which helps to understand how the response variable(presence/absence of a disease) is effected by other predictor variables. However, most of the times we encounter computational problems in terms of multicollinearity and high dimensionality in data which might give us misleading results.

With this analysis, we wanted to take a look at the regularization techniques which might help us to obtain more accurate models and stable coefficient estimates.

Dataset: Pima Indian Diabetes dataset

The dataset includes data from **768** women with **8** characteristics

Predictor variables - Number of times pregnant, Plasma glucose concentration at 2 hours in an oral glucose tolerance test, Diastolic blood pressure , Triceps skin fold thickness , 2-Hour serum insulin ,BMI, Diabetes pedigree function,age

Response variable (binary) – A person is diabetic or not

Ridge	LASSO	Elastic Net
This is useful when most of the variables in the dataset are important	❖It works best when our dataset has less contributing variables. ❖Variable selection method	This overcomes the disadvantages of LASSO in case number of variables> number of observations
This has typically lower fit on training data than maximum likelihood, but is less prone to overfitting	❖If number of variables > number of observations, then LASSO selects at most n variables ❖If there is a group of variables with very high pairwise correlations, then LASSO tends to select only one variable from the group, not caring which one	

Methods

Logistic regression model: The unknown parameters are estimated by the maximum likelihood estimation method.

L= $\max\limits_{\beta_0,\beta} \sum_{i=1}^{n} \{y_i (\beta_0 + x_i^T\beta) - \log(1+\exp(\beta_0 + x_i^T\beta))\}$

Penalized logistic regression models:
Model parameters are estimated by maximizing : (-Log likelihood + Penalty term)
As a result, there is shrinkage in the the coefficients of the less contributive variables.
Method – 1.Ridge Regression
2.LASSO (Least Absolute Shrinkage and selection operator)
3.Elastic Net

Regularization technique	Penalty term
Ridge	$\sum_{j=1}^{p} \beta_j^2$
LASSO	$\sum_{j=1}^{p} \lvert\beta_j\rvert$
Elastic Net	$\lambda\left((1\text{-}\alpha) \sum_{j=1}^{m} \hat{\beta}_j^{\,2} + \alpha \sum_{j=1}^{m} \lvert\hat{\beta}_j\rvert \right)$

Results-coefficient shrinkage and goodness of fit measures

Predictor Variables	Logistic	Ridge	LASSO	Elastic Net
Intercept	-9.6489767991	-8.361930e+00	-8.511399601	-9.2758746221
Pregnant	0.1270608689	1.021882e-01	0.111145203	0.1201894176
Glucose	0.0370923369	2.919457e-02	0.032711648	0.0349054841
Pressure	0.0023492576	5.864331e-03	0.001220964	0.0032082966
Triceps	-0.0055337282	2.755907e-03	.	-0.0025412159
Insulin	-0.0009061723	4.378162e-05	.	-0.0006105627
Mass	0.0997692752	7.580730e-02	0.083051906	0.0923854462
Pedigree	0.9852287195	8.339425e-01	0.794565926	0.9401617428
Age	0.0100477173	1.038172e-02	0.007015281	0.0098923182

Model	Accuracy	Sensitivity	Specificity
Logistic	0.808	0.67	0.87
Ridge	0.804	0.51	0.93
LASSO	0.816	0.55	0.93
Elastic Net	0.820	0.57	0.93

Conclusion

From the results, we can clearly observe that the elastic net logistic regression model has shown the highest accuracy with 82.03% which is slightly more than all the other models

Here we demonstrated the use of regularization on a relatively smaller dataset. But in reality, whenever we have huge dataset with many correlated variables, we should check which of these three techniques will suit our needs and apply as required