MICROECONOMETRICS HW4

Haihua XIE(谢海花)-27720181153991

December 14, 2019

1 Introdeuction

When we do regression, or fit some regression models, we often need to filter the explanation variables. For example, we have thousands of candidate explanation variables. In general, instead of putting thousands of variables directly into the model for fitting training, we will use some methods to select some of these explanation variables, put them into the model, and form a list of variables into the model. So how do we select the input module variables?

The process of variable selecting (also called model selection in some cases) is a complex procedure, and many factors need to be take into consideration, such as: the prediction ability of variables, the correlation between variables, the simplicity of variables (easy to generate and use), the interpretability of variables in the business (the interpretability when challenged), etc. However, the most important and direct measure may be the prediction ability of variables.

2 Univariate variable selection

2.1 Based on correlation coefficients

Pearson correlation coefficient

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E\left(\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right)}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E\left(X^2\right)-E^2(X)}\sqrt{E\left(Y^2\right)-E^2(Y)}}$$

$$\rho_{X,Y}=\frac{n\sum_i^n X_iY_i-\sum_i^n X_i\sum_i^n Y_i}{\sqrt{n\sum_i^n X_i^2-\left(\sum_i^n X_i\right)^2}\sqrt{n\sum_i^n Y_i^2-\left(\sum_i^n Y_i\right)^2}}$$

Pearson correlation coefficient is commonly used, the value of coefficient is always in $[-1,1]$, the variables close to 0 are uncorrelated, and the variables close to 1 or -1 are called to have strong correlation. But it can only describe the linear relationship between two variables. If there is a nonlinear relationship between the covariates and the target variable, Pearson correlation is not applicable.

Rank correlation coefficient[1]

An increasing rank correlation coefficient implies increasing agreement between rankings. The coefficient is also inside the interval $[-1, 1]$, assumes the value equal to 0 if the rankings are completely independent.

Let r_{xi}, r_{yi} denote the ranks of the ith member according to the x quality and y quality respectively, then we can define $a_{ij} = \text{sign}(r_{xj} - r_{xi})$ and $b_{ij} = \text{sign}(r_{yj} - r_{yi})$. Then the $\sum a_{ij}b_{ij}$ is the number of concordant pairs minus the number of discordant pairs. While $\sum a_{ij}^2 = n(n-1)/2$, which is the number of a_{ij}, the same as b_{ij}.

Spearman's ρ rank correlation coefficient is:

$$\rho = \frac{\sum(r_{xj} - r_{xi})(r_{yj} - r_{yi})}{\sum(r_{xj} - r_{xi})^2}$$

Kedall's τ rank correlation coefficient is:

$$\tau = \frac{2(\text{number of concordant pairs} - \text{number of discordant pairs})}{\sqrt{n(n-1)}\sqrt{n(n-1)}}$$

If there are no identical two elements in the two set of two variables, then when the one variable can be expressed as a very good monotone function of another variable (i.e., the same trend of two variables), the absolute value of rank correlation coefficient between the two variables can reach 1.

2.2 Based on the entropy

Information entropy: information quantity measures the information brought about by a specific event, while entropy is the expectation of the information quantity that may be generated before the result comes out considering all possible values of the random variable, that is, the expectation of the information quantity brought about by all possible events. And the information entropy can be defined as follows:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

[1] Conover, W. J. Rank Tests for One Sample, Two Samples, and k samples Without the Assumption of a Continuous Distribution Function[J]. Annals of Statistics, 1(6):1105-1125.

Information entropy can also be used as a measure of the complexity of a system. If the more complex the system is, the more kinds of different situations appear, then its information entropy is relatively large. Information entropy is often used as a quantitative index of the information content of a system or a variable, so it can be further used as a criterion for the variable selection. As we have learned in class that the information entropy can be used to select variables in the classification tree.

3 Variable selection and model selection

3.1 Subset selection

In general, given a lot of predictors, we can use CV or information criteria to choose the best subset of predictors. This is also called variable selection. The idea is intuitive, we try all possible variable subsets, and then use a selection criteria, such as AIC, BIC, CV, to select the best variable set. **Forward stepwise selection, Backward stepwise selection**[2]**, Hybrid selection(combine the former two methos)** we have learned in this class are the main methods in subset selection. However, for large p, subset selection is not computationally feasible

3.2 Information criterion

We can use information criteria for model selection. These are metrics that make adjustment to the training error in order to account for the bias due to overfitting.

Akaike' s information criterion (AIC)

$$\text{AIC} = -2\log L + 2p = \text{Deviance} + 2p$$

where L is the value of the likelihood function, and p is the number of the unknown parameters.

Bayesian information criterion (BIC)

$$\text{BIC} = -2\log L + p\log N = \text{Deviance} + p\log N$$

where N is the number of data points.

For both AIC and BIC, smaller values are better. We can use them for model selection by choosing the model with the minimum AIC or BIC. The penalty $p\log N$ in BIC is usually

[2] https://github.com/jiamingmao/data-analysis

lager than that $2p$ in AIC, then BIC favors simpler models. And the penalties in AIC and BIC can both be called ℓ_0 penalty. Referring to the information criterion of information theory, MDL (minimum description length) is also used to select models.

4 Regularization and variable selection

In machine learning, when there are a lot of sample features and a relatively small number of samples, the model is easy to fall into overfitting. In order to alleviate the overfitting problem, we usually use regularization to select variables and models simultaneously.

In the regularization scheme, we often have an objective function of the form "loss+penalty"

$$\min_{f \in F}\left[\frac{1}{n} \sum_{i=1}^{n} L\left(y_{i}, f\left(x_{i}\right)\right)+\lambda J(f)\right]$$

where the $L(\cdot)$ is the loss function, $J(\cdot)$ is some penalty function, and λ is the tuning parameter. The principle of preventing over fitting: the penalty term is generally a monotonic increasing function of the model complexity, while the empirical risk is responsible for minimizing the error, so that the model deviation is as small as possible, the less the empirical risk is, the more complex the model is, the greater the value of the regularization term is. If the regularization term is small, the complexity of the model is limited, so the over fitting can be effectively prevented.

LASSO (Least Absolute Shrinkage and Selection Operator)[3]

The most popular regularization method in the past twenty years is LASSO method. Lasso was first proposed by Robert Tibshirani in 1996. It constructs a ℓ_1 penalty function to get a more refined model, it requires the sum of the absolute value of the forced coefficient is less than a fixed value, at the same time, it sets some regression coefficients as exactly zero.

$$\min \frac{1}{n} \sum_{i=1}^{n}\left(y_{i}-\beta^{T} x_{i}\right)^{2}+\lambda\|\beta\|_{1} \quad(\lambda>0)$$

The lasso can be rescaled so that it becomes easy to anticipate and influence what degree of shrinkage is associated with a given value of λ. Lasso regularized models can be fit using a variety of techniques including subgradient methods, least-angle regression (LARS), and proximal gradient methods.

[3] Tibshirani, Robert. Regression Shrinkage and Selection Via the Lasso[J]. Journal of the Royal Statistical Society, 58(1):267-288.

A number of lasso variants have been created in order to remedy certain limitations of the original technique and to make the method more useful for particular problems, such as adaptive lasso, fused lasso,group lasso etc. Almost all of these focus on respecting or utilizing different types of dependencies among the covariates.

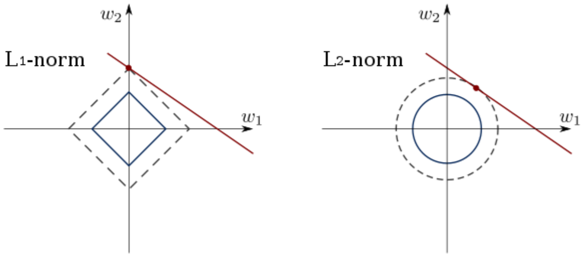

Figure 1: Forms of the constraint regions for lasso and ridge regression.

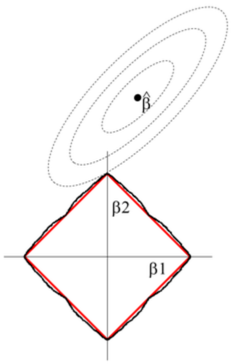

Figure 2: Forms of the constraint regions for lasso in two dimension.

Ridge regression

Ridge regression is corresponding to a ℓ_2 penalty:

$$\min \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\beta^T x_i\right)^2+\lambda\|\beta\|_2 \quad (\lambda>0)$$

Ridge regression is equivalent to artificially add a non negative factor to the main diagonal element of the information matrix composed of independent variables, which makes the matrix determinant non singular to reduce the error of regression coefficient estimation, improve the accuracy of estimation and the stability of the model. However, the ridge regression can only do shrinkage, and cannot make variable selection.

Elastic Net[4]

In 2005, Zou and Hastie introduced the elastic net to address several shortcomings of lasso. When $p > n$ (the number of covariates is greater than the sample size) lasso can select only n covariates (even when more are associated with the outcome) and it tends to select only one covariate from any set of highly correlated covariates. Additionally, even when $n > p$, if the covariates are strongly correlated, ridge regression tends to perform better.

$$\min \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \beta^T x_i\right)^2 + \lambda(\rho\|\beta\|_1 + (1-\rho)\|\beta\|_2) \quad (\lambda > 0)$$

So the result of the elastic net penalty is a combination of the effects of the lasso and Ridge penalties.

SCAD (Smoothly Clipped Absolute Deviation Penalty)[5]

Let $p(\cdot)$ denote the penalty function, Fan et. al introduced the SCAD penalty with the deviation:

$$p^{SCAD}(t)' = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\}, t \geq 0, \text{a} > 2$$

The penalty function is no longer convex, a family of penalty functions are introduced actually. SCAD can not only reduce the coefficients to zero exactly, but also obtain asymptotically unbiased estimates for larger coefficients, thus, the SCAD model has some orale properties. However, SCAD model has complex form and relative high computation cost. In the case of low noise level, the performance is better, but in the case of high noise level, the performance may be worse.

MCP (Minimax Concave Penalty)[6]

Zhang describes that the former methods do not address the uniqueness of the solution or provide methodologies for finding or approximating the local minimizer with the stated properties, among potentially many local minimizers. A major cause of computational and analytical difficulties in these studies of nearly unbiased selection methods is the nonconvexity

[4] Hui Zou, Trevor Hastie. Addendum: Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society Series B, 2005, 67.

[5] Fan J , Li R . Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties[J]. Publications of the American Statistical Association, 2001, 96(456):1348-1360.

[6] Zhang, Cunhui. Nearly unbiased variable selection under minimax concave penalty[J]. Annals of Statistics, 38(2):894-942.

of the minimization problem. So Zhang introduces a fast algorithm for nearly unbiased concave penalized selection in the linear model. And the MCP is defined as:

$$\rho(t;\lambda)=\lambda\int_0^t(1-x/(\gamma\lambda))_+dx$$

with a parameter $\gamma>0$.

The MCP provides the convexity of the penalized loss in sparse regions to the greatest extent given certain thresholds for variable selection and unbiasedness, at a universal penalty level, the MCP has high probability of matching the signs of the unknowns, and thus correct selection, without assuming the strong irrepresentable condition required by the LASSO.

5 Bi-level selection (group selection)

In 2006, Yuan and Lin introduced the **Group LASSO** [7] [8] in order to allow predefined groups of covariates to be selected into or out of a model together, so that all the members of a particular group are either included or not included. The objective function for the group lasso is a natural generalization of the standard lasso objective

$$\min_{\beta\in\mathbb{R}^p}\left\{\left\|y-\sum_{j=1}^{p}X_j\beta_j\right\|_2^2+\lambda\sum_{j=1}^{p}\|\beta_j\|_{K_j}\right\},\quad \|z\|_{K_j}=\left(z^tK_jz\right)^{1/2}$$

The penalty is the sum over the different subspace norms, as in the standard lasso, the constraint has some non-differential points, which correspond to some subspaces being identically zero. It can set the coefficient vectors corresponding to some subspaces to zero, while only shrinking others. However, it is possible to extend the group lasso to the so-called sparse group lasso, which can select individual covariates within a group.

A lot of such group style penalty functions have been introduced in the past twenty years, especially in the application in integrative analysis, we may have multi-view datasets, then we need to choose some group of variables, and also shrink some coefficients to exact zero to ensure simplicity and interpretability. If you are interested in such researches, you

[7] Ming Yuan, Yi Lin. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society, 68(1):49-67.

[8] Ogutu, Joseph O, Piepho, Hans Peter. Regularized group regression methods for genomic prediction: Bridge, MCP, SCAD, group bridge, group lasso, sparse group lasso, group MCP and group SCAD[J]. BMC Proceedings, 8(5 Supplement):S7.

may refer to group MCP [9], group SCAD[10], sparse group LASSO[11], etc. In some cases, we may encounter some super-high-dimensional dataset, especially in the gene studies, where $p >> n$, then the above method might not be very efficient when applied in such a dataset directly. An optional choice is to cut this dataset into some blocks, and using the methods mentioned before, it might be useful.

[9] Liu, Jin, Huang, Jian, Ma, Shuangge. Incorporating Network Structure in Integrative Analysis of Cancer Prognosis Data[J]. Genetic Epidemiology, 37(2):173-183.

[10] Wang, L, Chen, G, Li, H. Group SCAD regression analysis for microarray time course gene expression data[J]. Bioinformatics, 23(12):1486-1494.

[11] Simon, Noah, Friedman, Jerome, Hastie, Trevor. A Sparse-Group Lasso[J]. Journal of Computational and Graphical Statistics, 22(2):231-245.