ST 443: Group Project Instruction

Due by **4pm, 10 December, 2020**

1 Problems

1.1 Real World Data

The first part of the project is to apply statistical machine learning techniques on some real world data sets. The students are expected to find the data sets they are interested in from any resource and evaluate the sample performance of different regression or (and) classification approaches we have covered in class. I suggest the report includes

- Description of the data and the questions that you are interested in answering.
- Review of some of the approaches you tried.
- Summary of the final approach you used and the reason why you chose that approach.
- Summary of the results and conclusion.

1.2 Coordinate descent algorithm for solving the lasso problems

Tseng (1988) [see also Tseng (2001)] considers minimizing functions of the form

$$f(\beta_1,\dots,\beta_p) = g(\beta_1,\dots,\beta_p) + \sum_{j=1}^{p} h_j(\beta_j),$$

where $g(\cdot)$ is differentiable and convex, and the $h_j(\cdot)$'s are convex. Here each β_j cannot be overlapping. The author shows that coordinate descent converges to the minimizer of f. The project is to apply coordinate descent type of algorithm on penalized regression problems, e.g. the lasso in (1) and elastic net in (2).

Suppose the p predictors and response are standardized to have mean zero and variance 1. Take the lasso problem in (1) as an example.

$$\frac{1}{2n}\sum_{i=1}^{n}\big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\big)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{1}$$

Algorithm 1 summarizes the procedure for solving the ℓ_1 penalized optimization problem. See further details in Friedman et al. (2007).

Algorithm 1 Coordinate Descent Algorithm for Solving the Lasso

```
1. Initialize all the $\beta_j = 0, j = 1, \ldots, p$.
2. Repeat until convergence for $j = 1, \ldots, p$.
    (a) Compute the partial residuals $r_{ij} = y_i - \sum_{k \neq j} x_{ik}\beta_k$.
    (b) Compute the simple least squares coefficient of these residuals on the $j$-th predictor: $\beta_j^* = \frac{1}{n}\sum_{i=1}^n x_{ij} r_{ij}$.
    (c) Update $\beta_j$ by soft thresholding, see (2.f) in Homework 3 for details.
            $\beta_j = \text{sign}(\beta_j^*)(|\beta_j^*| - \lambda)_+,$
        where $(x)_+ = \max(x, 0)$.
```

Although the lasso is shown to be advantageous in many situations, it has some limitations. Consider the following three scenarios

1. In the case of $p > n$, the lasso selects at most n variables before it saturates.

2. If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and ignore the others, see Exercise 5 in Chapter 6 of ISLR for details.

3. For the usual case of $n > p$, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression (Tibshirani, 1996).

To overcome these limitations, the elastic net regularization considers adding both ℓ_1 and ℓ_2 penalties in the form of

$$\frac{1}{2n}\sum_{i=1}^n \Big(y_i - \sum_{j=1}^p x_{ij}\beta_j\Big)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2. \qquad (2)$$

The target of this project consists of

- Learn the "one-at-a-time" coordinate descent algorithm to solve the lasso problem in (1). Write the R code to implement the algorithm. (**You need to code by yourself without using existing R packages while solving the lasso**.)

- In analogy to Algorithm 1, develop coordinate descent algorithm to solve the elastic net in (2), see (2.h) in Homework 3. Write the R code to implement your developed algorithm. (**You need to code by yourself without using existing R packages while solving the elastic net problem**.)

- The simulation is to show that the elastic net not only dominates the lasso in terms of prediction accuracy, but also do a good job in variable selection. The simulated data is from the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \mathbf{I}_n).$$

The simulated data consists of a training set, an independent validation set and an independent test set. Models were fitted on training data only, the validation data were used to select the tuning parameters, e.g. $\hat{\lambda}$ in (1) and $\hat{\lambda}_1, \hat{\lambda}_2$ in (2). We compute the test error (the mean-squared error) on the test data. E.g., one can simulate 50 data sets consisting of $n_{train} = n_{validation} = 20$, $n_{test} = 200$ observations and eight predictors. Let $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma = 3$. The pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ was set to be $\text{corr}(i, j) = 0.5^{|i-j|}$. Report the mean (standard error) of **mean-squared errors**, **number of estimated nonzero coefficients** and other possible measures for the lasso and the elastic net, respectively.

- To possibly obtain a higher mark, you can also try other simulation settings (e.g. different values of n, p, σ, sparsity level of $\boldsymbol{\beta}$ and etc.) to demonstrate the superiority of elastic net over the lasso in many scenarios.

I suggest the report includes

- Your developed coordinate descent algorithms for the lasso and the elastic net.
- How you select the regularization parameters, e.g. $\hat{\lambda}$ in (1) and $\hat{\lambda}_1, \hat{\lambda}_2$ in (2).
- Describe the simulation settings and report the numerical results.
- Summarize your findings.

2 Timeline

- Week 5-7: Contact group members, decide who contributes to which part, search for the real world dataset and learn the coordinate descent approach. From the 7th week, please visit two GTAs' or my office hours to let us know the real data you decide to work on.
- Week 8-10: Analyse the data and write the R code to implement coordinate descent algorithm for the lasso and elastic net on simulated examples.
- Week 10-11: Write and submit the report.

3 Submission and assessment

3.1 Written report

The written report for the first part of the real data problem should be maximum of **4 pages**. The written report for the second part will not have any page limit, but I expect it to be less than **10 pages**.

Note the students are required to submit the reproducible R code for both parts of the project. Also the students should put additional figures, tables, mathematical derivations in the appendix.

Deadline to submit your coursework report: **4:00pm, 10th December 2020, Thursday of the 11th week.** Further details are to be announced.

3.2 Mark scheme

- Real data (50%)
- Coordinate descent approach (50%)

The grade for each part will be based on illustration of the dataset, the machine learning approaches you tried, discussion of the results, the readability and efficiency of R code, quality of layout and use of language and etc. The grade for each group member will be a function of the contribution of each group member using the following equation.

$$\text{member grade } = \text{report grade } \times \frac{\text{member contribution}}{\text{maximum contribution}}.$$

For example, for a group with 5 members contributing, 30%, 20%, 20%, 20%, 10% and the report grade is 75 (out of 100), the individual grades are

$$75 \times \frac{30}{30} = 75, \quad 75 \times \frac{20}{30} = 50, \quad 75 \times \frac{20}{30} = 50, \quad 75 \times \frac{20}{30} = 50, \quad 75 \times \frac{10}{30} = 25.$$

References

Friedman, J., Hastie, T. and Hofling, H. (2007). Pathwise coordinate optimization, *The Annals of Applied Statistics* **1**: 302–332.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of Royal Statistical Society, Series B* **58**: 267–288.

Tseng, P. (1988). Coordinate ascent for maximizing nondifferentiable concave functions, *Technical Report* .

Tseng, P. (2001). Convergence of block coordinate descent method for nondifferentiable maximiation, *J. Opt. Theory Appl.* **109**: 474–494.